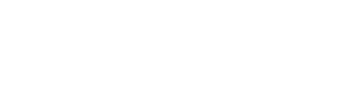
A blue and white logo with a person on a horse

Description automatically generated



T.C

DOKUZ EYLÜL ÜNİVERSİTESİ

FEN FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

**Breast Cancer Wisconsin (Diagnostic) Verisi**

**Sınıflandırma Algoritmaları**

**Ahmet ÖZTÜRK**

**Ali Egehan KIVRAK**

**Danışman: Prof. Dr. Selma Gürler**

**Nisan / 2025**

**İZMİR**

İÇERİK

**1. SINIFLANDIRMA YÖNTEMLERİ** ................................................................................................................ 4  
 1.1. PCA (Temel Bileşenler Analizi) .................................................................................................... 4  
 1.2. K-En Yakın Komşu (KNN) ............................................................................................................. 4  
 1.3. Support Vector Machine (SVM) .................................................................................................. 5  
 1.4. Decision Tree ve Random Forest ................................................................................................ 5  
  1.4.1. Karar Ağaçları (Decision Tree) .............................................................................................. 5  
  1.4.2. Rastgele Ormanlar (Random Forest) .................................................................................... 6  
 1.5. Boosting Algoritmaları ................................................................................................................ 6  
  1.5.1. AdaBoost .............................................................................................................................. 6  
  1.5.2. Gradient Boosting ................................................................................................................ 7  
  1.5.3. XGBoost ............................................................................................................................... 7

**2. VERİ SETİ TANITIMI VE VERİ ÖN İŞLEME** ................................................................................................ 8  
 2.1. Analizin Amacı ........................................................................................................................... 8  
 2.2. Veri Seti Özellikleri .................................................................................................................... 8  
  2.2.1. Değişkenler ve Açıklamaları ............................................................................................... 8

**3. TANIMLAYICI İSTATİSTİKLER** ................................................................................................................. 9  
 3.1. Değişkenlerin Boxplot Grafikleri .............................................................................................. 9  
 3.2. Değişkenlerin Histogramları ................................................................................................... 10  
 3.3. Korelasyon Matrisi ................................................................................................................. 12  
 3.4. Temel Bileşenler Analizi (PCA) ............................................................................................... 13

**4. SINIFLANDIRMA ALGORİTMALARI UYGULAMASI** ............................................................................... 15  
 4.1. KNN Algoritması ………………………….......................................................................................... 15  
 4.2. SVM Uygulaması ..................................................................................................................... 17  
 4.3. Lojistik Regresyon ................................................................................................................... 20  
 4.4. Decision Tree ve Random Forest ............................................................................................ 22  
 4.5. Boosting Yöntemleri: AdaBoost, Gradient Boosting ve XGBoost ........................................... 26

**5. MODEL PERFORMANSLARININ KARŞILAŞTIRILMASI** .......................................................................... 33  
 5.1. Doğruluk, ROC-AUC Ölçümleri …………….................................................................................. 33  
 5.2. En Başarılı Modelin Seçimi ..................................................................................................... 34

**6. SONUÇ VE ÖNERİLER** ........................................................................................................................... 35

**7. KAYNAKÇA** .......................................................................................................................................... 36

**ÖZET**

Bu çalışmada, meme kanseri teşhisinde kullanılabilecek makine öğrenmesi yöntemleri karşılaştırmalı olarak değerlendirilmiştir. UCI Machine Learning Repository’de yer alan **Breast Cancer Wisconsin (Diagnostic) Data Set** kullanılarak, çeşitli sınıflandırma algoritmalarının tanı başarımı analiz edilmiştir. Veri ön işleme aşamasında eksik değer kontrolü, özellik seçimi, dönüştürme işlemleri ve eğitim-test verilerinin ayrımı gerçekleştirilmiştir. Ardından **Lojistik Regresyon, K-En Yakın Komşu (KNN), Karar Ağaçları, Rastgele Orman (Random Forest)** ve **Destek Vektör Makineleri (SVM)** gibi yöntemler uygulanmış; modellerin doğruluk, duyarlılık, özgüllük, F1 skoru ve ROC-AUC gibi metriklerle performansları karşılaştırılmıştır. Ayrıca KNN algoritması hem ham verilerle hem de **Temel Bileşenler Analizi (PCA)** sonrası düşük boyutlu temsillerle değerlendirilmiştir. Elde edilen sonuçlara göre, en yüksek başarıyı sağlayan modeller belirlenmiş ve bu yöntemlerin meme kanseri teşhisinde nasıl kullanılabileceği üzerine çıkarımlar yapılmıştır.

**1.Bölüm**

**SINIFLANDIRMA YÖNTEMLERİ**

**1.1. PCA (Temel Bileşenler Analizi)**

Başlıca Bileşenler Analizi (PCA), özellikle çok sayıda değişken içeren yüksek boyutlu veri setlerinde, veri yapısını bozmadan boyut indirgeme amacıyla kullanılan güçlü bir yöntemdir. PCA, orijinal değişkenleri doğrusal kombinasyonlarla birleştirerek, veriyi varyansı en yüksek olan yeni eksenler (bileşenler) etrafında yeniden yapılandırır. Bu yeni bileşenler, veri setinin bilgi içeriğini mümkün olduğunca koruyacak şekilde oluşturulur.

Çok boyutlu veri setlerinde birçok değişkenin birbiriyle yüksek korelasyon göstermesi, analizlerde bilgi fazlalığına ve yorum karmaşasına yol açabilir. PCA bu durumu minimize ederek, korelasyonlu değişkenleri daha bağımsız ve özet bir yapıya dönüştürür. Bu da özellikle görselleştirme, model karmaşıklığını azaltma ve gürültüyü filtreleme gibi görevlerde avantaj sağlar.

PCA’nın temel amaçları şunlardır:

* **Boyut İndirgeme:** Yüksek boyutlu veri setini daha az bileşenle temsil ederek analizleri sadeleştirmek.
* **Özellik Dönüştürme:** Mevcut değişkenlerin daha anlamlı ve bilgi yoğun bileşenlere dönüştürülmesini sağlamak.
* **Korelasyon Azaltma:** Birbiriyle ilişkili değişkenleri bağımsız bileşenler haline getirerek modelin öğrenmesini kolaylaştırmak.
* **Veri Görselleştirme:** Karmaşık ve çok boyutlu verileri 2 veya 3 boyutlu düzlemde anlaşılabilir hale getirmek.

Bu çalışmada PCA, KNN algoritması ile birlikte değerlendirilmiş ve boyut indirgeme sonrası sınıflandırma performansındaki değişim analiz edilmiştir. Amaç, daha sade ve anlamlı bir özellik uzayında model başarımının nasıl etkilendiğini gözlemlemektir.

**1.2. K-En Yakın Komşu (KNN)**

K-En Yakın Komşu (K-Nearest Neighbors, KNN) algoritması, denetimli öğrenme yöntemleri arasında yer alan, özellikle sınıflandırma problemlerinde yaygın olarak kullanılan bir algoritmadır. Temel mantığı, sınıflandırılması gereken yeni bir verinin, eğitim veri setindeki en yakın **k adet komşusunun** sınıf bilgilerine bakarak sınıflandırılmasıdır. KNN, örnekler arasındaki benzerliği ölçmek için genellikle **mesafe metrikleri** (örneğin, Öklidyen, Manhattan, Minkowski) kullanır.

KNN algoritması, **parametrik olmayan (non-parametric)** ve **tembel öğrenme (lazy learning)** yaklaşımına sahiptir. Bu, modelin eğitim aşamasında herhangi bir genel kural öğrenmediği, tahmin sürecinde doğrudan eğitim verisini kullandığı anlamına gelir. Bu sebeple eğitim süreci çok hızlıdır; ancak test süresi, tüm eğitim verileriyle kıyaslama yapılması gerektiğinden daha uzun olabilir.

Algoritmanın başarısı seçilen **k değeri**, kullanılan **mesafe metriği** ve verinin **ölçeklendirilip ölçeklendirilmediği** gibi faktörlere bağlıdır. Özellikle değişkenlerin farklı ölçeklerde olması, mesafe hesaplamalarını etkileyebileceği için KNN uygulamalarında verilerin standartlaştırılması önerilir.

KNN, yorumlanabilirliği yüksek, basit yapılı ve uygulanması kolay bir algoritma olması nedeniyle, özellikle temel sınıflandırma problemlerinde güçlü bir alternatif olarak tercih edilmektedir.

**1.3. Support Vector Machine (SVM)**

Destek Vektör Makineleri (Support Vector Machine - SVM), sınıflandırma ve regresyon problemleri için kullanılan güçlü ve esnek bir makine öğrenmesi algoritmasıdır. SVM'in temel amacı, veriler arasındaki sınıfları en iyi ayıracak olan optimum ayrım çizgisini (hiperdüzlemi) bulmaktır. Bu ayrım çizgisi, iki sınıf arasında maksimum marj (margin) sağlayacak şekilde yerleştirilir. Marj, sınıflara en yakın veriler olan destek vektörlerine olan uzaklık ile belirlenir.

SVM, özellikle yüksek boyutlu verilerde etkili performans göstermesiyle bilinir. Ayrıca, sınıflar doğrusal olarak ayrılabilir değilse, veriyi daha yüksek boyutlu bir uzaya taşıyarak doğrusal ayrım yapılmasını mümkün kılan kernel (çekirdek) fonksiyonları kullanılır. En yaygın kullanılan kernel türleri; lineer, polinomial, RBF (Radial Basis Function) ve sigmoid fonksiyonlarıdır.

SVM algoritmasında model karmaşıklığını ve hata toleransını kontrol eden bazı önemli parametreler bulunur:

* **C (ceza parametresi):** Hatalı sınıflandırmaları ne kadar tolere edeceğini belirler. Küçük C değeri daha geniş marjlı ancak daha çok hata içeren modeller üretirken, büyük C değeri hatayı azaltmaya odaklanır ama marjı daraltır.
* **Gamma (γ):** RBF gibi kernel fonksiyonlarında, bir veri noktasının etkisinin ne kadar uzağa taşınacağını belirler. Yüksek gamma, daha lokal karar sınırları; düşük gamma ise daha genelleştirilmiş sınıflandırma sağlar.

SVM, overfitting riskine karşı oldukça dirençli olup, özellikle karmaşık ve dengesiz veri setlerinde güçlü sınıflandırma sonuçları elde edebilir. Bununla birlikte, büyük veri setlerinde eğitim süresi uzun olabilir, bu yüzden uygulamalarda hesaplama verimliliği de göz önünde bulundurulmalıdır.

**1.4. Decision Tree ve Random Forest**

**1.4.1. Karar Ağaçları (Decision Tree)**

Karar ağaçları, sınıflandırma ve regresyon problemlerinde yaygın olarak kullanılan, veriyi dallara ayrılan kurallarla sınıflandıran veya tahmin eden bir makine öğrenmesi algoritmasıdır. Ağaç yapısında her düğüm, bir özelliğe göre veri kümesini böler ve her dal, o özelliğin belirli bir değer aralığını temsil eder. Bu süreç, veriler tamamen ayrılana veya önceden belirlenmiş durdurma kriterlerine (ağaç derinliği, minimum yaprak sayısı gibi) ulaşılana kadar devam eder.

Karar ağaçları, veri setindeki karmaşık ilişkileri açık ve anlaşılır biçimde modelleyebilir. Bölme işlemi için genellikle Gini indeks veya Entropi (bilgi kazancı) gibi kriterler kullanılır. Ancak, karar ağaçları yüksek derinliklerde eğitildiğinde, model eğitim verisine aşırı uyum sağlayarak (overfitting) genelleme performansında düşüş yaşayabilir. Bu nedenle, aşırı öğrenmeyi önlemek amacıyla ağaç derinliği sınırlanabilir veya budama teknikleri uygulanabilir.

Avantajları arasında kolay yorumlanabilirlik, hızlı eğitim süreci ve özelliklerin önemli olup olmadığını göstermesi yer alırken, dezavantajları ise yüksek varyans ve aşırı öğrenmeye yatkınlık olarak özetlenebilir.

**1.4.2. Rastgele Ormanlar (Random Forest)**

Random Forest, karar ağaçlarından oluşan bir topluluk (ensemble) yöntemidir ve tek bir karar ağacının sınırlamalarını aşmak için geliştirilmiştir. Model, eğitim verisinden rastgele örneklemelerle (bootstrap sampling) birçok karar ağacı oluşturur. Her ağaç, eğitim setinin farklı bir alt kümesi ve özelliklerin rastgele seçilen bir alt kümesi üzerinde eğitilir. Bu çeşitlilik, ağaçlar arasında bağımsızlık yaratır ve modelin genelleme yeteneğini artırır.

Sonuç olarak, tüm ağaçların oyları (sınıflandırmada çoğunluk oyu) birleştirilerek nihai tahmin yapılır. Bu yöntem, tek karar ağacına göre çok daha sağlam ve hataya karşı dayanıklı modeller ortaya koyar. Random Forest, özellikle gürültülü veri setlerinde, yüksek boyutlu veri kümelerinde ve karmaşık ilişkilerin bulunduğu durumlarda başarılı performans sergiler.

Ayrıca, modelin hiperparametreleri (örneğin ağaç sayısı, maksimum derinlik, her düğümde değerlendirilen özellik sayısı) ayarlanarak performans optimize edilebilir. Random Forest’ın en önemli avantajlarından biri, overfitting’i azaltması ve değişken önemini ölçerek özellik seçimine katkıda bulunmasıdır.

**1.5 Boosting Algoritmaları**

Boosting, makine öğrenmesinde kullanılan etkili bir topluluk (ensemble) yöntemidir. Temel hedefi, performansı düşük olan bir dizi zayıf öğreniciyi ardışık olarak eğitip, bunları birleştirerek daha güçlü ve doğru bir tahmin modeli elde etmektir.

Boosting’in çalışma prensibi şudur: Genellikle basit modeller (örneğin küçük karar ağaçları) kullanılır ve her yeni model, önceki modellerin yaptığı hatalara odaklanarak onları düzeltmeye çalışır. Bu süreçte, yanlış sınıflandırılan verilerin ağırlıkları artırılır, böylece sonraki model bu zorluklara daha fazla önem verir. Son aşamada, tüm modellerin çıktıları belirli ağırlıklarla birleştirilerek nihai karar oluşturulur.

**1.5.1 Adaboost**

AdaBoost (Adaptive Boosting) ise bu yaklaşımın en yaygın kullanılan versiyonudur. AdaBoost, birden fazla zayıf sınıflandırıcıyı —genellikle karar ağaçlarını— ardışık şekilde eğiterek, hataları azaltan ve genel doğruluğu artıran güçlü bir sınıflandırıcı yaratır. AdaBoost, özellikle bireysel zayıf modellerin tek başına yeterince başarılı olmadığı durumlarda performansı büyük ölçüde yükseltir.

Algoritmanın işleyişi şu şekilde gerçekleşir: Başlangıçta tüm eğitim örneklerine eşit ağırlık verilir. İlk model eğitildikten sonra, hatalı sınıflandırılan örneklerin ağırlıkları yükseltilir. Böylece, sonraki model bu örneklere daha fazla odaklanır. Bu işlem birkaç iterasyon boyunca devam eder ve sonunda tüm zayıf modellerin tahminleri, kendi başarılarına bağlı olarak ağırlıklandırılarak birleştirilir.

AdaBoost, hata oranı yüksek olan basit modellerin performansını artırmak ve genel tahmin doğruluğunu yükseltmek için tercih edilen güçlü ve esnek bir yöntemdir.

**1.5.2 Gradient Boosting**

Gradient Boosting, makine öğrenmesinde güçlü ve esnek bir *ensemble* (birleşik) yöntemdir. Bu yöntem, hataları minimize etmeye odaklı bir şekilde ardışık modeller kurarak, zayıf öğrenicilerden (genellikle karar ağaçları) oluşan bir yapı ile yüksek doğrulukta tahminler yapmayı amaçlar.

Gradient Boosting’in temel mantığı, her yeni modelin bir önceki modelin yapamadığı doğru tahminleri hedeflemesidir. Özellikle önceki modellerin yaptığı hatalara odaklanılır ve bu hatalar, bir kayıp fonksiyonu (loss function) üzerinden *gradyan iniş* (gradient descent) yaklaşımıyla azaltılmaya çalışılır. Bu süreçte, model sürekli olarak hatayı en aza indirmeye yönelik iyileştirmeler yapar.

**Gradient Boosting’in Temel Adımları:**

1. İlk model veri üzerinde eğitilir ve hata (residual) hesaplanır.
2. Sonraki model, bu hataları öğrenmeye çalışır.
3. Elde edilen yeni tahminler, önceki tahminlerle birleştirilir.
4. Bu işlem belirli sayıda yinelemeyle devam eder.
5. Sonuçta tüm modellerin çıktıları birleştirilerek güçlü bir model elde edilir.

Bu yöntem, sınıflandırma ve regresyon problemlerinde yüksek doğruluk sağlar. Bununla birlikte, model karmaşık hale gelebildiği için **overfitting (aşırı öğrenme)** riski taşıyabilir. Bu nedenle, öğrenme oranı (*learning rate*), ağaç derinliği ve iterasyon sayısı gibi hiperparametreler dikkatle ayarlanmalıdır

**1.5.3 XGBoost**

XGBoost, Gradient Boosting algoritmasının geliştirilmiş ve optimize edilmiş bir versiyonudur. Özellikle hız, performans ve genel doğruluk açısından güçlü sonuçlar vermesiyle öne çıkar. Büyük veri setleri ve karmaşık yapılar üzerinde çalışırken oldukça etkili bir sınıflandırma ve regresyon yöntemidir.

XGBoost’un temel amacı, hata oranını en aza indirirken modeli aşırı karmaşıklaştırmamak ve eğitim süresini kısaltmaktır. Bunu başarmak için çeşitli teknikler kullanır:

* **Düzenlileştirme (Regularization):** Aşırı öğrenmeyi (overfitting) önlemek için modelin karmaşıklığını kontrol eder.
* **Eksik Verilerle Baş Etme:** Eksik değerleri otomatik olarak işler.
* **Ağaç Yapısında Optimizasyon:** Daha verimli ağaç yapıları oluşturur.
* **Paralel Hesaplama:** Eğitim sürecini hızlandırmak için çok çekirdekli işlemeyi destekler.

XGBoost, hem doğruluğu yüksek hem de hesaplama açısından oldukça verimlidir. Ayrıca sınıflandırma, regresyon ve sıralama gibi birçok farklı problem türünde etkili bir şekilde kullanılabilir.

**2. VERİ SETİ TANITIMI VE VERİ ÖN İŞLEME**

**2.1. Analizin Amacı**

Bu analiz çalışmasının temel amacı, meme kanseri teşhisinde kullanılabilecek en etkili makine öğrenmesi algoritmasını belirlemektir. Uygulamada kullanılan **“Breast Cancer Wisconsin”** veri seti; hücre çekirdek özelliklerine dayalı olarak tümörlerin malign (kötü huylu) ya da benign (iyi huylu) olarak sınıflandırılmasını mümkün kılmaktadır.

Bu bağlamda, farklı sınıflandırma algoritmaları (K-En Yakın Komşu (KNN), Karar Ağaçları, Destek Vektör Makineleri (SVM), Random Forest, AdaBoost, Gradient Boosting ve XGBoost gibi) veri seti üzerinde uygulanarak, her bir modelin doğruluk oranı, hata oranı ve sınıflandırma başarıları karşılaştırılmıştır. Ayrıca bazı yöntemlerde **PCA (Principal Component Analysis)** gibi boyut indirgeme tekniklerinden faydalanılarak, model performansının iyileştirilmesi amaçlanmıştır.

Sonuç olarak, farklı algoritmaların teşhis doğruluğu üzerindeki etkileri değerlendirilecek ve tıbbi karar destek sistemleri geliştirmede en etkili sınıflandırma yöntemi ortaya konacaktır.

**2.2. Veri Seti Özellikleri**

**Meme Kanseri Tanı Verisi (Breast Cancer Wisconsin Dataset),** meme kanseri tanısında kullanılan biyopsi verileri üzerinden iyi huylu (benign) ve kötü huylu (malignant) tümörleri sınıflandırmayı amaçlayan önemli bir veri setidir. Klinik çalışmalar, tıbbi araştırmalar ve makine öğrenmesi tabanlı teşhis sistemleri geliştirmek isteyen araştırmacılar için sıklıkla başvurulan bir kaynaktır. Bu veri seti, hücre çekirdeği morfolojisine ait sayısal özellikleri temel alarak bir tümörün tipini belirlemeye yardımcı olur.

Veri seti, toplamda 569 hastaya ait gözlemlerden oluşmakta ve her bir gözlem bir hastaya ait hücresel görüntüleme sonuçlarını içermektedir. Hedef değişken olan “Teşhis” (Diagnosis), tümörün iyi huylu (B) ya da kötü huylu (M) olduğunu belirtir.

**2.2.1. Değişkenler ve Açıklamaları**

* **ID Number (Kimlik Numarası):** Her hasta için benzersiz bir tanımlayıcıdır. Analizlerde kullanılmaz, sadece kayıt tutma amacı taşır.
* **Diagnosis (Teşhis):** Veri setinin hedef (bağımlı) değişkenidir. Tümörün türünü belirtir. “M” kötü huylu (malignant), “B” iyi huylu (benign) tümörleri temsil eder.
* **Hücre Özellikleri (Bağımsız Değişkenler):** Hücre çekirdeği üzerine yapılan ölçümlerden elde edilen 30 sayısal değişkenden oluşur. Bu özellikler şunlardır:
  + **Radius (Yarıçap):** Çekirdek merkezinden kenarlara olan ortalama uzaklık.
  + **Texture (Doku):** Gri tonlamalı piksellerin standart sapması; yapısal düzensizlikleri temsil eder.
  + **Perimeter (Çevre):** Hücre çekirdeğinin çevresinin uzunluğu.
  + **Area (Alan):** Hücre çekirdeğinin kapladığı toplam alan.
  + **Smoothness (Düzgünlük):** Yüzeyin düzgünlüğü; daha düşük değerler daha düzgün yüzeyleri temsil eder.
  + **Compactness (Sıkılık):** Hücre çekirdeğinin yoğunluk düzeyidir (çevre² / alan – 1).
  + **Concavity (İçe Çöküklük):** Hücre kenarlarının içe çökük bölümlerinin büyüklüğü.
  + **Concave Points (İçe Çökük Noktalar):** Hücre sınırındaki çöküntü noktalarının sayısı.
  + **Symmetry (Simetri):** Hücrenin simetrik yapısını temsil eder.
  + **Fractal Dimension (Fraktal Boyut):** Hücre kenarlarının karmaşıklığını ölçen bir katsayı.

Her bir temel hücre özelliği için üç farklı ölçüm sunulmuştur:

1. **Mean (Ortalama):** Özelliğin genel ortalama değeri.
2. **Standard Error (Standart Hata):** Özelliğin ölçümündeki belirsizliği ifade eder.
3. **Worst (En Yüksek Değer Ortalaması):** İlgili özelliğin en yüksek üç değerinin ortalamasıdır.

Toplamda bu yapı sayesinde 30 farklı sayısal değişken elde edilmiştir. Veri setinde eksik değer bulunmamaktadır ve tüm değişkenler dört ondalık hassasiyetle sayısal olarak kodlanmıştır. Bu detaylı yapı, özellikle sınıflandırma algoritmalarıyla çalışmak için oldukça elverişlidir.

**3. TANIMLAYICI İSTATİSTİKLER**

**3.1. Değişkenlerin Boxplot Grafikleri**

A graph with a purple square

AI-generated content may be incorrect.

A chart of different colored squares

AI-generated content may be incorrect.

Bu bölümde veri setindeki değişkenlerin görselleştirilmesi için boxplot kullanılmıştır. Veri setinin orijinal halinin boxplotını ve standartlaştırılmış(scale) boxplotı görebilirsiniz.

**3.2. Değişkenlerin Histogramları** A group of blue bars

AI-generated content may be incorrect.

A group of blue and white graphs

AI-generated content may be incorrect.

Several blue and white bars

AI-generated content may be incorrect.

A graph with blue squares

AI-generated content may be incorrect.

Bu kısımda veri setindeki değişkenlerin dağılımını görmek için çizilmiş histogramları görebilirsiniz.

**3.3. Korelasyon Matrisi**

A close-up of a data sheet

AI-generated content may be incorrect.

Veri setindeki değişkenler arasındaki doğrusal ilişkileri incelemek için korelasyon matrisi oluşturulmuştur. Korelasyon matrisi, değişken çiftleri arasındaki ilişkiyi Pearson korelasyon katsayısı ile ölçer ve -1 ile 1 arasında bir değer alır:

• 1: Güçlü pozitif ilişkiyi ifade eder; bir değişken artarken diğeri de artar.

• 0: İki değişken arasında hiçbir doğrusal ilişki olmadığını gösterir.

• -1: Güçlü negatif ilişkiyi ifade eder; bir değişken artarken diğeri azalır.

Korelasyon matrisi, veri setindeki önemli değişkenler arasındaki ilişkiyi şu şekilde özetlemiştir:

**Radius\_mean** ile **Perimeter\_mean** arasında çok güçlü ve pozitif bir ilişki vardır. Bu durum, tümörün boyutunu ölçen bu iki değişkenin birlikte artma eğiliminde olduğunu ve birbirini desteklediğini göstermektedir.

**Texture\_mean** ile **Smoothness\_mean** arasında zayıf bir pozitif ilişki gözlenmiştir. Bu, tümörün dokusal özellikleri ile yüzey pürüzlülüğü arasında belirgin bir bağlantı olmadığını işaret etmektedir.

**Compactness\_mean** ve **Concavity\_mean** değişkenleri arasında yüksek bir pozitif korelasyon bulunmuştur. Bu, tümörün şekil bozukluklarını ifade eden bu iki değişkenin birlikte hareket ettiğini, yani tümörün yapısal düzensizliklerinin birlikte arttığını göstermektedir.

Ayrıca **Fractal\_dimension\_mean** ile **Radius\_mean** arasında anlamlı negatif bir ilişki vardır. Bu, tümörün karmaşıklık düzeyi arttıkça genel boyutunun küçüldüğünü düşündürmektedir.

**3.4. Temel Bileşenler Analizi (PCA)**

Bu çalışmada, orijinal veri setindeki çok sayıda değişkenin boyutunu azaltmak ve verideki temel yapıyı daha iyi anlamak amacıyla Temel Bileşenler Analizi (PCA) uygulanmıştır. PCA, birbirleriyle yüksek korelasyona sahip değişkenlerden oluşan veri setlerinde, bu değişkenlerin varyansını en iyi şekilde açıklayan yeni, bağımsız temel bileşenler oluşturur. Böylece, verideki bilgi kaybı minimum düzeyde tutulurken, değişken sayısı azaltılarak analizlerin daha etkin ve anlaşılır hale gelmesi sağlanır.

A graph with a number of blue bars

AI-generated content may be incorrect.

Temel bileşenler analizi uyguladığımız verimize baktığımızda, ilk 5 bileşenin varyansın %84.82’sini açıkladığını görebiliyoruz. Bu durum, verinin büyük bir kısmının ilk iki bileşen ile temsil edilebileceğini ve analizde daha az sayıda bileşenle çalışmanın uygun olduğunu göstermektedir. Bu bileşenler, verideki temel desenleri ve değişkenler arasındaki ilişkileri özetleyerek, kümelerin ayrışmasını daha net bir şekilde ortaya koymuştur. A diagram of a graph

AI-generated content may be incorrect.

PCA sonuçları iki boyutlu bir grafik üzerinde görselleştirilmiştir. Burada değişkenlerin yüklerinin (vektörlerinin) farklı yönlerde veya açılarda yer alması, değişkenler arasındaki ilişkiler hakkında önemli bilgiler sunmaktadır: 90 derece açıda (dik) yer alan değişkenler aralarında anlamlı bir ilişki bulunmadığını, yani birbirlerinden bağımsız olduklarını ifade ediyor.

**4. SINIFLANDIRMA ALGORİTMALARI UYGULAMASI**

**4.1. KNN Algoritması**

**Model Kurulumu ve Parametre Seçimi**

Modelin performansını artırmak için 10 katlı çapraz doğrulama yöntemi kullanıldı. Bu sayede farklı *k* değerleri denenerek, en uygun *k* değeri belirlendi. İncelenen tek sayılar arasından *k=9* en iyi sonuçları verdi. Bu, modelin hem aşırı uyum yapmasını engelledi hem de yeterince komşu göz önüne alarak karar vermesini sağladı.

**Model Performansı ve Değerlendirme**

Test verisi üzerindeki değerlendirmede KNN modeli yüksek bir başarı gösterdi:

* Doğruluk (Accuracy): %95.29
* Kappa: 0.8985
* Duyarlılık (Sensitivity): %97.20
* Özgüllük (Specificity): %92.06
* Dengelenmiş Doğruluk (Balanced Accuracy): %94.63
* Pozitif Sınıf: B

Model, özellikle pozitif sınıfı (B) çok iyi tespit etti ve genel sınıflandırma hataları oldukça düşüktü.

**ROC Eğrisi ve AUC Skoru**

Modelin ayırma gücünü değerlendirmek için ROC eğrisi oluşturuldu ve AUC değeri hesaplandı. KNN modelinin AUC skoru 0.982 seviyesinde bulunarak, sınıflar arasındaki ayrımı oldukça başarılı biçimde gerçekleştirdiği görüldü.

A graph of a function

AI-generated content may be incorrect.

**Farklı Mesafe Metrikleri ile Performans Karşılaştırması**

Model, farklı mesafe ölçütleri kullanılarak da test edildi ve performansları karşılaştırıldı:

* **Euclidean (Öklid) Mesafesi:** Doğruluk %95.29, Kappa 0.8985, Dengelenmiş Doğruluk %94.63
* **Manhattan Mesafesi:** Doğruluk %95.88, Kappa 0.9114, Dengelenmiş Doğruluk %95.42
* **Minkowski Mesafesi (p=3):** Doğruluk %94.12, Kappa 0.8739, Dengelenmiş Doğruluk %93.70

Bu karşılaştırma, Manhattan mesafesinin bu veri setinde en iyi performansı verdiğini ortaya koydu. Euclidean mesafesi de güçlü sonuçlar verirken, Minkowski metriği (p=3) biraz daha düşük başarı gösterdi.

**Genel Değerlendirme**

KNN algoritması, doğru *k* değeri ve uygun mesafe metriği seçildiğinde güçlü ve güvenilir sınıflandırma sonuçları vermektedir. Özellikle Manhattan mesafesi kullanılarak, model hem hassas hem de dengeli bir şekilde sınıflandırma yapabilmiştir. KNN, basit yapısına rağmen bu veri setinde yüksek performans sergileyerek etkili bir alternatif olarak ön plana çıkmıştır.

**PCA Sonrası k-NN Model Performansı**

PCA sonrası elde edilen k-NN modeli, boyut indirgeme sayesinde karmaşıklığı azaltılmış ve model daha sade hale getirilmiştir. 10 katlı çapraz doğrulama ile en iyi k değeri 13 olarak seçilmiştir. Bu parametre, komşu sayısını yeterli düzeyde tutarak aşırı öğrenmeyi önlemiş ve dengeli bir sınıflandırma sağlamıştır.

Test seti üzerinde yapılan değerlendirmede modelin performansı şu şekildedir:

* **Doğruluk (Accuracy):** %87.06 Modelin genel olarak doğru sınıflandırma oranı yüksek.
* **AUC:** 0.935 Sınıflar arasında güçlü ayrım yeteneği olduğunu gösterir.
* **Duyarlılık (Sensitivity):** %96.26 Pozitif sınıf (örneğin “B” yani benign) doğru tespit oranı oldukça yüksek.
* **Özgüllük (Specificity):** %71.43 Negatif sınıfın da iyi ayırt edildiği görülmekte, ancak duyarlılığa kıyasla biraz daha düşük.
* **Kappa:** 0.71 Modelin tesadüfen değil, gerçek anlamda iyi bir sınıflandırma yaptığına işaret eder.

A graph of a function

AI-generated content may be incorrect.

Ek olarak, farklı mesafe fonksiyonları denendiğinde Euclidean mesafe performans açısından en uygun seçim olarak öne çıktı. Manhattan mesafede doğruluk biraz düşerken, Minkowski (p=3) daha zayıf sonuç verdi. Bu durum PCA sonrası veride Euclidean mesafenin daha tutarlı ve anlamlı olduğunu gösteriyor.

**PCA Öncesi ve Sonrası Genel Karşılaştırma**

* PCA sonrası modelde değişken sayısı azaltılarak modelin işlem yükü ve karmaşıklığı ciddi oranda düşürülmüştür.
* Performans açısından küçük bir düşüş yaşanmıştır, fakat %87 doğruluk ve 0.93 AUC, boyut indirgeme sonrasında da güçlü bir model kurulduğunu gösterir.
* PCA öncesinde, model daha yüksek boyutlu ve gürültülü veriye maruz kaldığından parametre seçimi ve model kararlılığı daha değişken olabilmektedir.
* Mesafe fonksiyonları PCA öncesinde farklı sonuçlar verirken, PCA sonrası Euclidean mesafe tutarlı sonuçlar sunmuştur.
* Genel anlamda, PCA boyut indirgeme işlemi, küçük bir performans kaybına rağmen modelin daha hızlı ve stabil çalışmasını sağlamaktadır.

**4.2. SVM Uygulaması**

**Uygulama Adımları**  
Modelleme sürecinde üç farklı kernel fonksiyonu kullanılarak SVM performansı değerlendirilmiştir: linear, radial ve polynomial. Her bir model, ölçeklenmiş eğitim verisi ile eğitilmiş ve test verisi üzerinde doğruluk oranlarına göre karşılaştırılmıştır. Aşağıda her bir çekirdek fonksiyonu için elde edilen doğruluk sonuçları yer almaktadır:

* **Linear kernel**: %98.82 doğruluk
* **Radial kernel**: %98.82 doğruluk
* **Polynomial kernel**: %92.35 doğruluk

Yukarıdaki sonuçlar doğrultusunda en yüksek doğruluk oranına ulaşan linear çekirdek fonksiyonu en iyi model olarak seçilmiştir. Bu kernel ile hiperparametre optimizasyonu yapılmıştır. Özellikle cost parametresi için 0.1 ile 3 arasında bir aralıkta tarama yapılmış ve en iyi sonuç cost = 0.1 değeriyle elde edilmiştir.

**Final Model Performansı**  
Optimum parametreler ile oluşturulan final model, test verisi üzerinde %99.41 doğruluk oranına ulaşmıştır. Modelin duyarlılık (sensitivity) değeri %100, özgüllük (specificity) değeri ise %98.41 olarak hesaplanmıştır. Bu da modelin hem pozitif sınıfları doğru tahmin etmede hem de negatif sınıfları doğru şekilde ayırt etmede oldukça başarılı olduğunu göstermektedir.

Ayrıca modelin ROC eğrisi çizilmiş ve AUC (Area Under Curve) değeri hesaplanmıştır. AUC değeri **0.9988** olarak bulunmuş, bu da modelin sınıfları ayırmadaki genel performansının oldukça yüksek olduğunu göstermektedir.

**A graph of a function

AI-generated content may be incorrect.**

**Sonuç**  
SVM algoritması, özellikle linear çekirdek fonksiyonu ile uygulandığında yüksek doğruluk, duyarlılık ve AUC değerleriyle etkili bir sınıflandırma performansı sergilemiştir. Hiperparametre ayarlaması ile modelin başarısı daha da optimize edilmiştir. Elde edilen bu sonuçlar, SVM'nin bu veri seti için uygun ve güvenilir bir yöntem olduğunu ortaya koymaktadır.

**PCA Sonrası SVM Model Performansı ve Değerlendirme**

PCA sonrası boyut indirgeme uygulanmış veri ile kurulan SVM modelleri farklı kernel türleri (linear, radial, polynomial) ile denenmiştir. Performans karşılaştırmalarında en iyi doğruluk (accuracy) %55.29 ile **linear kernel** elde edilmiştir.

* **Accuracy (Doğruluk):** %55.29 (Linear kernel)
* **Sensitivity (Duyarlılık):** %53.27 (Pozitif sınıfın doğru tespiti)
* **Specificity (Özgüllük):** %58.73 (Negatif sınıfın doğru tespiti)
* **Kappa:** 0.11 (Orta düşük, modelin tesadüfi sınıflandırmadan biraz daha iyi olduğunu gösterir)
* **AUC:** 0.56 (Sınıflar arası ayrım gücü oldukça düşük, sınıflandırma gücü zayıf)

Model için en iyi kernel linear seçilmiş ve cost=0.8 parametresiyle tuning yapılmıştır. Final model doğrulukta çok az düşüş yaşasa da (yaklaşık %54), aşırı uyumdan kaçınmak ve modeli basitleştirmek adına uygun bir parametre ayarıdır.

A graph of a function

AI-generated content may be incorrect.**PCA Öncesi ve Sonrası SVM Karşılaştırması**

PCA öncesi veri ile model genellikle daha yüksek boyut ve gürültü içerdiğinden, modelin kararlılığı ve doğruluğu daha değişkendir. PCA sonrası modelde boyut indirgeme sayesinde model sadeleşmiş, ancak doğruluk ve AUC değerleri düşük kalarak sınıflandırma başarısı sınırlı olmuştur.

* PCA sonrası **SVM modelinin doğruluğu %54-55** seviyesindedir; bu değer sınıflandırmanın biraz üzerinde olsa da pratikte yeterli performans değildir.
* PCA öncesi modelde daha yüksek doğruluk veya AUC elde edilebiliyorsa, boyut indirgeme aşamasında önemli bilgi kaybı olmuş olabilir.
* PCA sonrası modelde linear kernelin radial ve polynomial kernelden daha başarılı olması, indirgenmiş veri yapısının doğrusal sınıflandırma için daha uygun olduğunu gösterir.

**Sonuç**

PCA sonrası SVM modeli, boyut indirgemeyle veri yapısını sadeleştirirken, model performansında belirgin bir azalma gözlenmiştir. En iyi parametrelerle linear kernel tercih edilmiş olsa da, modelin AUC ve doğruluk değerleri sınıflandırma problemini yeterince iyi çözmediğini işaret etmektedir. Bu durumda, PCA sonrası veri için SVM yerine alternatif modeller (örneğin k-NN, Random Forest) denenebilir veya PCA bileşen sayısı ve parametrik ayarlar yeniden gözden geçirilebilir.

**4.3. Lojistik Regresyon**

**Model Performansı ve Değerlendirme**  
Model eğitimi sırasında, bazı gözlemlerin sınıflar arasında neredeyse tamamen ayrılması nedeniyle, algoritmanın yakınsama uyarısı verdiği görülmüştür. Bu, modelin bazı sınıfları çok net ayırdığına işaret etmektedir. Buna rağmen, test verisi üzerinde yüksek başarı göstermiştir.

* **Doğruluk (Accuracy):** %94.12
* **Duyarlılık (Sensitivity):** %96.26
* **Özgüllük (Specificity):** %90.48
* **Kappa:** 0.8731
* **Balanced Accuracy:** %93.37
* **Pozitif Sınıf:** B

Bu sonuçlar, modelin her iki sınıfı da dengeli ve etkili bir şekilde ayırt edebildiğini göstermektedir. Özellikle yüksek duyarlılık değeri, modelin pozitif sınıf olan "iyi huylu" tümörleri başarılı bir şekilde tanıdığını ortaya koymaktadır.

**ROC Eğrisi ve AUC Skoru**  
Modelin genel sınıflandırma performansı ROC eğrisi altında kalan alan (AUC) değeriyle de değerlendirilmiştir. Elde edilen:

* **AUC skoru:** 0.9749

Bu değer, modelin hem doğru pozitifleri hem de doğru negatifleri yüksek bir başarıyla tahmin ettiğini göstermektedir. ROC eğrisi, ideal sınıflayıcıya oldukça yakın bir şekilde, başarılı bir ayrım gücüne işaret etmektedir.

A graph of a logistic plot

AI-generated content may be incorrect.

**Sonuç**  
Lojistik regresyon modeli, bu çalışmada yüksek doğruluk, denge ve ayrım gücü ile etkili bir sınıflandırma algoritması olduğunu göstermiştir. Basit yapısına rağmen güçlü bir performans sergileyen model, özellikle yorumlana bilirliğin önemli olduğu durumlar için ideal bir tercih olmayı sürdürmektedir.

**PCA Sonrası Lojistik Regresyon Model Performansı ve Değerlendirme**

PCA sonrası boyut indirgeme uygulanmış veri ile kurulan lojistik regresyon modeli, sınıflandırma performansında sınırlı bir başarı göstermiştir. Modelin doğruluk (accuracy) oranı yaklaşık %54 civarında olup, sınıflar arasındaki ayrım gücü oldukça düşüktür.

* **Accuracy (Doğruluk):** %54.12 (Modelin genel doğruluk oranı)
* **Sensitivity (Duyarlılık):** %52.34 (Pozitif sınıfın doğru tespiti)
* **Specificity (Özgüllük):** %57.14 (Negatif sınıfın doğru tespiti)
* **Kappa:** 0.0879 (Modelin rasgele sınıflandırmadan sadece biraz daha iyi olduğunu gösterir)
* **AUC:** 0.57 (Modelin sınıflar arasındaki ayrım gücünün zayıf olduğunu gösterir)

Modelde 0.5 eşik değeriyle sınıflandırma yapılmıştır. PCA sonrası veri yapısı nedeniyle model sadeleşmiş, ancak sınıflandırma başarısı istenilen seviyeye ulaşamamıştır.

A graph of a function

AI-generated content may be incorrect.

**PCA Öncesi ve Sonrası Lojistik Regresyon Karşılaştırması**

PCA öncesi modelde yüksek boyutlu ve gürültülü veriler nedeniyle model performansı dalgalanabilir. Boyut indirgeme sonrasında model daha basit hale gelmiş, ancak bu sadeleşme beraberinde önemli bir bilgi kaybını getirmiş olabilir. Bu nedenle, PCA sonrası model doğruluk ve AUC değerleri düşük kalmış ve sınıflandırma başarısı sınırlı olmuştur.

PCA sonrası lojistik regresyon modelinin sınıflandırma doğruluğu ve AUC değeri, modeli pratikte kullanmak için yetersiz kalmaktadır.

**Sonuç**

PCA sonrası lojistik regresyon modeli, boyut indirgeme sayesinde veri yapısını sadeleştirirken, performansta belirgin bir düşüş yaşamıştır. Model, sınıflar arasındaki ayrımı yeterince yapamamaktadır. Bu durumda, PCA sonrası veri için lojistik regresyon yerine farklı modeller denenebilir ya da PCA bileşen sayısı ve model parametreleri yeniden optimize edilebilir.

Alternatif olarak, PCA sonrası k-NN veya SVM gibi modellerle karşılaştırma yapılarak en uygun yaklaşım belirlenmelidir.

**4.4. Decision Tree ve Random Forest**

**Karar Ağaçları (Decision Tree)**

**Model Performansı ve Değerlendirme**  
Karar ağacı modeli, doğruluk oranı açısından tatmin edici bir performans sergilemiştir. Her ne kadar daha karmaşık modeller kadar yüksek başarı sağlamasa da, basitliği ve açıklanabilirliğiyle önemli avantajlara sahiptir.

* **Doğruluk (Accuracy):** %91.18
* **Duyarlılık (Sensitivity):** %90.65
* **Özgüllük (Specificity):** %92.06
* **Kappa:** 0.8139
* **Balanced Accuracy:** %91.36
* **Pozitif Sınıf:** B

Modelin hem pozitif hem negatif sınıflarda dengeli sonuçlar ürettiği görülmektedir. Duyarlılığın yüksek olması, iyi huylu tümörlerin büyük çoğunluğunun doğru şekilde tespit edildiğini gösterirken, özgüllük değeri de kötü huylular için makul düzeyde başarıya işaret etmektedir.

**ROC Eğrisi ve AUC Skoru**  
Modelin genel sınıflandırma başarısı, ROC eğrisi ve altında kalan alan (AUC) değeri ile değerlendirilmiştir.

* **AUC skoru:** 0.940

Bu değer, karar ağacının sınıfları ayırt etme yetisinin yüksek olduğunu ve modelin güvenilir tahminler yaptığını ortaya koymaktadır.

A graph of a tree

AI-generated content may be incorrect.

**Sonuç**  
Karar ağacı modeli, oldukça tatmin edici bir doğrulukla çalışmış, hem görsellik hem de yorumlanabilirlik açısından avantaj sunmuştur. Karmaşık modeller kadar yüksek performans göstermese de, sade yapısıyla anlaşılır ve uygulanabilir bir sınıflandırma çözümü sağlamıştır.

**Rastgele Ormanlar (Random Forest)**

**Model Performansı ve Değerlendirme**  
Random Forest modeli yüksek başarı oranı ile dikkat çekmiştir. Hem benign (iyi huylu) hem de malignant (kötü huylu) sınıfları oldukça dengeli şekilde sınıflandırabilmiştir.

* **Doğruluk (Accuracy):** %95.88
* **Duyarlılık (Sensitivity):** %96.26
* **Özgüllük (Specificity):** %95.24
* **Kappa:** 0.912
* **Balanced Accuracy:** %95.75
* **Pozitif Sınıf:** B

Bu sonuçlar, modelin genel olarak sağlam ve güvenilir tahminlerde bulunduğunu göstermektedir. Yanlış sınıflandırma oranı oldukça düşüktür ve model, veri setinin sınıflarını iyi ayırt edebilmektedir.

**ROC Eğrisi ve AUC Skoru**  
Modelin sınıflandırma gücü ROC analizi ile de değerlendirilmiş ve son derece başarılı bir sonuç elde edilmiştir.

* **AUC skoru:** 0.989

Bu skor, sınıflar arasında ayırt ediciliğin neredeyse mükemmel düzeyde olduğunu göstermektedir.

A graph of a function

AI-generated content may be incorrect.

**Değişken Önem Düzeyleri**  
Random Forest ayrıca, hangi değişkenlerin sınıflandırma sürecinde daha etkili olduğunu belirlemek için değişken önem grafiği (Variable Importance Plot) sunar. Bu grafik, modelin kararlarını hangi özelliklerin yönlendirdiğini görselleştirmek açısından değerlidir.

A graph of a graph with a line of points

AI-generated content may be incorrect.

**Sonuç**  
Random Forest, yüksek doğruluk oranı, istikrarlı sınıflandırma performansı ve yorumlanabilir değişken önem çıktıları ile güçlü bir modelleme tekniğidir. Karmaşık yapılı veri setlerinde bile başarılı sonuçlar üretebilir. Bu çalışmada da, gerek ROC eğrisi gerekse genel metriklerde öne çıkan modellerden biri olmuştur.

**PCA Sonrası Random Forest Model Performansı ve Değerlendirme**

PCA sonrası boyut indirgeme uygulanmış veri ile kurulan Random Forest modeli, sınıflandırma performansında sınırlı bir başarı göstermiştir. Modelin doğruluk (accuracy) oranı yaklaşık %44 civarında olup, sınıflar arasındaki ayrım gücü oldukça düşüktür.

* **Accuracy (Doğruluk):** %44.12 (Modelin genel doğruluk oranı)
* **Sensitivity (Duyarlılık):** %28.04 (Pozitif sınıfın doğru tespiti)
* **Specificity (Özgüllük):** %71.43 (Negatif sınıfın doğru tespiti)
* **Kappa:** -0.0045 (Modelin rasgele sınıflandırmanın bile altında olduğunu gösterir)
* **AUC:** 0.63 (Modelin sınıflar arasındaki ayrım gücünün zayıf olduğunu gösterir)

Modelde, PCA sonrası veri yapısı nedeniyle sınıflandırma başarısı istenilen seviyeye ulaşamamıştır.

A graph of a graph

AI-generated content may be incorrect.

**PCA Öncesi ve Sonrası Random Forest Karşılaştırması**

PCA öncesi modelde yüksek boyutlu ve gürültülü veriler nedeniyle model performansı dalgalanabilir. Boyut indirgeme sonrasında model daha sade hale gelmiş, ancak bu sadeleşme beraberinde önemli bir bilgi kaybını getirmiş olabilir. Bu nedenle, PCA sonrası model doğruluk ve AUC değerleri düşük kalmış ve sınıflandırma başarısı sınırlı olmuştur.

**Sonuç**

PCA sonrası Random Forest modeli, boyut indirgeme sayesinde veri yapısını sadeleştirirken performansta belirgin bir düşüş yaşamıştır. Model, sınıflar arasındaki ayrımı yeterince yapamamaktadır. Bu durumda, PCA sonrası veri için Random Forest yerine farklı modeller denenebilir ya da PCA bileşen sayısı ve model parametreleri yeniden optimize edilebilir. PCA sonrası Random Forest modeli, doğruluk ve AUC değerleri açısından lojistik regresyon ve SVM modellerine kıyasla daha düşük performans göstermiştir. Bu nedenle, PCA sonrası veri için Random Forest yerine lojistik regresyon veya SVM gibi modeller tercih edilebilir veya model parametreleri yeniden optimize edilmelidir.

Alternatif olarak, PCA sonrası k-NN, SVM veya lojistik regresyon gibi modellerle karşılaştırma yapılarak en uygun yaklaşım belirlenmelidir.

**4.5. Boosting Yöntemleri: AdaBoost, Gradient Boosting ve XGBoost**

**Adaboost Model Performansı ve Değerlendirme**  
AdaBoost modeli, yüksek başarı oranıyla öne çıkmıştır. Hem benign (B) hem de malignant (M) sınıfları çok başarılı bir şekilde ayırt edebilmiştir.

* **Doğruluk (Accuracy):** %96.47
* **Duyarlılık (Sensitivity):** %98.13
* **Özgüllük (Specificity):** %93.65
* **Kappa:** 0.9238
* **Balanced Accuracy:** %95.89
* **Pozitif Sınıf:** B

Model, sınıflar arasında oldukça dengeli ve etkili bir ayrım yapmıştır. Özellikle benign vakaları tespit etmede çok yüksek bir duyarlılık sergilemiştir.

**ROC Eğrisi ve AUC Skoru**  
AdaBoost modeli, ROC eğrisi açısından da olağanüstü bir performans göstermiştir.

* **AUC skoru:** 0.993

Bu değer, modelin sınıflar arasında neredeyse mükemmel bir ayrım yapabildiğini göstermektedir. Bu seviye bir AUC, pratikte üst düzey sınıflandırma performansına işaret eder.

A graph of a function

AI-generated content may be incorrect.

**Parametre Ayarları (Caret ile)**  
Modelin en iyi sonuç verdiği parametreler aşağıdaki gibidir:

* **mfinal:** 100 (toplam zayıf öğrenici sayısı)
* **maxdepth:** 3 (her bir ağacın maksimum derinliği)
* **coeflearn:** Freund (klasik AdaBoost öğrenme kuralı)

Bu parametrelerle kurulan model, test verisinde güçlü bir genelleme yeteneği göstermiştir.

**Sonuç**  
AdaBoost, hatalardan ders çıkararak öğrenme prensibi sayesinde hem doğruluk hem de AUC açısından en yüksek performans gösteren modellerden biri olmuştur. Dengeli sınıflandırma yeteneği ve düşük hata oranı ile uygulamada güvenilir bir alternatif olarak değerlendirilebilir.

**PCA Sonrası AdaBoost Model Performansı ve Değerlendirme**

PCA sonrası boyut indirgeme uygulanmış veri ile kurulan AdaBoost modeli, sınıflandırma performansı açısından tatmin edici sonuçlar vermiştir. Modelin doğruluk (accuracy) oranı yaklaşık %76 civarındadır ve pozitif sınıfı tanıma (duyarlılık) oranı oldukça yüksektir.

* **Accuracy (Doğruluk):** %76.47 (Modelin genel doğruluk oranı)
* **Sensitivity (Duyarlılık):** %90.65 (Pozitif sınıfın doğru tespiti)
* **Specificity (Özgüllük):** %52.38 (Negatif sınıfın doğru tespiti)
* **Kappa:** 0.4604 (Modelin rasgele sınıflandırmadan daha anlamlı şekilde başarılı olduğunu gösterir)
* **AUC:** 0.73 (Modelin sınıflar arasındaki ayrım gücünün güçlü olduğunu gösterir)

Modelde 0.5 eşik değeriyle sınıflandırma yapılmıştır. En iyi parametre kombinasyonu (mfinal = 100, maxdepth = 3, coeflearn = "Zhu") **ROC değerini maksimize edecek şekilde çapraz doğrulama ile seçilmiştir**. PCA sonrası sadeleşmiş veri yapısına rağmen AdaBoost modeli, sınıflar arası ayrımda yüksek başarı sağlamış ve pratikte kullanılabilir düzeyde bir performans göstermiştir.

A graph of a number of patients with a number of patients with a number of patients with a number of patients with a number of patients with a number of patients with a number of patients with a

AI-generated content may be incorrect.

**PCA Öncesi ve Sonrası AdaBoost Karşılaştırması**

PCA öncesi modelde yüksek boyutlu veri nedeniyle aşırı öğrenme (overfitting) riski olabilir. Boyut indirgeme sayesinde model daha sade hale gelirken, AdaBoost’un topluluk (ensemble) yapısı sayesinde bilgi kaybı performansa ciddi bir zarar vermemiştir. Sonuç olarak, PCA sonrası AdaBoost modeli yüksek duyarlılık ve makul özgüllükle en iyi sonucu veren modellerden biri olmuştur.

**Sonuç**

PCA sonrası AdaBoost modeli, sadeleştirilmiş veri yapısıyla bile yüksek sınıflandırma başarısı göstermiştir. Model, özellikle pozitif sınıfın tahmininde güçlüdür. Bu nedenle PCA sonrası veri için AdaBoost, diğer yöntemlere kıyasla daha avantajlı bir seçenek olabilir. Ancak yine de modelin parametreleri ve PCA bileşen sayısı optimize edilerek performans daha da iyileştirilebilir.

**Gradient Boosting Model Performansı ve Değerlendirme**  
Gradient Boosting modeli, oldukça yüksek bir sınıflandırma başarısı göstermiştir. Hem benign hem malignant sınıfları başarıyla ayırt edebilmiştir.

* **Doğruluk (Accuracy):** %97.06
* **Duyarlılık (Sensitivity):** %98.13
* **Özgüllük (Specificity):** %95.24
* **Kappa:** 0.9367
* **Balanced Accuracy:** %96.68
* **Pozitif Sınıf:** B

Model, her iki sınıfı da yüksek doğrulukla sınıflandırmış; özellikle benign (B) sınıfı neredeyse hatasız şekilde tespit edilmiştir.

**ROC Eğrisi ve AUC Skoru**  
Modelin sınıflandırma başarımı ROC eğrisiyle de desteklenmiştir.

* **AUC skoru:** 0.991

Bu skor, modelin sınıflar arasındaki ayrımı neredeyse kusursuz yaptığını göstermektedir.

A graph of a graph

AI-generated content may be incorrect.

**En İyi Parametreler (Caret ile Ayarlanmış):**

* **n.trees:** 200
* **interaction.depth:** 5
* **shrinkage:** 0.1
* **n.minobsinnode:** 10

Bu parametreler, çapraz doğrulama ile belirlenmiş ve modelin genel doğruluğunu maksimize etmiştir.

**Sonuç**  
Gradient Boosting modeli, hem doğruluk hem AUC açısından çok başarılı sonuçlar vermiştir. Boosting ailesinin güçlü bir temsilcisi olarak, hem dengeli hem de etkili bir sınıflandırma sunmuştur. Gerçek dünya verileri üzerinde yüksek performans gösteren bu model, uygulamada güvenle kullanılabilir.

**PCA Sonrası Gradient Boosting Model Performansı ve Değerlendirme**

PCA ile boyut indirgeme uygulandıktan sonra kurulan **Gradient Boosting** modeli, sınıflandırma performansında orta düzeyde bir başarı göstermiştir. Modelin doğruluk oranı yaklaşık %60 civarında olup, pozitif sınıfı tanıma gücü yüksek fakat negatif sınıfı ayırt etme gücü sınırlıdır.

* **Accuracy (Doğruluk):** %60.59 (Modelin genel doğruluk oranı)
* **Sensitivity (Duyarlılık):** %65.42 (Pozitif sınıfın doğru tespiti)
* **Specificity (Özgüllük):** %52.38 (Negatif sınıfın doğru tespiti)
* **Kappa:** 0.174 (Modelin rastgele sınıflandırmaya göre bir miktar daha iyi olduğunu gösterir)
* **AUC:** 0.63 (Modelin sınıflar arasındaki ayrım gücünün orta düzeyde olduğunu gösterir)

Modelde 0.5 eşik değeriyle sınıflandırma yapılmıştır. En iyi parametre kombinasyonu (n.trees = 200, interaction.depth = 1, shrinkage = 0.1, n.minobsinnode = 10) **ROC değerini maksimize edecek şekilde 5 katlı çapraz doğrulama** ile seçilmiştir. Model, aşırı karmaşık olmadan sınıflar arası ayrımı makul düzeyde yapabilmiştir.

A graph of a graph

AI-generated content may be incorrect.

**PCA Öncesi ve Sonrası Gradient Boosting Karşılaştırması**

PCA sonrası veri, boyut olarak sadeleşmesine rağmen Gradient Boosting modeli bilgi kaybı nedeniyle sınırlı performans sergilemiştir. Özellikle modelin özgüllüğü düşük kalmış, yani negatif sınıfı ayırt etmede zorlanmıştır. PCA öncesinde daha yüksek boyutlu veride modelin daha fazla varyans yakalama potansiyeli olduğundan, doğrudan karşılaştırma yapılırken dikkatli olunmalıdır. Yine de model, aşırı öğrenme riskini azaltarak genel performansta istikrar sunmuştur.

**Sonuç**

Gradient Boosting modeli, PCA sonrası veri üzerinde sade ama makul düzeyde bir performans göstermiştir. Sınıflar arası ayrım gücü lojistik regresyon ve rastgele orman modellerinden daha iyi, ancak AdaBoost modelinden düşüktür. Bu bağlamda, AdaBoost gibi daha güçlü ensembıl yöntemleri tercih etmek, sınıflandırma başarısını artırmak açısından daha uygun bir strateji olabilir.

**XGBoost Model Performansı ve Değerlendirme**  
XGBoost modeli, veri setinde oldukça başarılı bir sınıflandırma yapmıştır.

* **Doğruluk (Accuracy):** %94.12
* **Duyarlılık (Sensitivity):** %93.46
* **Özgüllük (Specificity):** %95.24
* **Kappa:** 0.8755
* **Balanced Accuracy:** %94.35
* **Pozitif Sınıf:** B

Model, özellikle her iki sınıfı da dengeli şekilde ayırmakta başarılıdır. Yanlış sınıflandırmalar düşük seviyededir.

**ROC Eğrisi ve AUC Skoru**  
Modelin ayrım gücü ROC eğrisi ile doğrulanmıştır.

* **AUC skoru:** 0.988

Bu yüksek AUC değeri, modelin hem pozitif hem negatif sınıfları çok iyi ayırt ettiğini göstermektedir.

**A graph of a graph

AI-generated content may be incorrect.**

**Model Parametreleri**

* booster = "gbtree"
* objective = "binary:logistic"
* eval\_metric = "auc"
* eta = 0.1 (learning rate)
* max\_depth = 6
* min\_child\_weight = 1
* subsample = 0.8
* colsample\_bytree = 0.8
* gamma = 0

Bu parametreler, modelin performansını optimize etmek için belirlenmiştir.

**Sonuç**  
XGBoost, yüksek doğruluk ve güçlü genelleme kabiliyetiyle güçlü bir sınıflandırma modeli olarak öne çıkmaktadır. Özellikle ROC AUC değeri ile sınıf ayrımındaki başarısını kanıtlamış, veri setinde dengeli ve güvenilir sonuçlar vermiştir.

**PCA Sonrası XGBoost Model Performansı ve Değerlendirme**

PCA ile boyut indirgeme sonrası kurulan **XGBoost** modeli, yüksek duyarlılık ve dengeli sınıflandırma başarısı ile güçlü bir performans göstermiştir. Modelin doğruluk oranı yaklaşık %75 olup, pozitif sınıfı ayırt etmede çok başarılıdır.

* **Accuracy (Doğruluk):** %75.29 (Modelin genel doğruluk oranı)
* **Sensitivity (Duyarlılık):** %90.65 (Pozitif sınıfın doğru tespiti)
* **Specificity (Özgüllük):** %49.21 (Negatif sınıfın doğru tespiti)
* **Kappa:** 0.429 (Modelin rastgele sınıflandırmaya göre belirgin şekilde daha iyi olduğunu gösterir)
* **AUC:** 0.69 (Modelin sınıflar arası ayrım gücünün iyi seviyede olduğunu gösterir)

Model, 0.5 eşik değeriyle sınıflandırma yapmıştır. **Parametreler** (eta = 0.1, max\_depth = 6, min\_child\_weight = 1, subsample = 0.8, colsample\_bytree = 0.8, gamma = 0) modelin AUC skorunu en çok artıran yapı olarak seçilmiş ve **early stopping** yöntemiyle en iyi 23. iterasyonda durmuştur.

**A graph of a graph

AI-generated content may be incorrect.**

**PCA Öncesi ve Sonrası XGBoost Karşılaştırması**

PCA sonrasında model daha sade hale gelmesine rağmen, XGBoost yüksek performans göstermeye devam etmiştir. Bu durum, algoritmanın az sayıda özelliğe rağmen karmaşık ilişkileri yakalayabildiğini göstermektedir. Ancak özgüllüğün sınırlı kalması, modelin negatif sınıfı ayırt etmede zorlandığını da ortaya koymaktadır.

**Sonuç**

XGBoost modeli, PCA sonrası veri ile hem doğruluk hem de AUC açısından tatmin edici bir performans sergilemiştir. Model, özellikle pozitif sınıfı ayırt etmede güçlü olup, AdaBoost’un ardından en yüksek başarıyı sağlayan yöntemlerden biri olmuştur. Negatif sınıf üzerindeki başarısını artırmak için sınıf ağırlıkları veya eşik ayarları yeniden düzenlenebilir.

**5. MODEL PERFORMANSLARININ KARŞILAŞTIRILMASI**

**5.1. Doğruluk, ROC-AUC Ölçümleri**

Bu çalışmada, PCA öncesi ve sonrası olmak üzere yedi farklı sınıflandırma algoritması (KNN, SVM, Lojistik Regresyon, Random Forest, AdaBoost, Gradient Boosting, XGBoost) değerlendirilmiş; her birinin başarımı doğruluk (Accuracy), ROC-AUC ve diğer temel ölçütlerle analiz edilmiştir. Aşağıda bu sonuçların genel değerlendirmesi yer almaktadır.

| **Model** | **PCA Öncesi Accuracy** | **PCA Öncesi AUC** | **PCA Sonrası Accuracy** | **PCA Sonrası AUC** |
| --- | --- | --- | --- | --- |
| **KNN** | **%95.29** | **0.982** | **%87.06** | **0.935** |
| **SVM** | **%99.41** | **0.9988** | **%55.29** | **0.562** |
| **Lojistik Reg.** | **%94.12** | **0.9749** | **%54.12** | **0.57** |
| **Random Forest** | **%95.88** | **0.989** | **%55.12** | **0.627** |
| **AdaBoost** | **%96.47** | **0.993** | **%76.47** | **0.734** |
| **Gradient Boosting** | **%97.06** | **0.991** | **%60.59** | **0.629** |
| **XGBoost** | **%94.12** | **0.988** | **%75.29** | **0.694** |

**PCA Sonrası Performans Değerlendirmesi**

PCA uygulandıktan sonra, boyut indirgeme sonucunda modellerin başarılarında genel olarak düşüş gözlemlenmiştir. Bu durum, PCA’nın bazı önemli değişken bilgilerini azaltarak modellerin ayırt ediciliğini düşürdüğünü göstermektedir.

Özellikle:

* SVM ve Lojistik Regresyon modelleri PCA sonrası ciddi performans kaybına uğramıştır. Bu durum, bu modellerin orijinal özellikler üzerinde daha iyi çalıştığını ortaya koymaktadır.
* AdaBoost ve XGBoost ise PCA’ya karşı daha dirençli davranmış ve %75 üzeri doğruluk değerlerini koruyabilmiştir. Bu iki model, düşük boyutlu veriyle dahi öğrenme kapasitesini koruyabilmektedir.

**5.2. En Başarılı Modelin Seçimi**

* **SVM**, PCA öncesinde %99.41 doğruluk ve 0.9988 AUC değeri ile en yüksek başarıyı göstermiştir. Ancak PCA sonrası performansı ciddi şekilde düşmüş (Accuracy: %55.29, AUC: 0.562), bu da modelin boyut indirgeme işlemlerine karşı duyarlılığını ortaya koymuştur.
* **AdaBoost**, PCA öncesinde yüksek performans (Accuracy: %96.47, AUC: 0.993) sergilemiş, PCA sonrasında ise doğruluk oranı %76.47’ye, AUC değeri ise 0.734’e düşmüştür. Performansını büyük ölçüde korumuş olsa da, PCA sonrası en iyi başarıyı göstermemiştir.
* **KNN**, PCA öncesi %95.29 doğruluk ve 0.982 AUC değeri ile iyi bir performans göstermiştir. Ancak özellikle **PCA sonrası en yüksek doğruluğa (%87.06)** ve oldukça yüksek bir AUC değerine (0.935) ulaşarak, **boyut indirgeme sonrasında en başarılı model olmuştur**. Bu durum, KNN'nin düşük boyutlu veri temsilinde dahi güçlü genelleme yapabildiğini göstermektedir.
* **XGBoost** ve **Gradient Boosting** modelleri de genel olarak iyi sonuçlar vermiştir. Ancak PCA sonrası AdaBoost ile benzer şekilde performans kayıpları yaşamışlardır.

Bütün bu karşılaştırmalar ışığında, **KNN algoritması en başarılı model olduğuna karar verdik.**  
Bunun nedenleri:

* PCA öncesi zaten yüksek olan doğruluk ve AUC değerlerinin, PCA sonrası da en yüksek seviyelerde kalması,
* Boyut indirgeme gibi ön işleme adımlarından **diğer modellere kıyasla daha az etkilenmesi**,
* **Hem doğruluk hem de AUC açısından PCA sonrası açık ara en yüksek değerlere sahip olması**,
* Genel olarak **kararlı, dayanıklı ve veri indirgeme tekniklerine toleranslı** bir yapı sergilemesi.

Bu yönüyle KNN modeli, yalnızca bu çalışma özelinde değil, benzer sınıflandırma problemlerinde de **gerçek dünya uygulamaları için etkili, güvenilir ve uygulanabilir** bir seçenek olarak öne çıkmaktadır.

**6. SONUÇ VE ÖNERİLER**

Bu çalışmada, meme kanseri verisi kullanılarak çeşitli makine öğrenmesi algoritmalarının sınıflandırma başarısı değerlendirilmiştir. İlk aşamada orijinal veriler üzerinde, ardından boyut indirgeme tekniği olarak Temel Bileşenler Analizi (PCA) uygulandıktan sonra modeller yeniden eğitilerek karşılaştırmalar yapılmıştır. Accuracy ve ROC-AUC gibi temel performans ölçütlerine göre modellerin başarıları analiz edilmiştir.

PCA uygulanmadan önce SVM modeli %99.41 doğruluk ve 0.9988 AUC ile en yüksek başarıyı gösteren model olmuştur. Ancak PCA uygulandıktan sonra bu başarı ciddi oranda düşmüş; doğruluk oranı %55.29’a, AUC ise 0.562’ye gerilemiştir. Bu durum, SVM modelinin veri boyutunun indirgenmesinden olumsuz etkilendiğini göstermektedir. Diğer yandan, **K-Nearest Neighbors (KNN) algoritması PCA öncesi%95.29 doğruluk ve 0.982 AUC değeri sonrasında da %87.06 doğruluk ve 0.935 AUC değeri ile en iyi performansı sergilemiştir**. Bu nedenle, PCA sonrası en başarılı model olarak KNN öne çıkmıştır. Bu sonuç, KNN algoritmasının daha düşük boyutlu temsillerle bile güçlü genelleme yapabildiğini ve PCA ile bilgi kaybından görece daha az etkilendiğini ortaya koymaktadır.

Model performanslarına bakıldığında, boyut indirgeme işleminin her model için olumlu sonuçlar doğurmadığı, bazı modellerde ciddi performans kayıplarına neden olduğu görülmektedir. Bu nedenle, PCA gibi tekniklerin uygulanmadan önce veri yapısına ve modelin doğasına uygunluğu dikkatle değerlendirilmelidir.

Çalışmadan elde edilen sonuçlara dayanarak öneriler şunlardır:

* Gerçek dünyada meme kanseri teşhis sistemlerinde KNN algoritması gibi basit ama etkili modeller, özellikle ön işleme sonrası güçlü bir alternatif olabilir.
* Boyut indirgeme gibi tekniklerin mutlaka model bazlı olarak test edilmesi ve etkilerinin dikkatlice analiz edilmesi gereklidir.
* Veri kümesine daha fazla örnek eklemek, özellikle azınlık sınıflar için denge sağlamak, modellerin genelleme kabiliyetini artırabilir.
* Son olarak, bu çalışma yalnızca sınıflandırma başarısına odaklanmıştır; gelecekte farklı veri türleriyle (örneğin genetik veriler veya görüntüler) yapılan çoklu modelleme yaklaşımları, teşhis doğruluğunu daha da artırabilir.

Bu analizler, veri ön işleme ve algoritma seçiminin model performansı üzerindeki etkilerini ortaya koyarak, sağlık alanında daha doğru ve güvenilir makine öğrenmesi uygulamaları geliştirilmesine katkı sağlamayı hedeflemektedir.

**KAYNAKÇA**

An Introduction to Statistical Learning with Applications in R -Second Edition Gareth James , Daniela Witten , Trevor Hastie , Robert Tibshirani

Machine Learning with R, the tidyverse, and mlr HEFIN I. RHYS

[Principal Component Analysis in R: prcomp vs princomp - Articles - STHDA](https://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/#google_vignette)

[**https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data**](https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data)