

**İST405 - İSTATİSTİKSEL VERİ MADENCİLİĞİ FİNAL ÖDEVİ**

ÇANKAYA, ANKARA

GRUP NO: 15

ŞİMAL VİZE - 2210329027

İREM TARIM – 2210329047

DR. ÖĞR. ÜYESİ ONUR TOKA

BOOSTING ALGORİTMALARI KULLANILARAK WISCONSIN MEME KANSERİ VERİSİNİ SINIFLANDIRMA

2024

İçindekiler Tablosu

[1. GİRİŞ 3](#_Toc186140634)

[2. Veri Seti 5](#_Toc186140635)

[3. Veriye Ait Açıklayıcı Görselleştirme 6](#_Toc186140636)

[3.1. Korelasyon Matrisi 6](#_Toc186140637)

[3.2. ScatterPlot 8](#_Toc186140638)

[3.3. Violin Plot 12](#_Toc186140639)

[3.4. Pair Plot 12](#_Toc186140640)

[4. Veri Düzenleme 14](#_Toc186140641)

[5. Karar Ağaçları Modeli 15](#_Toc186140642)

[5.1. Boosting Algoritmalar ile İlişkisi 15](#_Toc186140643)

[5.2. Karar Ağacı Sonuçları 16](#_Toc186140644)

[5.3. Adaboosting ile Karar Ağacının Desteklenmesi 20](#_Toc186140645)

[6. Farklı Boosting Algoritmaları Kullanımı 22](#_Toc186140646)

[6.1. Adaboosting Nedir? 22](#_Toc186140647)

[6.2. Adaboosting Uygulaması 22](#_Toc186140648)

[6.3. Gradient Boosting Nedir? 24](#_Toc186140649)

[6.4. Gradient Boosting Uygulaması 25](#_Toc186140650)

[6.5. XGBoost Nedir? 27](#_Toc186140651)

[6.6. XGBoost Uygulaması 27](#_Toc186140652)

[6.7. LightGBM Nedir? 29](#_Toc186140653)

[6.8. LightGBM Uygulaması 30](#_Toc186140654)

[6.9. Catboosting Nedir? 32](#_Toc186140655)

[6.10. Catboosting Uygulaması 32](#_Toc186140656)

[7. En İyi Model Seçimi 34](#_Toc186140657)

[8. Kaynakça 35](#_Toc186140658)

# GİRİŞ

Meme kanseri, dünya genelinde kadınlar arasında en sık teşhis edilen kanser türlerinden biri olup, erken teşhis ve etkili sınıflandırma, hastalığın tedavisinde kritik bir öneme sahiptir. Dünya Sağlık Örgütü’nün (DSÖ) 2020 yılı verilerine göre, dünya çapında her yıl yaklaşık 2,3 milyon kadına meme kanseri teşhisi konulmakta ve meme kanseri, kadınlar arasında kansere bağlı ölümlerin %15'inden fazlasını oluşturmaktadır. Türkiye’de ise her 8 kadından birinin hayatının bir döneminde meme kanseri ile karşılaşacağı tahmin edilmektedir. BRCA1 veya BRCA2 mutasyonu gibi belirli gen mutasyonlarına sahip kadınlarda meme kanseri riski daha yüksektir. Meme kanseri teşhisi koyulduğunda kanser hücrelerinin meme içinde veya vücudun diğer bölgelerine yayılıp yayılmadığını da gözlemlemek için bazı testler yapılır. Evreleme işlemi birincil tümörün ne kadar büyüdüğüne, yayıldığı yerlere (yakındaki lenf düğümlerine veya vücudun diğer bölgelerine yayılıp yayılmadığına), derecesine ve ayrıca belirli tümör biyobelirteçlerin mevcut olup olmadığı incelenerek yapılır. Meme kanserinin teşhis ve sınıflandırma süreçleri, doğru tedavi planının oluşturulabilmesi için büyük bir hassasiyetle yürütülmelidir. Son yıllarda, tıbbi veri analizi alanında makine öğrenimi ve yapay zekâ tabanlı yaklaşımlar önemli bir ilerleme kaydetmiş ve bu teknolojiler sağlık sektöründe yeni bir dönemin kapılarını açmıştır.

Meme kanseri, çeşitli klinik belirtileri olan karmaşık bir hastalıktır. Göğüsteki hücrelerin anormal ve kontrolsüz bir şekilde büyüyüp tümörlü dokulara dönüşmesiyle meydana gelir. Çoğunlukla meme kanseri nedenleri arasında hareketsiz yaşam, yaş faktörü, genetik, sigara-alkol tüketimi, obezite ve radyasyon maruziyeti gibi durumlar yer alır. Bu durumların her biri meme kanseri riskini artıran unsurlar arasında yer alır. Özellikle ailesinde meme kanseri geçirmiş olan bir kişi, yaş olarak 40’ın da üstündeyse düzenli olarak kontrole gitmelidir. Meme tümörlerinin iyi huylu veya kötü huylu olarak doğru bir şekilde sınıflandırılması, uygun eylem planını belirlemek için önemlidir. Tıbbi görüntüleme ve genetik profillemeden toplananlar gibi yüksek boyutlu veri setlerinin ortaya çıkması, bu sınıflandırma görevine yardımcı olmak için gelişmiş makine öğrenimi yöntemlerinin araştırılmasına yol açmıştır. Tümör özelliklerinin çeşitli yönlerini kapsayan 30 farklı özellikten oluşan Wisconsin Meme Kanseri veri seti, tahmini modeller geliştirmek ve değerlendirmek için ideal bir test ortamı sağlar.

Göğüs kanseri teşhisinde farklı makine öğrenmesi algoritmaları kullanılmakla birlikte yaygın olarak kümeleme, yapay sinir ağları, destek vektör makineleri, bulanık ve yapay bulanık mantık ve hibrit teknikler kullanılmaktadır. Örneğin, Ravdin ve Clark düşük ve yüksek risk taşıyan göğüs kanseri hastalarının tespitine yönelik bir yapay sinir ağı modeli ortaya koymuşlardır. Mangasarian vd. ise kötü huylu tümörler için tekrarlamayan vakaları, tekrarlayan vakalar için ise tekrarlama zamanlarını tahmin etmeye yönelik doğrusal programlama tabanlı bir sistem geliştirmişlerdir. Ravi ve Zimmermann tümör veri seti üzerinde geliştirdikleri üç fazlı bulanık veri işleme modelinde önce özellik uzayında boyut indirgeme yoluna gitmiş, ardından bulanık kuralları otomatik olarak oluşturup daha az kuralla daha yüksek bir sınıflama gücü elde etmişlerdir. Delen vd. geniş bir göğüs kanseri veri seti üzerinde iki popüler veri madenciliği algoritması olan yapay sinir ağları ve karar ağaçlarını kullanarak tahmin modelleri geliştirmişlerdir. Karar ağaçları ile %93.6, yapay sinir ağı modeli ile %91.2 doğruluk elde etmişlerdir. Polat ve Güneş, En küçük kare destek vektör makinesi (LS-SVM) sınıflama algoritması kullanarak göğüs kanseri verileri üzerinde %98 oranında başarı elde etmişlerdir. Khan vd. geliştirdikleri bulanık karar ağaçları ile göğüs kanseri verilerini sınıflamış ve bağımsız sınıflayıcılara göre daha başarılı olduklarını ortaya koymuşlardır. Chauhan vd. yapay sinir ağlarında parametre ayarlamaları için diferansiyel evrim modeli kullanarak gerçekleştirdikleri sistemi, göğüs kanseri veri seti dâhil üç farklı veri seti ile test ederek, geleneksel yapay sinir ağı modelinden daha başarılı olduğunu ortaya koymuşlardır. Karabatak ve İnce, ilişki kuralları ile yapay sinir ağlarını birleştirerek ürettikleri hibrit model ile göğüs kanseri verilerini sınıflamışlar ve modellerinin %95.6 oranda doğru sınıflama yaptığını ortaya koymuşlardır. Powel vd. Kaliforniya’da yaşayan ve içerisinde yüksek oranda göğüs kanseri olan, hiç doğum yapmamış ve geç doğum yapmış kadınlara ait veriler içeren veri seti ile Breast Cancer Risk Assessment Tool (BCRAT), International Breast Intervention Study (IBIS) ve BRCAPRO göğüs kanseri risk değerlendirme modellerini 5 yıl boyunca yaptıkları uygulamalarla karşılaştırmışlardır ve performanslarını test etmişlerdir.

Bu çalışmada Wisconsin Üniversitesi hastanesinde Dr. William H. Wolberg tarafından toplanan ve araştırmalar için paylaşılan, göğüs kanseri bulgularını içeren 569 örnekten oluşan veri seti kullanılmıştır. Veriler %80 eğitim seti, %20 test seti olacak şekilde rastgele bölünerek beş farklı boosting algoritması ile sınıflanarak test edilmiştir. Adaboosting, GBM, XGBoost, LightGBM, Catboost algoritmalarının test başarıları karşılaştırılmıştır. Modellerin oluşturulması ve test işlemleri Python programlama dili ile gerçekleştirilmiştir.

# Veri Seti

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Tümör Türlerinin Frekanslar

Kullanılan veri seti 569 adet örnek içermektedir. Her bir veri örneği için 30 tanımlayıcı özellik ve bir adet teşhis sınıfı olmak üzere toplam 31 özellik bulunmaktadır. Göğüste görülen kitle resimlerinin sayısallaştırılması ile elde edilen dokunun çapı, şekli, pürüzsüzlüğü, yüzey alanı, sıkışıklığı, içbükeyliği, içbükey nokta sayısı, simetrisi, kenar kırıklığı gibi verilerden oluşmaktadır. 30 adet özelliğin 10 tanesi tümör hücresinin çekirdeği üzerinden direkt olarak ölçümlenmiş, 20 tanesi ise bunlara bağlı olarak hesaplanmış sayısal değerlerdir. Tümörün iyi huylu veya kötü huylu olduğunu belirten B (benign) ve M (malignant) etiketi ile ifade edilen teşhis sınıfı yer almaktadır. 569 verinin sınıf dağılımı ise 357 iyi huylu, 212 kötü huylu şeklindedir. Yüzdelik olarak değerleri e Şekil 2’deki gibidir.

metin, ekran görüntüsü, diyagram, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Yüzdelik Olarak Gösterim

# Veriye Ait Açıklayıcı Görselleştirme

## Korelasyon Matrisi

ekran görüntüsü, renklilik, metin, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Veriye Ait Korelasyon Matrisi

Bu görselleştirme, Wisconsin Meme Kanseri veri setindeki sayısal değişkenler arasındaki ilişkileri anlamak ve analiz etmek amacıyla oluşturulmuştur. Üçgen bir korelasyon matrisi kullanılarak yapılan bu ısı haritası, değişkenler arasındaki doğrusal ilişkiyi görselleştirmek için etkili bir yöntem sunar.

Grafikte, bazı değişken grupları arasında çok güçlü pozitif korelasyonlar olduğu görülmektedir. Ortalama yarıçap (mean Radius) ortalama çevre (mean perimeter) (r=1.00) ve ortalama alan (mean area) ortalama çevre (mean perimeter) (r=0.99) arasındaki ilişki oldukça güçlüdür. Benzer bir şekilde, en kötü yarıçap (worst Radius) en kötü çevre (worst perimeter) ve en kötü alan (worst area) arasındaki yüksek korelasyonlar da tümör boyutlarına dair ölçümlerin uyum içinde olduğunu işaret etmektedir.

Bazı değişkenler arasında düşük korelasyonlar dikkat çekmektedir. Örneğin, ortalama alan (mean area) ile ortalama simetri (mean symmetry) (r=0.15) arasında zayıf bir ilişki görülmektedir. Bu durum, bu değişkenlerin daha bağımsız özellikleri temsil ettiğini düşündürmektedir. Negatif korelasyonlar genellikle nadir olmakla birlikte, ortalama çevre (mean perimeter) ile doku hatası (texture error) arasında r=−0.09 düzeyinde düşük bir negatif ilişki olduğu gözlemlenmiştir.

Değişkenler gruplar halinde değerlendirildiğinde, ortalama ile başlayan değişkenlerin birbirleriyle yüksek korelasyon gösterdiği, hata ile biten değişkenlerin ise genellikle orta düzeyde korelasyonlara sahip olduğu görülmektedir. Ayrıca, en kötü ile başlayan değişkenler arasında güçlü pozitif korelasyonlar dikkat çekmektedir. Örneğin, ortalama çukur noktaları (mean concave points) ile ortalama çukur (mean concavity) (r=0.83) arasındaki yüksek ilişki, tümör yüzeyindeki girinti ve çıkıntı özelliklerinin birbiriyle uyumlu olduğunu göstermektedir. Öte yandan, en kötü fraktal boyut (worst fractal dimension) genellikle diğer değişkenlerle düşük korelasyona sahiptir ve bu da bu ölçümün bağımsız bir özellik taşıyabileceğini düşündürmektedir.

## ScatterPlot

metin, ekran görüntüsü, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Ortalama Doku Pürüzlülüğü ve Tümör Yarıçapı Dağılım Grafiği

Mean radius (ortalama yarıçap) ve mean texture (ortalama doku pürüzlülüğü) arasındaki ilişkiyi incelemek ve bu özelliklerin verideki iki sınıf (Malignant/Kötü Huylu ve Benign/İyi Huylu) üzerindeki ayrışımını göstermektedir.

Malignant (kötü huylu) tümörler genellikle daha yüksek ortalama yarıçapve ortalama doku pürüzlülüğüdeğerlerine sahiptir. Bu durum, kötü huylu tümörlerin boyutlarının ve yüzey doku pürüzlülüğünün daha belirgin olduğu bilgisini destekler. Benign (iyi huylu) tümörler ise genellikle daha düşük ortalama yarıçapve ortalama dokupürüzlülüğü değerlerine sahip olduğu söylenebilir. Sınıflar arasında kısmi bir ayrışma mevcut olsa da özellikler tamamen ayrık değil ve kısmi bir örtüşme söz konusudur.

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Tümörün Ortalama Çevresi ve Ortalama Yarıçapı için Dağılım Grafiği

Malignant (kötü huylu) örnekler, genellikle daha yüksek Ortalama çevre ve ortalama yarıçap değerlerine sahiptir. Benign (iyi huylu) örnekler, daha düşük değerlere yoğunlaşmıştır. Bu, sınıfların ayrışmasını kolaylaştıran bir yapı oluşturur. Özellikle düşük ortalama çevre ve ortalama yarıçap değerlerinde sınıfların kısmen örtüştüğü gözlemlenmektedir. Bu, bazı iyi huylu ve kötü huylu örneklerin benzer ölçümlere sahip olabileceğini göstermektedir. Ortalama çevre ve ortalama yarıçap arasında belirgin bir doğrusal ilişki olduğu gözlemlenmektedir. Bu, daha büyük çevreye sahip tümörlerin aynı zamanda daha geniş bir alana sahip olduğunu göstermektedir.

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Ortalama Çevre ve En Yüksek Alan Ölçümü için Dağılım Grafiği

Grafikte dikkat çeken ilk unsur, iki değişken arasında pozitif bir ilişki bulunmasıdır; ortalama çevre arttıkça en kötü alanda genellikle artmaktadır. Bu ilişki, özellikle kötü huylu sınıfta daha belirgin bir şekilde görülmektedir. İyi huylu sınıfına ait noktalar genellikle ortalama çevre ve düşük en kötü alan değerlerinde kümelenmiş durumdadır. Bu sınıfın verileri -1 ile 1 arasında yoğunlaşmıştır ve oldukça dar bir dağılım göstermektedir. Öte yandan, kötü huylu sınıfına ait noktalar, ortalama çevre ve en kötü alan değişkenlerinin yüksek değerlerinde daha yaygın olup, daha geniş bir aralıkta yer almaktadır. Bu durum, kötü huylu tümörlerin büyüklük açısından daha belirgin varyasyonlar gösterebildiğini ortaya koymaktadır. Aykırı değerlere bakıldığında, kötü huylu sınıfında ortalama çevre ve en kötü alan için oldukça yüksek değerlere sahip birkaç nokta dikkat çekmektedir. Bu noktalar, verideki ekstrem vakaları temsil edebilir ve sınıflandırma modellerinde önemli etkiler yaratabilir. Ayrıca, kötü huylu ve iyi huylu sınıfları arasında bazı noktaların örtüşmesi, bu iki değişkenin yalnız başına kesin bir ayrım için yetersiz olabileceğini işaret etmektedir.

metin, ekran görüntüsü, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Ortalama Alan ile Ölçülen Çevrede Ortaya Çıkan Hata için Dağılım Grafiği

Benign (iyi huylu) örnekler, çevre hatası ve ortalama alan değerlerinde daha düşük aralıklarda kümelenmektedir. Genellikle düşük "perimeter error" (0-2) ve düşük "mean area" (-1 ile 1 arasında) değerlerinde yoğunlaştığı görülmektedir. Malignant (kötü huylu) örnekler, bu aralıkların dışına çıkarak daha geniş bir dağılım göstermektedir. Malignant örnekler, genelde daha yüksek çevre hatası ve ortalama alan değerlerine sahiptir. Bu durum, kötü huylu örneklerin, iyi huylu örneklere kıyasla daha büyük ve düzensiz alanlara sahip olduğunu gösterebilir. Çevre hatası (perimeter error) değişkeni için 10’un üzerinde yer alan birkaç veri noktası dikkati çekiyor. Bu noktalar, veri kümesindeki genelden oldukça farklı özellikler taşıyan örnekleri temsil ediyor olabilmektedir. 4 ve üzerinde yer alan birkaç veri noktası, veri kümesindeki genel dağılımdan oldukça farklıdır.

## Violin Plot

diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Ortalama Düzgünlük için Tümör Çeşidine Göre Violin Plot

Ortalama düzgünlük için kötü huylu tümörler (malignant) için dağılım özellikle pozitif değerlere doğru daha geniştir, bu kötü huylu tümörlerde ortalama düzgünlük özelliğinin daha yüksek değerlere çıkabileceğini gösterir. İyi huylu tümörler (benign) için dağılım daha simetriktir ve değerler kötü huylu tümörlere göre daha dar bir aralıkta yoğunlaşmıştır. Her iki sınıf arasında ortalama düzgünlük değerlerinde bir miktar ayrışma olsa da grafikte iki sınıfın dağılımlarının önemli ölçüde örtüştüğü bölgeler bulunmaktadır. Bu, ortalama düzgünlük özelliğinin tek başına kullanıldığında tanı koyma açısından sınırlı bir ayrım gücüne sahip olabileceğini düşündürür.

## Pair Plot

Bu grafik, veri setinin "mean radius" (ortalama yarıçap), "worst radius" (en uzun yarıçap) ve "Radius error" (yarıçap varyasyonu) adlı üç önemli özelliğinin birbirleriyle olan ilişkisini ve bu özelliklerin tanıya ("Diagnosis") göre nasıl farklılaştığını analiz etmek için oluşturulmuş bir pair plot grafiğidir.

metin, diyagram, harita içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Ortalama, En Kötü ve Yarıçap Varyasyonu için Pair Plot

Mean radius ve worst radius arasında çok güçlü bir pozitif korelasyon bulunmaktadır. Bu, bir tümörün ortalama çapı arttığında en büyük çapının da arttığını göstermektedir. Malignant (mavi) sınıfındaki noktalar, büyük mean radius ve worst radius değerlerinde yoğunlaşmıştır. Bu, kötü huylu tümörlerin genellikle daha büyük çaplara sahip olduğunu göstermektedir. Benign (turuncu) sınıfındaki noktalar, bu eksenlerde daha düşük değerlere sahip ve kötü huylu sınıftan ayrışmaktadır.

mean radius ile radius error arasında belirgin bir ilişki var, ancak korelasyon düzeyi daha düşüktür. Kötü huylu sınıfta (malignant), bu iki değişken daha geniş bir aralığa yayılarak daha değişken bir yapı sergilemektedir. İyi huylu sınıfta (benign), bu iki değişken daha dar bir aralıkta yoğunlaşmaktadır.

Kötü huylu (mavi) sınıfın dağılımı, iyi huylu (turuncu) sınıfa göre daha sağa kaymış (daha büyük değerlere sahip). Bu durum, mean radius özelliğinin sınıflar arasında güçlü bir ayırt edici olduğunu göstermektedir. worst radius için de benzer bir durum geçerli. Kötü huylu sınıf, genelde daha büyük değerlere sahip. İyi huylu sınıf bu eksende düşük değerlerde toplanmış. Radius error için de aynı yorum yapılabilir.

# Veri Düzenleme

Bu kısımda, Wisconsin Meme Kanseri veri seti kullanılarak yapılan veri analizi ve ön işleme adımlarını kapsamaktadır. Veri seti, meme kanseri teşhisi için kullanılan çeşitli özellikler içermekte ve sınıflandırma problemi için uygundur. Bu raporda, veri seti üzerinde RobustScaler yöntemi uygulanarak özelliklerin ölçeklendirilmesi ele alınmıştır. Ayrıca, veri setinin genel yapısı incelenmiştir. load\_breast\_cancer fonksiyonu kullanılarak scikit-learn kütüphanesinden yüklenmiştir. Bu veri seti, iki sınıfa (Malignant: kötü huylu, Benign: iyi huylu) ayrılmış meme kanseri verilerini içerir.

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Hedef Değişkenin Düzenlenmesi

Hedef sütun (Diagnosis) daha anlamlı hale getirilmiş ve sınıflar 1: Benign (iyi huylu), 0: Malignant (kötü huylu) olacak şekilde ayarlanmıştır.

metin, yazı tipi, ekran görüntüsü, cebir içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Ölçeklendirme İşlemi

Özelliklerin farklı ölçeklerde olması, makine öğrenimi modellerinde ağırlıklandırma sorunlarına yol açabilir. Bu nedenle, veri setindeki sayısal sütunlara ölçeklendirme uygulanmıştır. RobustScaler, verilerin medyanını ve çeyrekler arası açıklığını dikkate alarak ölçeklendirme yapılmıştır. Bu yöntem ile uç değerlerden (outlier) daha az etkilenmiştir ve verinin daha dengeli bir dağılıma sahip olmasını sağlamıştır.

# Karar Ağaçları Modeli

Karar ağacı, belirli bir soruna yönelik tüm potansiyel çözümleri haritalandıran akış şeması benzeri bir diyagramdır. Genellikle kuruluşlar tarafından, bir dizi karar almanın tüm olası sonuçlarını karşılaştırarak en uygun hareket tarzını belirlemeye yardımcı olmak için kullanılır.

Karar ağacı, her bir iç düğümün bir özelliği (veya niteliği), her bir dalın bir karar kuralını ve her bir yaprak düğümünün sonucu temsil ettiği akış şeması benzeri bir ağaç yapısıdır. Bir karar ağacındaki en üst düğüm kök düğüm olarak bilinir ve herhangi bir alt düğümü olmayan düğümler yaprak düğüm olarak adlandırılır. Karar ağacı algoritması kök düğümden başlar ve bir yaprak düğüme ulaşana kadar girdi özellik değerlerine göre bir karar vererek ağaç boyunca ilerler. Yaprak düğümdeki değer, tahmin edilen çıktı değerini temsil eder.

Karar ağaçları, verilerdeki kalıpları belirlemeye yardımcı olan bir tür makine öğrenimi modelidir. Bir dizi girdi değeri alarak ve ardından ağacın mevcut veriler için en iyi kararın ne olduğuna inandığına bağlı olarak bunları farklı dallara ayırarak çalışır. Ağacın verdiği her karar, mümkün olan en iyi seçeneğin belirlenmesinde bir adım olarak düşünülebilir. Bir karar ağacının nihai sonucu genellikle her bir girdi değerinin (önceden belirlenmiş) çıktı değerlerinden birine karşılık gelme olasılığını yansıtan belirli bir çıktı değerleri kümesidir. Karar ağaçlarının diğer modellere göre sahip olduğu önemli bir avantaj, karmaşık veri setleriyle hızlı bir şekilde başa çıkabilmeleridir. Karar ağaçlarının bir diğer büyük avantajı da çok yönlü olmalarıdır.

## Boosting Algoritmalar ile İlişkisi

Boosting algoritmaları, zayıf öğrenicilerden (genellikle karar ağaçları) güçlü bir model oluşturmak için kullanılan bir toplama (ensemble) yöntemidir. Boosting algoritmalarının temel amacı, modelin hatalarını minimize ederek daha güçlü ve doğru bir tahmin yapabilen bir model oluşturmaktır. Karar ağaçları, özellikle düşük derinlikli (genellikle 1-3 katmanlı) ağaçlar, zayıf öğrenici olarak kullanılır. Bu tür karar ağaçları, sınıflandırma ya da regresyon görevlerinde çok güçlü modeller oluşturmazlar, ancak Boosting algoritmalarının iteratif yapısı sayesinde hatalarını düzeltmek için birbirleriyle birleştirildiklerinde çok güçlü bir model ortaya çıkar.

Boosting algoritmalarında karar ağaçları, temel öğrenici olarak sıklıkla tercih edilir:

**Hızlı Öğrenme:** Karar ağaçları, veriyi hızlı bir şekilde bölüp sınıflandırma yapabilen basit modellere sahiptir. Bu, boosting algoritmalarının çok sayıda modelin hatalarını ardışık olarak düzeltme amacına uygundur.

**Esneklik:** Karar ağaçları hem sayısal hem de kategorik verilerle çalışabilir ve özellikler arasındaki doğrusal olmayan ilişkileri de öğrenebilir. Bu esneklik, farklı veri türleriyle çalışmak için büyük avantaj sağlar.

**Aşırı Uyum Kontrolü:** Boosting algoritmalarında kullanılan karar ağaçları genellikle derinlikleri sınırlanmış ağaçlardır (örneğin, derinliği 1 veya 2 olan ağaçlar kullanılır). Bu, aşırı uyum (overfitting) riskini azaltır ve modellerin genelleme kabiliyetini artırır.

## Karar Ağacı Sonuçları

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Karar Ağacı Modeli Oluşturma

İlk adım olarak X ve Y değişkenleri belirlenmiştir. Kötü huylu (malignant) 0 olarak sayısal değere dönüştürülürken iyi huylu (benign) 1 olarak elde edilir. Test boyutu %20 olarak seçilmiştir. stratify=y, bu parametre, hedef değişkenin (y) dağılımının eğitim ve test setlerinde eşit olmasını sağlar. Model oluşturulup eğitilir ve gerekli doğruluk gibi değerler elde edilmesi sağlanır.

Tablo . Sınıflandırma Raporu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Tablo 1’e göre kesinlik değeri kötü huylu (malignant) sınıfı için tahmin ettiği örneklerin %85'inin gerçekten 'Malignant' olduğunu gösterir. Yani model, çoğunlukla doğru şekilde kötü huylu sınıfı tahmin etmektedir. Duyarlılık değeri gerçek 'Malignant' örneklerinin %93'ünün model tarafından doğru bir şekilde tahmin edildiği anlamına gelir. Yani model, kötü huylu sınıfı tespit etmede oldukça başarılıdır. Benign sınıfı içinse kesinlik değeri tahmin ettiği örneklerin %96'sı gerçekten 'Benign' (iyi huylu) olduğu anlamına gelir. Gerçek 'Benign' örneklerinin %90'ı doğru bir şekilde tahmin edilmiştir. Yani model, kötü huylu tümörü tespit etmekte daha güçlüdür. Genel model için yorum yapıldığında ise modelin tüm test setinde doğru sınıflandırdığı örneklerin oranıdır. %91.23 doğruluk, genel olarak modelin güçlü olduğunu ve çoğu örneği doğru tahmin ettiğini gösterir. %91 gayet başarılı bir sonuçtur fakat yapılacak boosting algoritmaları ile burada elde edilen sonucu ne kadar etkilediklerini incelemesi yapılacaktır.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Karışıklık Matrisi

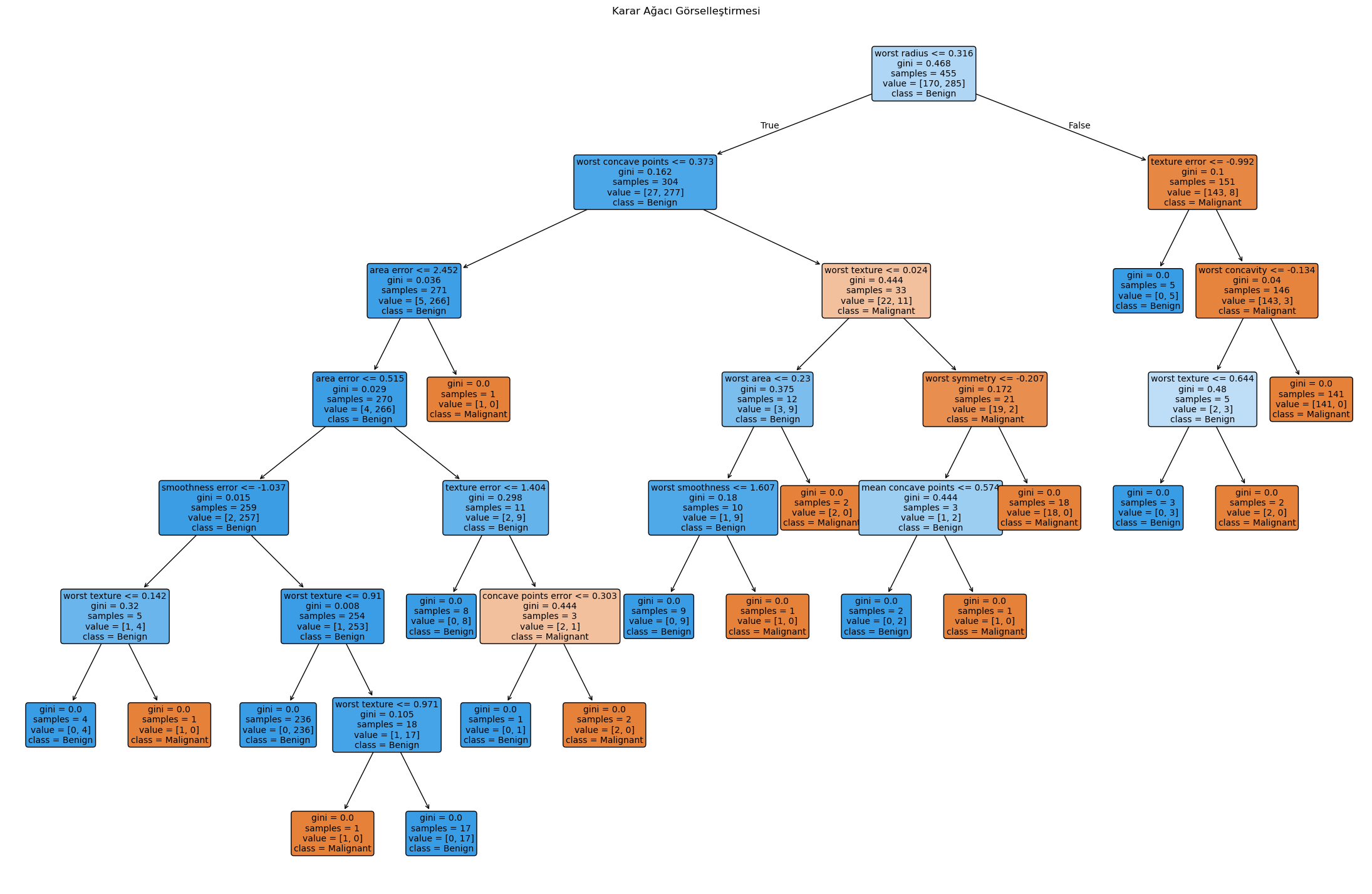
Gerçekten 'Malignant' olan ve modelin doğru şekilde 'Malignant' olarak tahmin ettiği 39 örnek, gerçekten 'Benign' olan ve modelin doğru şekilde 'Benign' olarak tahmin ettiği 65 örnek bulunurken gerçekten 'Benign' olan, ancak modelin yanlış bir şekilde 'Malignant' olarak tahmin ettiği 7 örnek ve gerçekten 'Malignant' olan, ancak modelin yanlış bir şekilde 'Benign' olarak tahmin ettiği 3 örnek bulunmaktadır.

metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

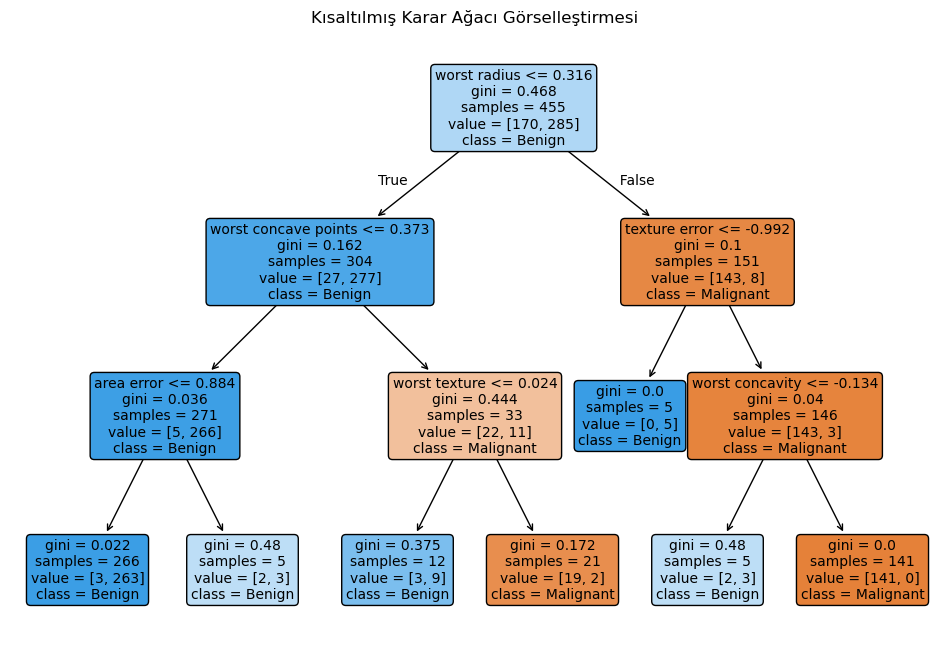
Şekil . Model için ROC Eğrileri

Şekil 6’daki grafik bir ROC grafiğidir ve sınıflandırma modelinin performansını değerlendirmek için kullanılmıştır. Eğri, modelin pozitif sınıfları doğru şekilde tahmin etme oranını (True Positive Rate) yanlış pozitif oranına (False Positive Rate) karşı gösterir. Doğruluk değeri ROC eğrisinin altındaki alan (AUC), modelin performansını değerlendirir. Bu durumda, AUC = 0.92, modelin oldukça iyi bir performansa sahip olduğunu gösterir. Eğer eğri (1,1) noktasına ne kadar yakınsa o kadar iyi model olduğunu gösterir. Eğitim setinde model mükemmel bir doğruluk sergiliyor. Bu genellikle modelin eğitim verisine çok fazla uyum sağladığını ve genel performans açısından dikkatli olunması gerektiğini işaret edebilir.



Şekil . Model Sonucu Ortaya Çıkan Ağaç Modeli

Ayrıntılı olarak her değişken için bir veri elde edilse de daha kısaltılmış hali yorumlanacaktır. Sadece en son dalda her iki düğümün de her türlü tümörün rengini içerdiğinden (turuncu: kötü huylu, mavi: iyi huylu) doğru sonuçlar elde edildiği söylenebilir.



Şekil . Kısaltılmış Ağaç Modeli

Başlangıç noktası yani kök düğüm olarak seçilen en kötü yarıçaptır. Veriye en uygun olarak bu terim en uzun yarıçap olarak değiştirilebilir. Bu değişkeni seçme sebebi veri setinin en iyi ayrıştırılmasını sağlayan özelliği olduğunu tespit etmesindendir. Bu düğüme göre eğer bu değişken 0.316 küçük ya da eşitse gini katsayısı 0.468 olarak hesaplanmıştır. Katsayı 0’a ne kadar yakınsa düğümdeki sınıflar o kadar homojendir. Düğüm iyi huylu olup olmadığını göstermektedir. Sol dala geçilmesi şart sağlandığını gösterir ve en kötü içbükey noktaların (worst concave points) değeri 0.373’ten küçük olması sonucunda gini 0.162’ye düşerek homojenliğin arttığı görülmüştür. Sınıf sayısı 304’e düşmüş ve 27’si iyi huylu, 277 tanesi kötü huyludur. Alt dallarına bakıldığında tümörün alan varyasyonunun 0.884’ten küçük (area error) kısmında gini katsayısı 0.036 olması çok saf bir düğüm oluşturduğunu gösterir. En yüksek doku pürüzlülüğünün (worst texture) 0.024’ten küçük olduğu durumda gini katsayısı 0.444 olarak gelmiş, homojenlik pek sağlanamadığı görülmektedir. Sağ dala geçiş yapıldığında yani şart sağlanmadığında 1. düğüm tümörün doku pürüzlülüğü varyasyonu (texture error) -0.992’ten küçükse gini katsayısı 0.1 çıktığından kötü huylu sınıfı için yüksek homojenlik görüldüğü söylenebilir. Alt dallara geçildiğinde gini 0 olduğundan o dal tamamen iyi huylu grubudur. En kötü çukurluk değeri (worst concavity) -0.134 küçük olduğunda gini 0.04 gelerek homojenlik sağlanmış yorumu yapılabilir.

Genel bir yorum yapıldığında bu ağacın genel performansı, veri setinde Benign ve Malignant sınıflarını iyi şekilde ayırabildiğini gösteriyor.

## Adaboosting ile Karar Ağacının Desteklenmesi

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Adaboosting ile Desteklenen Model Kurulumu

Bu kod, AdaBoost algoritmasını kullanarak karar ağacı modelini eğitmek ve test etmek için yazılmıştır. AdaBoost, zayıf öğrenicileri (weak learners) birleştirerek güçlü bir model oluşturur. Bu model çıktısı sonucunda elde edilecek değerler ile makine öğrenmesi modeli eğer boosting algoritması ile desteklenirse ne kadar fark olacağını görmek amaçlı yapılmıştır. estimator=base\_estimator\_tree parametresiyle, zayıf öğrenici olarak base\_estimator\_tree yani karar ağacı belirlenmiştir.

Tablo . Sınıflandırma Tablosu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Tablo 2’ye göre kesinlik değeri kötü huylu (malignant) sınıfı için tahmin ettiği örneklerin %97'nin gerçekten 'Malignant' olduğunu gösterir. Yani model, çoğunlukla doğru şekilde kötü huylu sınıfı tahmin etmektedir. Duyarlılık değeri gerçek 'Malignant' örneklerinin %90'nın model tarafından doğru bir şekilde tahmin edildiği anlamına gelir. Yani model, kötü huylu sınıfı tespit etmede oldukça başarılıdır. Benign sınıfı içinse kesinlik değeri tahmin ettiği örneklerin %94'ünün gerçekten 'Benign' (iyi huylu) olduğu anlamına gelir. Gerçek 'Benign' örneklerinin %98'ini doğru bir şekilde tahmin edilmiştir. Yani model, iyi huylu tümörü tespit etmekte daha güçlüdür. Genel model için yorum yapıldığında ise modelin tüm test setinde doğru sınıflandırdığı örneklerin oranıdır. %95.61 doğruluk, genel olarak modelin güçlü olduğunu ve çoğu örneği doğru tahmin ettiğini gösterir. Sonuç olarak modelin boosting ile desteklenmesi olumlu bir katkı sağlamıştır.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Karışıklık Matrisi

Gerçekten 'Malignant' olan ve modelin doğru şekilde 'Malignant' olarak tahmin ettiği 38 örnek, gerçekten 'Benign' olan ve modelin doğru şekilde 'Benign' olarak tahmin ettiği 71 örnek bulunurken gerçekten 'Benign' olan, ancak modelin yanlış bir şekilde 'Malignant' olarak tahmin ettiği 1 örnek ve gerçekten 'Malignant' olan, ancak modelin yanlış bir şekilde 'Benign' olarak tahmin ettiği 4 örnek bulunmaktadır. Karışıklık matrisinden de görüleceği üzere hata yapılma oranı azalmıştır.

metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . ROC Eğrisi

Eğitim performansı, Mavi çizgi, modelin eğitim setinde mükemmel sonuç verdiğini gösteriyor. AUC'nin 1.00 olması, aşırı öğrenme (overfitting) potansiyelini düşündürse de test setinde yüksek performans sergilendiğinden bu durum kontrol altında görünüyor. Test performansı, Kırmızı çizginin yüksek AUC değeri (0.98), modelin genelleme yeteneğinin oldukça iyi olduğunu gösterir. Overfitting, Eğitim ve test AUC değerleri arasında küçük bir fark vardır (1.00 ve 0.98), bu da aşırı öğrenmenin düşük düzeyde olduğunu düşündürmektedir.

# Farklı Boosting Algoritmaları Kullanımı

## Adaboosting Nedir?

AdaBoost algoritması, zayıf sınıflandırıcıların bir araya gelmesiyle ortaya çıkan güçlü bir sınıflandırıcıyı temsil eden topluluk sınıflandırıcısı olarak isimlendirilmektedir. Modelin genel çalışma mantığı, her aşamada bir önceki aşamanın sonucunda yapılan yanlış tahminlerin ağırlığını artırarak sınıflandırıcının tekrar çalıştırılması ile başlamaktadır. Yapılan bu işlemlerle yanlış yapılan tahminlere odaklanıp, oluşturulan modelin sınıflandırmadaki doğruluk oranını yükseltmek amaçlanmaktadır. AdaBoost, 1995 yılında Yoav Freund ve Robert Schapire tarafından geliştirilmiştir. Temel sınıflandırıcıların ağırlıklı bir oylama sistemi kullanarak bir araya getirilmesi prensibine dayanır.

## Adaboosting Uygulaması

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Model Kurulumu

Verinin %20'si test seti, %80'i eğitim seti olarak ayrılır. Model, her iterasyonda zayıf bir öğrenici eğitir ve bu öğrenicilerin ağırlıklarını hatalarına göre ayarlar. Elde edilecek sonuçlarla adaboosting kendi kullanıldığında mı daha iyi yoksa base estimator karar ağacı olarak alındığında daha iyi o karşılaştırma yapılabilecektir.

Tablo . Sınıflandırma Tablosu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Tablo 3’e göre kesinlik değeri kötü huylu (malignant) sınıfı için tahmin ettiği örneklerin %97'nin gerçekten 'Malignant' olduğunu gösterir. Yani model, çoğunlukla doğru şekilde kötü huylu sınıfı tahmin etmektedir. Duyarlılık değeri gerçek 'Malignant' örneklerinin %95'nin model tarafından doğru bir şekilde tahmin edildiği anlamına gelir. Yani model, kötü huylu sınıfı tespit etmede oldukça başarılıdır. Benign sınıfı içinse kesinlik değeri tahmin ettiği örneklerin %97’sinin gerçekten 'Benign' (iyi huylu) olduğu anlamına gelir. Gerçek 'Benign' örneklerinin %98'ini doğru bir şekilde tahmin edilmiştir. Yani model, iyi huylu tümörü tespit etmekte daha güçlüdür. Genel model için yorum yapıldığında ise modelin tüm test setinde doğru sınıflandırdığı örneklerin oranıdır. %97.37 doğruluk, genel olarak modelin güçlü olduğunu ve çoğu örneği doğru tahmin ettiğini gösterir.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Karışıklık Matrisi

Gerçekten 'Malignant' olan ve modelin doğru şekilde 'Malignant' olarak tahmin ettiği 41 örnek, gerçekten 'Benign' olan ve modelin doğru şekilde 'Benign' olarak tahmin ettiği 70 örnek bulunurken gerçekten 'Benign' olan, ancak modelin yanlış bir şekilde 'Malignant' olarak tahmin ettiği 1 örnek ve gerçekten 'Malignant' olan, ancak modelin yanlış bir şekilde 'Benign' olarak tahmin ettiği 2 örnek bulunmaktadır. Karışıklık matrisinden de görüleceği üzere hata yapılma oranı diğerlerine göre çok azdır.

metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . ROC Eğrisi

Eğitim ve test eğrilerinin AUC değerlerinin ikisininde 1 çıkması modelin eğitim verisinde mükemmel bir performans gösterdiği anlamına gelir. Eğitim ve test eğrilerinin aynı derecede mükemmel performans göstermesi, modelin eğitim setine fazla uyum sağlamadığını ve **overfitting** yapmadığını düşündürür. Ancak, bu kadar yüksek bir performans gerçek veri setlerinde genellikle nadirdir. Özellikle AUC'nin 1 olması, veri setinde belirgin bir örüntü olduğu ya da modelin fazla sade bir veri seti ile eğitildiği anlamına gelebilir. Model çok basit bir örüntüyü ezberlemiş olabilir yorumu yapılabilir.

## Gradient Boosting Nedir?

Gradyan artırma, regresyon ve sınıflandırma problemleri için bir makine öğrenmesi tekniğidir. Bu, zayıf tahmin modellerinin bir araya gelmesiyle tipik olarak karar ağaçlarının oluşturduğu bir model oluşturur. Denetlenen herhangi bir öğrenme algoritmasının amacı, bir kayıp fonksiyonu tanımlamak ve en aza indirmektir. Gradient Boosting’de öncelikli olarak ilk yaprak (initial leaf) oluşturulur. Sonrasında tahmin hataları göz önüne alınarak yeni ağaçlar oluşturulur. Bu durum karar verilen ağaç sayısına ya da modelden daha fazla gelişme kaydedilemeyinceye kadar devam eder.

## Gradient Boosting Uygulaması

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . GBM Modeli Kurulumu

Bir ensemble öğrenme yöntemidir ve zayıf öğrenicileri (örneğin, karar ağaçları) ardışık olarak birleştirerek güçlü bir sınıflandırıcı oluşturur. 50 adet karar ağacı oluşturacaktır. Bu, modelin karmaşıklığını ve öğrenme kapasitesini etkiler.

Tablo . Sınıflandırma Raporu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Tablo 4’e göre kesinlik değeri kötü huylu (malignant) sınıfı için tahmin ettiği örneklerin %95'nin gerçekten 'Malignant' olduğunu gösterir. Yani model, çoğunlukla doğru şekilde kötü huylu sınıfı tahmin etmektedir. Duyarlılık değeri gerçek 'Malignant' örneklerinin %93'ünü model tarafından doğru bir şekilde tahmin edildiği anlamına gelir. Yani model, kötü huylu sınıfı tespit etmede oldukça başarılıdır. Benign sınıfı içinse kesinlik değeri tahmin ettiği örneklerin %95.8’inin gerçekten 'Benign' (iyi huylu) olduğu anlamına gelir. Gerçek 'Benign' örneklerinin %97'sini doğru bir şekilde tahmin edilmiştir. Yani model, iyi huylu tümörü tespit etmekte daha güçlüdür. Genel model için yorum yapıldığında ise modelin tüm test setinde doğru sınıflandırdığı örneklerin oranıdır. %95.61 doğruluk, genel olarak modelin güçlü olduğunu ve çoğu örneği doğru tahmin ettiğini gösterir. Buna göre GBM modeli adaboosting’e göre daha etkisiz kalmıştır.

metin, ekran görüntüsü, dikdörtgen, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Karışıklık Matrisi

Gerçekten 'Malignant' olan ve modelin doğru şekilde 'Malignant' olarak tahmin ettiği 40 örnek, gerçekten 'Benign' olan ve modelin doğru şekilde 'Benign' olarak tahmin ettiği 69 örnek bulunurken gerçekten 'Benign' olan, ancak modelin yanlış bir şekilde 'Malignant' olarak tahmin ettiği 2 örnek ve gerçekten 'Malignant' olan, ancak modelin yanlış bir şekilde 'Benign' olarak tahmin ettiği 3 örnek bulunmaktadır. Diğer modellere oldukça yakın sonuç vermiştir.

metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . ROC Eğrisi

ROC eğrisinin adaboosting uygulanırken elde edilen eğriyle aynı sonuç vermiştir. Eğitim ve test AUC değerlerinin 1 çıkması ya modelin mükemmel bir şekilde çalıştığına ya da veri setinde belirgin bir örüntü olduğu, modelin fazla sade bir veri seti ile eğitildiği anlamına gelebilir. Sağlıklı bir sonuç değildir.

## XGBoost Nedir?

XGBoost(eXtreme Gradient Boosting), Gradient Boosting algoritmasının çeşitli düzenlemeler ile optimize edilmiş yüksek performanslı halidir. Tianqi Chen ve Carlos Guestrin’in 2016 yılında yayınladıkları “XGBoost: A Scalable Tree Boosting System” adlı makale ile hayatımıza dahil olmuştur. Algortimanın en önemli özellikleri yüksek tahmin gücü elde edebilmesi, aşırı öğrenmenin önüne geçebilmesi, boş verileri yönetebilmesi ve bunları hızlı yapabilmesidir. Tianqi’ye göre XGBoost diğer popüler algoritmalardan 10 kat daha hızlı çalışmaktadır. Daha az kaynak kullanarak üstün sonuçlar elde etmek için yazılım ve donanım optimizasyon tekniklerini uygulanmıştır. Karar ağacı tabanlı algoritmaların en iyisi olarak gösterilir. Gradient Boosting ve XGBoost aynı prensiple çalışmaktadır. Aralarındaki farklar detaylardadır. XGBoost, farklı teknikler kullanarak daha yüksek tahmin başarısı gösterir ve büyük veri setlerinde çalışmak için optimize edilmiştir.

## XGBoost Uygulaması

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Model Kurulumu

XGBClassifier: XGBoost'un sınıflandırma problemleri için kullandığı algoritmadır. n\_estimators: Karar ağaçlarının sayısını belirtir (burada 50 ağaç kullanılacak). En sonda da gerekli olan kesinlik değerleri gibi sayısal veriler elde edilecektir.

Tablo . XGBoost için Sınıflandırma Raporu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Tablo 5’e göre kesinlik değeri kötü huylu (malignant) sınıfı için tahmin ettiği örneklerin %95'nin gerçekten 'Malignant' olduğunu gösterir. Yani model, çoğunlukla doğru şekilde kötü huylu sınıfı tahmin etmektedir. Duyarlılık değeri gerçek 'Malignant' örneklerinin %90'ını model tarafından doğru bir şekilde tahmin edildiği anlamına gelir. Yani model, kötü huylu sınıfı tespit etmede oldukça başarılıdır. Benign sınıfı içinse kesinlik değeri tahmin ettiği örneklerin %94.5’inin gerçekten 'Benign' (iyi huylu) olduğu anlamına gelir. Gerçek 'Benign' örneklerinin %97'sini doğru bir şekilde tahmin edilmiştir. Yani model, iyi huylu tümörü tespit etmekte daha güçlüdür. Genel model için yorum yapıldığında ise modelin tüm test setinde doğru sınıflandırdığı örneklerin oranıdır. %94.73 doğruluk, genel olarak modelin güçlü olduğunu ve çoğu örneği doğru tahmin ettiğini gösterir. Diğer modellere göre model doğruluğu azdır.

metin, ekran görüntüsü, dikdörtgen, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Karışıklık Matrisi

Gerçekten 'Malignant' olan ve modelin doğru şekilde 'Malignant' olarak tahmin ettiği 38 örnek, gerçekten 'Benign' olan ve modelin doğru şekilde 'Benign' olarak tahmin ettiği 70 örnek bulunurken gerçekten 'Benign' olan, ancak modelin yanlış bir şekilde 'Malignant' olarak tahmin ettiği 2 örnek ve gerçekten 'Malignant' olan, ancak modelin yanlış bir şekilde 'Benign' olarak tahmin ettiği 4 örnek bulunmaktadır. Diğer modellere oldukça yakın sonuç vermiştir.

metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . ROC Eğrisi

Eğitim ROC Eğrisi (Mavi): Eğitim verisi üzerinde modelin performansını gösterir. AUC (Area Under Curve) değeri 1.00, yani mükemmel bir performans göstermiştir. Test ROC Eğrisi (Kırmızı): Test verisi üzerinde modelin performansını gösterir. AUC değeri 0.99, yani model test verisinde de neredeyse mükemmel performans göstermiştir. Bu grafik, AUC değerinin hem eğitim hem de test verisi için çok yüksek olduğunu (%100 ve %99) göstermektedir. Bu durum, modelin hem eğitim verisine çok iyi uyum sağladığını hem de test verisinde genelleme kabiliyetinin yüksek olduğunu belirtir. Test ROC eğrisinin yüksek olması, modelin overfitting (aşırı öğrenme) yapmadığını ve genelleme gücünün iyi olduğunu göstermektedir.

## LightGBM Nedir?

LightGBM, Microsoft DMTK (Distributed Machine Learning Toolkit) projesi kapsamında 2017 yılında geliştirilmiş bir boosting algoritmasıdır. Diğer boosting algoritmaları ile karşılaştırıldığında yüksek işlem hızı, büyük verileri işleyebilmesi, daha az kaynak(RAM) kullanımı, yüksek tahmin oranı, paralel öğrenme ve GPU öğrenimini desteklemesi gibi avantajları vardır. Modelin tanıtıldığı “LightGBM: A Highly Efﬁcient Gradient Boosting Decision Tree” makalesine göre, yapılan çalışmalarda LightGBM’in diğer modellere göre 20 kat daha hızlı olduğu sonucuna ulaşılmıştır. LightGBM, histogram tabanlı çalışan bir algoritmadır. Sürekli değere sahip olan değişkenleri kesikli(discrete bin) hale getirerek hesaplama maliyetini azaltır. Karar ağaçlarının eğitim süresi yapılan hesaplama ve dolayısıyla bölünme sayısı ile doğru orantılıdır. Bu yöntem sayesinde hem eğitim süresi kısalmakta hem de kaynak kullanımı düşmektedir.

## LightGBM Uygulaması

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Model Kurulumu

Bu kod, LightGBM (Light Gradient Boosting Machine) algoritmasını kullanarak bir sınıflandırma modeli oluşturmak ve değerlendirmek için yazılmıştır. Veri seti, eğitim (%80) ve test (%20) olarak ayrılmıştır. stratify=y parametresi, hedef değişkenin sınıf dağılımının hem eğitim hem de test setinde korunmasını sağlamıştır. Bu, özellikle dengesiz veri setlerinde sınıfların orantısının korunmasını sağlamak için önemlidir.

Tablo . Sınıflandırma Raporu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Tablo 6’ya göre kesinlik değeri kötü huylu (malignant) sınıfı için tahmin ettiği örneklerin %95'inin gerçekten 'Malignant' olduğunu gösterir. Yani model, çoğunlukla doğru şekilde kötü huylu sınıfı tahmin etmektedir. Duyarlılık değeri gerçek 'Malignant' örneklerinin %92.8'nin model tarafından doğru bir şekilde tahmin edildiği anlamına gelir. Yani model, kötü huylu sınıfı tespit etmede oldukça başarılıdır. Benign sınıfı içinse kesinlik değeri tahmin ettiği örneklerin %95'inin gerçekten 'Benign' (iyi huylu) olduğu anlamına gelir. Gerçek 'Benign' örneklerinin %97'sini doğru bir şekilde tahmin edilmiştir. Yani model, iyi huylu tümörü tespit etmekte daha güçlüdür. Genel model için yorum yapıldığında ise modelin tüm test setinde doğru sınıflandırdığı örneklerin oranıdır. %95.61 doğruluk, genel olarak modelin güçlü olduğunu ve çoğu örneği doğru tahmin ettiğini gösterir. Sonuç olarak modelin LightGBM ile kurulması iyi sonuç elde edilmesini sağlamıştır.

metin, ekran görüntüsü, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Karışıklık Matrisi

Gerçekten 'Malignant' olan ve modelin doğru şekilde 'Malignant' olarak tahmin ettiği 39 örnek, gerçekten 'Benign' olan ve modelin doğru şekilde 'Benign' olarak tahmin ettiği 70 örnek bulunurken gerçekten 'Benign' olan, ancak modelin yanlış bir şekilde 'Malignant' olarak tahmin ettiği 2 örnek ve gerçekten 'Malignant' olan, ancak modelin yanlış bir şekilde 'Benign' olarak tahmin ettiği 3 örnek bulunmaktadır. Karışıklık matrisinden de görüleceği üzere hata yapılma oranı azalmıştır.

metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . ROC Eğrisi

Eğitim performansı, mavi çizgi, modelin eğitim setinde mükemmel sonuç verdiğini göstermektedir. AUC'nin 1 olması, aşırı öğrenme (overfitting) potansiyelini düşündürse de test setinde yüksek performans sergilendiğinden bu durum kontrol altında görünüyor. Test performansı, kırmızı çizginin yüksek AUC değeri (0.99), modelin genelleme yeteneğinin oldukça iyi olduğunu gösterir. Overfitting, Eğitim ve test AUC değerleri arasında küçük bir fark vardır (1.00 ve 0.99), bu da aşırı öğrenmenin düşük düzeyde olduğunu düşündürmektedir.

## Catboosting Nedir?

Catboost, Yandex şirketi tarafından geliştirilmiş olan Gradient Boosting tabanlı açık kaynak kodlu bir makine öğrenmesi algoritmasıdır. Gradient Boosting’in performansını arttırmak amacıyla geliştirilen XGBoost ve LightGBM’e alternatif olarak Nisan 2017 tarihinde “CatBoost: unbiased boosting with categorical features” makalesiyle tanıtılmıştır. Adı “Category” ve “Boosting” kelimelerinin birleşiminden gelmektedir. Yüksek öğrenme hızı hem sayısal hem kategorik hem de metin verileri ile çalışılabilmesi, GPU desteği ve görselleştirme seçenekleri sunması diğer algoritmalardan en çok ayrılan özellikleridir. CatBoost, veri hazırlığı evresini kısaltması ile önemli bir konuma sahiptir. Boş veriler ile başa çıkabilir, kategorik verilere kodlama(encoding) uygular. Kategorik veriler ile yüksek performanslı çalışabilmesinin nedeni kendine has bir kodlama metoduna sahip olmasıdır.

## Catboosting Uygulaması

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Model Kurulumu

Bu kod, CatBoost algoritmasını kullanarak bir sınıflandırma modeli oluşturur ve modelin performansını değerlendirir. CatBoost, özellikle kategorik verilerle iyi performans göstermesi ve otomatik optimizasyon kabiliyeti ile bilinir. CatBoostClassifier sınıfı ile bir CatBoost sınıflandırıcı modeli oluşturulur. Model, eğitim verisi (X\_train) ve hedef değişken (y\_train) kullanılarak eğitilir. accuracy\_score(y\_test, y\_pred\_catboost\_karar) fonksiyonu, modelin doğruluk oranını hesaplar. Bu oran, modelin test verilerindeki doğru tahminlerinin yüzdesini ifade eder.

Tablo . Sınıflandırma Raporu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Tabloya göre kesinlik değeri kötü huylu (malignant) sınıfı için tahmin ettiği örneklerin %97'inin gerçekten 'Malignant' olduğunu gösterir. Yani model, çoğunlukla doğru şekilde kötü huylu sınıfı tahmin etmektedir. Duyarlılık değeri gerçek 'Malignant' örneklerinin %90 'nın model tarafından doğru bir şekilde tahmin edildiği anlamına gelir. Yani model, kötü huylu sınıfı tespit etmede oldukça başarılıdır. Benign sınıfı içinse kesinlik değeri tahmin ettiği örneklerin %94’ünün gerçekten 'Benign' (iyi huylu) olduğu anlamına gelir. Gerçek 'Benign' örneklerinin %98'sini doğru bir şekilde tahmin edilmiştir. Yani model, iyi huylu tümörü tespit etmekte daha güçlüdür. Genel model için yorum yapıldığında ise modelin tüm test setinde doğru sınıflandırdığı örneklerin oranıdır. %95 doğruluk, genel olarak modelin güçlü olduğunu ve çoğu örneği doğru tahmin ettiğini gösterir. Sonuç olarak modelin CatBoost ile desteklenmesi olumlu bir katkı sağlamıştır.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Karışıklık Matrisi

Gerçekten 'Malignant' olan ve modelin doğru şekilde 'Malignant' olarak tahmin ettiği 38 örnek, gerçekten 'Benign' olan ve modelin doğru şekilde 'Benign' olarak tahmin ettiği 71 örnek bulunurken gerçekten 'Benign' olan, ancak modelin yanlış bir şekilde 'Malignant' olarak tahmin ettiği 1 örnek ve gerçekten 'Malignant' olan, ancak modelin yanlış bir şekilde 'Benign' olarak tahmin ettiği 4 örnek bulunmaktadır. Karışıklık matrisinden de görüleceği üzere hata yapılma oranı azalmıştır.

metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . ROC Eğrisi

Eğitim performansı, Mavi çizgi, modelin eğitim setinde mükemmel sonuç verdiğini gösteriyor. AUC'nin 1.00 olması, aşırı öğrenme (overfitting) potansiyelini düşündürse de test setinde yüksek performans sergilendiğinden bu durum kontrol altında görünüyor. Test performansı, Kırmızı çizginin yüksek AUC değeri (1.00), modelin genelleme yeteneğinin oldukça iyi olduğunu gösterir. Overfitting, Eğitim ve test AUC değerleri arasında fark yoktur. Bu da aşırı öğrenmenin olmadığını düşündürmektedir.

# En İyi Model Seçimi

Tablo . Doğruluk Değerleri

|  |  |
| --- | --- |
| **MODEL** | **DOĞRULUK SKORU** |
| ADABOOSTING+KARAR AĞAÇLARI | 95.61 |
| ADABOOSTING | 97.36 |
| GBM | 95.61 |
| XGBOOST | 94.7 |
| LIGHTGBM | 95.61 |
| CATBOOST | 95.61 |

En iyi model seçiminde kullanılacak olan doğruluk skorları Tablo 8’de verilmiştir. Doğruluk skorru yüksek olan model veri seti için en iyisi olandır denebilir. Buna göre tümörün türünü tespit etme de en iyi çalışan model adaboosting olacaktır.

# Kaynakça

Özgür, S. N., & Keser, S. (2021). Meme Kanseri Tümörlerinin Derin Öğrenme Algoritmaları ile Sınıflandırılması. *Türk Doğa ve Fen Dergisi*, 222.

Abu Naser, S., & Anas, A. S. (2023). Predictive Modeling of Breast Cancer Diagnosis Using Neural Networks:A Kaggle Dataset Analysis. *International Journal of Academic Engineering Research (IJAER)*, 1-9.

Agarap, A. (2018, Şubat 2). *On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset*. Aralık 2024 tarihinde ACMDL: https://dl.acm.org/doi/10.1145/3184066.3184080 adresinden alındı

Akel, U. (2020, Nisan 12). *Adaboost Algoritması*. Aralık 2024 tarihinde bilisimkitabi: https://bilisimkitabi.com/adaboost-algoritmasi adresinden alındı

*Breast Cancer Wisconsin (Diagnostic) Dataset*. (2024, Nisan 30). Aralık 2024 tarihinde geeksforgeeks: https://www.geeksforgeeks.org/breast-cancer-wisconsin-diagnostic-dataset/ adresinden alındı

Bulutistan. (2023, Ağustos 18). *Makine Öğrenmesi Karar Ağacı (Decision Tree) Nedir?* Aralık 2024 tarihinde Bulutistan: https://bulutistan.com/blog/makine-ogrenmesi-karar-agaci-decision-tree-nedir/ adresinden alındı

DevHunter. (tarih yok). *Gradyan Artırma(Gradient Boostin)*. Aralık 2024 tarihinde WordPress: https://devhunteryz.wordpress.com/2018/07/11/gradyan-arttirmagradient-boosting/ adresinden alındı

Güzel, K. (2020, Aralık 28). *Boosting Nedir? Adım Adım AdaBoost Algoritması*. Aralık 2024 tarihinde Medium: https://kadirguzel.medium.com/boosting-nedir-adım-adım-adaboost-algoritması-439cce20ab9a adresinden alındı

Karanam, S. (2022, Eylül 30). *Exploratory Data Analysis — Breast Cancer Wisconsin (Diagnostic) Dataset*. Aralık 2024 tarihinde Medium: https://medium.com/@shashmikaranam/exploratory-data-analysis-breast-cancer-wisconsin-diagnostic-dataset-6a3be9525cd adresinden alındı

*LogitBoost*. (tarih yok). Aralık 2024 tarihinde Wikipedia: https://en.wikipedia.org/wiki/LogitBoost adresinden alındı

Memorial Tıbbi Yayın Kurulu. (2024, Mayıs 16). *Meme Kanseri Nedir? Meme Kanseri Belirtileri, Nedenleri ve Tedavisi*. Aralık 2024 tarihinde MEMORIAL: https://www.memorial.com.tr/hastaliklar/meme-kanseri-belirtileri-tanisi-ve-tedavi-yontemleri#meme-kanseri-nedir adresinden alındı

Muratlar, E. R. (2020, Ağustos 6). *CatBoost Nedir? Diğer Boosting Algoritmalarından Farkı Nelerdir?* Aralık 2024 tarihinde VBO: https://www.veribilimiokulu.com/catboost-nedir-diger-boosting-algoritmalarindan-farki-nelerdir/#:~:text=CatBoost%2C%20veri%20hazırlığı%20evresini%20kısaltması,bir%20kodlama%20metoduna%20sahip%20olmasıdır adresinden alındı

Muratlar, E. R. (2020, Ocak 24). *Gradient Boosted Regresyon Ağaçları*. Aralık 2024 tarihinde VBO: https://www.veribilimiokulu.com/gradient-boosted-regresyon-agaclari/ adresinden alındı

Muratlar, E. R. (2020, Nisan 29). *LightGBM*. Aralık 2024 tarihinde VBO: https://www.veribilimiokulu.com/lightgbm/ adresinden alındı

Muratlar, E. R. (2020, Mart 22). *XGBoost Nasıl Çalışır? Neden İyi Performans Gösterir?* Aralık 2024 tarihinde VBO: https://www.veribilimiokulu.com/xgboost-nasil-calisir/ adresinden alındı

Muratlar, E. R. (2020, Mart 22). *XGBoost Nasıl Çalışır? Neden İyi Performans Gösterir?* Aralık 2024 tarihinde VBO: https://www.veribilimiokulu.com/xgboost-nasil-calisir/ adresinden alındı

Sevli, O. (2019, Mayıs 22). Göğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması. *Avrupa Bilim ve Teknoloji Dergisi*, 185.

Şirin, E. (2017, Temmuz 2017). *Python ile Karar Ağacı (Decision Tree with Python)*. Aralık 2024 tarihinde VBO: https://www.veribilimiokulu.com/siniflandirma-notlari-16-karar-agaci-python-uygulama/ adresinden alındı

Yıldırım, E. (2021, Şubat 17). *Temel Topluluk Öğrenimi ve Random Forest, Gradient Boosting Algoritmaları*. Aralık 2024 tarihinde Medium: https://ecehanyildirim.medium.com/temel-topluluk-öğrenimi-ve-random-forest-gradient-boosting-algoritmaları-e105f93b33f0 adresinden alındı

Yılmaz, H., & Kuncan, F. (2022). ANALYSIS OF DIFFERENT MACHINE LEARNING TECHNIQUES WITH PCA IN THE DIAGNOSIS OF BREAST CANCER. *Journal of Engineering Technology and Applied Sciences*, 195-205.