Breast Cancer Wisconsin (Diagnostic)

Göğüs Tümörü

Süleyman Yalçın

[suleyman.yalcin@gazi.edu.tr](mailto:suleyman.yalcin@gazi.edu.tr)

Abdulgaffar Gülice

[abdulgaffar.gulice@gazi.edu.tr](mailto:abdulgaffar.gulice@gazi.edu.tr)

Harun Reşit Mercan

[hresit.mercan@gazi.edu.tr](mailto:hresit.mercan@gazi.edu.tr)



**ÖZET**Bu çalışmada, meme dokusundaki tümörlerin malign (kötü huylu) ya da benign (iyi huylu) olarak sınıflandırılmasına yönelik bir analiz gerçekleştirilmiştir. Çalışmada, Breast Cancer Wisconsin (Diagnostic) veri seti kullanılmış ve toplamda sekiz farklı makine öğrenmesi algoritması değerlendirilmiştir. Veri seti, tümörlerin çeşitli morfolojik özelliklerini temsil eden 30 sayısal değişkenden oluşan 569 gözlem içermektedir. Ön işleme sürecinde yüksek korelasyonlu değişkenler çıkarılmış, veriler normalize edilmiş ve eğitim-test ayrımı yapılmıştır. Ardından; Karar Ağacı, Rastgele Orman, K-En Yakın Komşu, Naive Bayes, Lojistik Regresyon, XGBoost, Destek Vektör Makineleri (SVM) ve Yapay Sinir Ağı modelleri uygulanmıştır. Her bir model doğruluk, kesinlik, duyarlılık ve F1 skoru gibi performans metrikleriyle analiz edilmiştir. Elde edilen sonuçlar, özellikle XGBoost ve Yapay Sinir Ağı algoritmalarının yüksek doğruluk oranları ile öne çıktığını göstermektedir.

**Anahtar Kelimeler:** Meme Kanseri, Makine Öğrenmesi, Sınıflandırma, Tanı, Yapay Zekâ

**GİRİŞ**  
Meme dokusunda meydana gelen tümöral oluşumlar, özellikle kadınlar arasında yaygın görülen sağlık problemleri arasında yer almaktadır. Dünya genelinde kadınlarda en sık teşhis edilen kanser türlerinden biri olan meme kanseri, meme dokularında kötü huylu (kanserli) hücrelerin oluşmasıyla ortaya çıkan kanser türüdür. En yaygın meme kanseri türü, kanalların hücrelerinde başlayan duktal karsinomdur. Loblarda veya lobüllerde başlayan kanser lobüler karsinom olarak adlandırılırken, bu türün diğer meme kanseri türlerine göre her iki memede görülme sıklığı daha fazladır. Meme kanserinin en tipik belirtisi memede kitle veya meme dokusu şeklinde değişikliktir. Memede ele gelen kitle, meme cildi ve ucunda farklılıklar, meme başında akıntı gelmesi ve meme dokusu ve meme ucu şeklinde değişiklik ve içe çökme meme kanseri belirtileridir. Erken teşhiste tedavi edilebilen bir kanser olan meme kanseri, mamografi yöntemiyle teşhis edilir ve daha sonra tedavi sürecine geçilir.

Erken evrede tespit edildiğinde yüksek tedavi başarısı sunmaktadır. Bu bağlamda, erken tanı süreçlerinin geliştirilmesi hem hasta sağkalımı açısından kritik öneme sahiptir hem de sağlık sistemine olan yükün azaltılmasına katkı sağlamaktadır.

Son yıllarda, tıbbi tanı süreçlerinin daha hızlı, doğru ve otomatik hale getirilmesi amacıyla yapay zekâ ve makine öğrenmesi tekniklerinin kullanımı giderek artmaktadır. Makine öğrenmesi, geçmiş verilere dayanarak örüntüleri tanıma ve tahminlerde bulunma yeteneğine sahip algoritmalar bütünü olarak tanımlanabilir. Bu yöntemler, tıbbi görüntüleme, laboratuvar sonuçları ve hasta geçmişi gibi farklı veri kaynaklarını analiz ederek hekime karar destek sağlayabilecek niteliktedir.

Meme kanseri gibi karmaşık yapıya sahip hastalıkların tanısında, veriye dayalı sınıflandırma yaklaşımları öznellikten kaynaklanan hataları azaltarak karar süreçlerinin doğruluğunu artırabilmektedir. Bu çalışmada, Breast Cancer Wisconsin (Diagnostic) veri seti kullanılarak meme dokusundaki anormal yapıların malign ya da benign olup olmadığını belirlemeye yönelik çeşitli makine öğrenmesi algoritmaları uygulanmıştır. Algoritmaların başarımı doğruluk, kesinlik, duyarlılık ve F1 skoru gibi performans metrikleriyle değerlendirilmiş ve karşılaştırılmıştır. Elde edilen bulgular, hem literatüre katkı sağlamayı hem de gelecekte klinik karar destek sistemlerinin geliştirilmesine zemin oluşturmayı amaçlamaktadır.

**Literatür Taraması:**

### Özellik Seçimi ve Boyut İndirgeme Yöntemlerinin Model Performansına Etkisi

Makine öğrenmesi uygulamalarında model performansını artırmak, karmaşıklığı azaltmak ve genelleme yeteneğini geliştirmek amacıyla özellik seçimi ve boyut indirgeme yöntemlerine sıklıkla başvurulmaktadır. Özellikle çok sayıda özelliğe sahip veri setlerinde, aralarındaki yüksek korelasyon nedeniyle bazı değişkenlerin redundant hale gelmesi, modelin öğrenme sürecini olumsuz etkileyebilmektedir. Bu nedenle, yalnızca en bilgi verici değişkenlerin seçilmesi, modelin daha yalın ve daha etkili hale gelmesini sağlayabilmektedir.

Wisconsin Tanı Veri Seti örneğinde, 30 sayısal özelliğin bulunması ve bu özelliklerin bir kısmının birbirleriyle yüksek korelasyona sahip olması, boyut indirgeme ve özellik elemesi ihtiyacını gündeme getirmektedir. Bu doğrultuda, 2023 yılında **Juarto** tarafından gerçekleştirilen bir çalışmada, aykırı değerlerin temizlenmesi ve VIF (Variance Inflation Factor) analizine dayalı özellik eleme yöntemleri uygulanmıştır. Bu yöntemlerle yüksek korelasyona sahip değişkenlerin çıkarılması sonrasında, lojistik regresyon ve rastgele orman algoritmaları ile elde edilen doğruluk oranlarının sırasıyla %98.25’ten %99.12’ye yükseldiği rapor edilmiştir. Çalışmanın bulguları, multikorelasyonun azaltılmasının model performansı üzerinde doğrudan ve olumlu bir etkiye sahip olduğunu göstermektedir.

### Wisconsin Meme Kanseri Veri Seti Üzerine Son Yıllarda Yapılan Makine Öğrenmesi ve Derin Öğrenme Çalışmaları

2019 yılı sonrasında gerçekleştirilen akademik çalışmalar ve uygulamalı projelerde, *Breast Cancer Wisconsin (Diagnostic)* veri seti sıklıkla tercih edilen bir test ortamı haline gelmiştir. Bu veri seti üzerinde klasik denetimli öğrenme algoritmalarından lojistik regresyon, karar ağaçları, rastgele orman (Random Forest), destek vektör makineleri (Support Vector Machines, SVM), k-en yakın komşu (k-NN), Naive Bayes ve Boosting tabanlı yöntemler (örneğin, XGBoost, LightGBM, Gradient Boosting) yaygın olarak uygulanmıştır [1][2]. Birçok araştırma, bu algoritmaların birden fazlasını aynı veri seti üzerinde karşılaştırmalı biçimde değerlendirerek performans analizleri gerçekleştirmiştir.

Örneğin, bir çalışmada lojistik regresyon, k-NN, karar ağacı, rastgele orman, Gaussian Naive Bayes, SVM, LightGBM ve XGBoost algoritmaları aynı veri kümesi üzerinde eğitilmiş ve bu modellerin doğruluk, kesinlik (precision), geri çağırma (recall) gibi metrikler açısından karşılaştırılması yapılmıştır [2].

Bunun yanı sıra, yapay sinir ağlarına (YSA) dayalı yaklaşımlar da bu veri seti üzerinde uygulanmıştır. Küçük boyutlu bir veri seti olmasına rağmen bazı araştırmacılar, çok katmanlı algılayıcılar (Multilayer Perceptron, MLP) gibi derin öğrenme yapıları ile sınıflandırma görevlerini başarıyla gerçekleştirmiştir. Bu yaklaşımlarda, öznitelikler doğrudan sinir ağına girdi olarak verilmiş ve farklı mimarilerin başarımı incelenmiştir [2].

Daha ileri düzey çalışmalarda, veri setinin yeniden şekillendirilerek 1-boyutlu evrişimli sinir ağı (1D Convolutional Neural Network, CNN) uygulamaları da denenmiştir. Örneğin, 2024 yılında yürütülen bir projede, veri dengesizliği giderilmiş ve öznitelikler CNN mimarisiyle uyumlu hale getirilerek sınıflandırma başarımları ölçülmüştür [2].

Ansambl (ensemble) yöntemleri de literatürde dikkate değer bir yer edinmiştir. Nitekim, yakın tarihli bir çalışmada CWV-BANN-SVM adlı yöntemle, yapay sinir ağı ve destek vektör makinelerinin birleşiminden oluşan bir topluluk modeli geliştirilmiş ve bu modelin sınıflandırma performansı yüksek doğruluk oranlarıyla rapor edilmiştir [3].

Genel olarak, son yıllarda yapılan araştırmalar klasik denetimli öğrenme algoritmalarından derin öğrenme tabanlı modellere kadar geniş bir yöntem yelpazesi ile *Wisconsin Breast Cancer Diagnostic* veri setinin sınıflandırma problemlerinde etkin biçimde kullanılabileceğini ortaya koymuştur.

metin, çizgi, ekran görüntüsü, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

### Veri Ön İşleme Teknikleri (Preprocessing)

Breast Cancer Wisconsin (Diagnostic) veri setinde yüksek model performansı için kapsamlı ön işleme adımları uygulanmıştır:

* **Özellik Ölçekleme:** SVM, lojistik regresyon, k-NN gibi algoritmalar için standartlaştırma (z-score) veya min-max normalizasyon yaygındır. Bu, modellerin belirli özniteliklere aşırı duyarlılığını önler.
* **Eksik Veriler:** Diagnostic versiyonunda eksik değer yoktur. Ancak bazı çalışmalar, farklı sürümleri karıştırarak eksik kayıtları çıkarmıştır.
* **Sınıf Dengesizliği:** Veri setinde malign örnekler azınlıktadır (~%37). Çoğu model bu dağılımla çalışsa da bazı projeler **SMOTE** gibi yöntemlerle sınıfları dengelemiştir.
* **Aykırı Değerler:** Alt/üst çeyrekler dışındaki veriler çıkarılarak model doğruluğu artırılmıştır (ör. doğruluk %98.25 → %99.12).
* **Veri Bölünmesi & Çapraz Doğrulama:** Yaygın olarak %70–80 eğitim / %30–20 test ayrımı veya **10-fold cross-validation** kullanılmıştır. Özellikle küçük veri setlerinde çapraz doğrulama tercih edilmiştir.
* **Özellik Dönüşümü:** Log veya karekök dönüşümleri, yeni öznitelik türetme (feature engineering) gibi adımlar uygulanmıştır. Korelasyonlu öznitelikler birleştirilebilmiştir.
* **Kodlama:** Hedef değişken (‘M’, ‘B’) genellikle sayısal olarak 1 ve 0’a dönüştürülmüştür.

### Kullanılan Araçlar ve Kütüphaneler

* **Python** en yaygın kullanılan dildir.  
  + **scikit-learn:** Klasik algoritmalar (logistic regression, SVM, k-NN, Random Forest vs.), veri bölme ve ölçekleme (train\_test\_split, StandardScaler) gibi işlemler için temel araçtır.
  + **Keras/TensorFlow:** Derin öğrenme modellerinde (MLP, CNN) kullanılmıştır. Özellikle Keras, model tanımlama ve eğitme sürecinde tercih edilmiştir.
  + **PyTorch:** Daha az tercih edilmiştir.
  + **XGBoost & LightGBM:** scikit-learn arayüzü ile yaygın olarak kullanılmıştır.
  + **pandas, numpy, matplotlib, seaborn:** Veri işleme ve görselleştirme için standart kütüphanelerdir.
* **MATLAB:** Özellikle bulanık mantık veya karar destek sistemleri gibi özel çalışmalar için kullanılmıştır.
* **R:** Daha çok istatistiksel analizlerde, caret veya temel paketlerle kullanılmıştır.

### Modellerin Doğruluk, F1 Skoru ve AUC Gibi Metriklerle Karşılaştırılması

Birçok çalışma, farklı algoritmaları değerlendirerek en iyi performansı veren modeli belirlemeye çalışmıştır. Genel olarak, bu veri setinde algoritmalar oldukça yüksek doğruluklara ulaşabilmektedir. Doğruluk (accuracy) genellikle %95’in üzerinde rapor edilmiştir ve en iyi modeller %99’a varan başarılar göstermektedir. Örneğin, 2019 tarihli bir çalışmada bahsedilen CWV-BANN-SVM ansamblesi %100 doğruluk bildirmiştir1. Yine aynı derlemede, Random Forest algoritması ile %99.42 doğruluk elde eden bir model ikinci sırada yer almıştır1. 10-kat çapraz doğrulama uygulayan bir başka çalışmada ise belli başlı 5 özellik kullanılarak bulanık mantık tabanlı bir modelle %99.40 doğruluk yakalanmıştır2.

Bu düzeyde doğruluklar, veri setinin görece kolay ayrılabilir olduğunu göstermektedir. Aşağıdaki tablo, seçilmiş bazı çalışmaların farklı algoritmalarla elde ettikleri en iyi sonuçları örnek olarak özetlemektedir:

#### 

#### 

#### Seçili Çalışmalarda Kullanılan Modellerin Performans Özeti

| **Çalışma (Yıl)** | **Algoritmalar** | **En İyi Model (Accuracy)** | **Diğer Metrikler** | **Notlar / Özel Durumlar** |
| --- | --- | --- | --- | --- |
| Abdar & Makarenkov (2019)1 | ANN + SVM ansamblesi | %100 | – | 50/50 eğitim-test, CV yok (overfitting olası) |
| Hernández-Julio et al. (2019)1 | Bulanık karar modeli (Mamdani) | %99.40 | – | 10-kat CV, seçilmiş 5 özellik, MATLAB ortamı |
| Juarto (2023)2 | Lojistik Regresyon, RF (eşit) | %99.12 | – | VIF ile özellik seçimi sonrası |
| Kılıç & Karakoyun (2023)3 | K-NN, (2.: SVM & LR) | %99.3 | F1: %99.4, Precision: %98.9 | K-NN en yüksek başarı, özenli ön işleme |
| Strelcenia & Prakoonwit (2023)4 | SVM, LR, K-NN vs. | %99.3 | – | SVM özellik mühendisliği ile en iyi performansı vermiş |
| Bonat (2024)5 | LR, SVM (eşit) | %99.12 | F1: ≈ %99 | 80/20 veri ayrımı, hiperparametre optimizasyonu yapılmış |

Tablodan da görüldüğü gibi, SVM, rastgele orman ve lojistik regresyon gibi yöntemler sıkça en üst sırada yer almıştır. Örneğin, 2023 yılında yapılan bir karşılaştırmalı çalışmada SVM modeli %99.3 doğruluk ile en başarılı model olurken, lojistik regresyon %98.06 ve k-NN %97.35 doğrulukla onu izlemiştir4. Diğer çalışmalarda da benzer şekilde SVM ve RF modelleri öne çıkmaktadır.

F1 skoru, hassasiyet (precision), duyarlılık (recall) gibi ölçütler de çoğu zaman %95–100 arasında oldukça yüksektir. Örneğin Kılıç & Karakoyun (2023) çalışmasında K-NN modeli %100 recall ve %99.4 F1 skoru ile en iyi sonuçları vermiştir3.

Sonuç olarak, Wisconsin (Diagnostic) veri setinde kullanılan modellerin çoğu %95’in üzerinde doğruluk sağlamış, özellikle SVM, RF, LR ve K-NN gibi algoritmalar sıkça en başarılılar arasında yer almıştır. Ancak hangi modelin en iyi olduğu, veri ön işleme ve özellik seçimi gibi faktörlere bağlı olarak değişebilmektedir.

**MATERYAL VE YÖNTEM** Bu çalışmada, sınıflandırma problemi kapsamında UCI Machine Learning Repository üzerinde yayımlanan **Breast Cancer Wisconsin (Diagnostic)** veri seti kullanılmıştır. Veri seti, her biri tümörün morfolojik yapısını temsil eden 30 sayısal öznitelik içeren 569 gözlemden oluşmaktadır. Hedef değişken, tümörlerin malign (kötü huylu) ya da benign (iyi huylu) olarak sınıflandırılmasını temsil etmektedir.

Analiz öncesinde veriye çeşitli ön işleme adımları uygulanmıştır. İlk olarak, %90’dan yüksek korelasyon gösteren öznitelikler veri setinden çıkarılmış ve veriler standartlaştırılarak normalize edilmiştir. Ardından veri seti %80 eğitim, %20 test şeklinde bölünmüştür.

Çalışma kapsamında aşağıdaki sekiz makine öğrenmesi algoritması değerlendirilmiştir:

* Karar Ağacı (Decision Tree)
* Rastgele Orman (Random Forest)
* Lojistik Regresyon (Logistic Regression)
* K-En Yakın Komşu (K-Nearest Neighbors, KNN)
* Naive Bayes
* XGBoost
* Destek Vektör Makineleri (Support Vector Machines, SVM)
* Yapay Sinir Ağı (Artificial Neural Network)

Modellerin başarımı; doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru gibi yaygın sınıflandırma metrikleri ile değerlendirilmiş ve elde edilen sonuçlar karşılaştırmalı olarak analiz edilmiştir. **Model Performans Değerlendirme Ölçütleri**

Sınıflandırma modelinin performansını ölçütlerken karışıklık matrisindeki değerler, AUC değeri, F1 Skoru ölçütlerinin açıklamaları:

**Doğruluk (Accuracy)**

Modelin doğru tahmin ettiği tüm örneklerin toplam örneklere oranıdır.

**Kesinlik (Precision)**

Modelin pozitif olarak tahmin ettiği örneklerin ne kadarının gerçekten pozitif olduğunun oranıdır.

****

**Duyarlılık(Recall)**

Pozitif sınıfın doğru bir şekilde tanınma oranıdır ve gerçek pozitiflerin doğru tahmin edilme yüzdesini gösterir.

****

**F1 Skoru (F1 Score)**

Hassasiyet ve duyarlılığın harmonik ortalamasıdır ve dengesiz veri setlerinde modelin genel performansını ölçmek için kullanılır.

****

### 2.2.1. Decision Tree (Karar Ağaçları)

### Karar ağaçları, veriyi dallara ayırarak sınıflandırma yapan, görsel olarak yorumlanabilir ve sezgisel açıdan anlaşılır modellerdir. Her düğüm, bir özniteliğe ve belirli bir eşik değerine göre veri bölünmesini temsil ederken, yaprak düğümler sınıf etiketlerini içerir. Bu yapıda model, her adımda en yüksek bilgi kazancı sağlayan değişkeni seçerek ilerler. Bölme işlemleri genellikle Gini İndeksi veya Entropi gibi kriterlerle gerçekleştirilir.

### Uygulama: Model, eğitim verisi üzerinde oluşturulmuş ve test verisi ile doğrulama yapılmıştır. Elde edilen doğruluk oranı %92.08’dir. Modelin başarımını artırmak amacıyla etkisiz öznitelikler çıkarılmış ve yeniden eğitilmiştir. Performans değerlendirmesinde F1 skoru, precision, recall ve destek (support) değerleri analiz edilmiştir. Sonuçlar, modelin aşırı öğrenme (overfitting) eğilimi göstermediğini ortaya koymuştur. ROC-AUC skoru ise yaklaşık 0.92 olarak hesaplanmıştır.

### 2.2.2. Random Forest (Rastgele Orman)

### Random Forest algoritması, çok sayıda karar ağacının topluluk (ensemble) yapısı içinde bir araya getirilmesiyle oluşturulur. Her ağaç, rastgele seçilen alt örneklemler üzerinde eğitilir ve nihai sınıflandırma, çoğunluk oylaması ile belirlenir. Bu yapı, modelin aşırı öğrenmeye karşı daha dayanıklı olmasını sağlar.

### Uygulama: Model, 100’den fazla karar ağacı ile eğitilmiş ve test verisi üzerinde %96.49 doğruluk oranı elde etmiştir. ROC eğrisinin altında kalan alan (AUC) değeri 1’e oldukça yakın hesaplanmıştır. İlk aşamada overfitting riski dikkate alınmış, ancak yapılan analizler modelin genelleme kapasitesinin yüksek olduğunu göstermiştir.

### 2.2.3. Naive Bayes (Naif Bayes Sınıflandırıcısı)

### Naive Bayes, Bayes teoremi temelli ve değişkenlerin birbirinden bağımsız olduğu varsayımına dayanan bir sınıflandırma algoritmasıdır. Basit yapısı ve düşük hesaplama maliyeti ile dikkat çeker. Ancak, öznitelikler arasında korelasyon bulunduğunda performansı düşebilir.

### Uygulama: Model, eğitim-test ayrımı yapıldıktan sonra Gaussian Naive Bayes yöntemiyle uygulanmıştır. Elde edilen doğruluk oranı %92.98’dir. Hızlı uygulanabilirliği avantaj sağlarken, değişken bağımlılığı nedeniyle diğer yöntemlere kıyasla daha düşük başarı göstermiştir.

### 2.2.4. K-Nearest Neighbors (k-En Yakın Komşu

### KNN algoritması, test örneklerinin sınıfını, eğitim verisindeki en yakın “k” komşuya göre belirler. Sınıflandırma işlemi, genellikle Öklidyen mesafesi temel alınarak yapılır. Parametre olarak seçilen “k” değeri model başarımını doğrudan etkiler.

### Uygulama: Çeşitli k değerleri denenmiş ve en iyi sonuç k=7 için elde edilmiştir. Mesafe hesaplamalarında Öklidyen metriği kullanılmıştır. Model, test verisinde %95.61 doğruluk ve yaklaşık 0.98 AUC değeriyle yüksek performans göstermiştir.

**2.2.5. Logistic Regression (Lojistik Regresyon)**

Lojistik regresyon, özellikle ikili sınıflandırma problemlerinde yaygın olarak kullanılan, doğrusal karar sınırları oluşturan bir modeldir. Girdi değişkenleri ile hedef değişken arasındaki ilişkiyi olasılık temelli olarak modellendirir. Model çıktısı, sigmoid fonksiyonu ile 0 ile 1 arasında bir olasılık değeri olarak üretilir.

**Uygulama:**  
Bu çalışmada lojistik regresyon modeli, tümörün malign veya benign olup olmadığını tahmin etmek üzere eğitilmiştir. Elde edilen doğruluk oranı %97.36 olup, ROC eğrisi altında kalan alan (AUC) 0.98 olarak hesaplanmıştır. Ayrıca, modelin öznitelik katsayıları incelenerek her bir değişkenin sınıflandırmaya katkısı yorumlanabilir hale getirilmiştir. Bu yönüyle model, hem yüksek performans hem de açıklanabilirlik sağlamaktadır.

**2.2.6. Linear Regression (Doğrusal Regresyon)**

Doğrusal regresyon temelde sürekli değişkenleri tahmin etmek için geliştirilmiş bir yöntemdir. Ancak bu çalışmada, ikili sınıflandırma bağlamında doğrusal sınırlar oluşturmak amacıyla uyarlanarak kullanılmıştır. Bu durumda model çıktıları belirli bir eşik değerine göre sınıflandırmaya dönüştürülmüştür.

**Uygulama:**  
Modelin çıktıları round() fonksiyonu ile 0 veya 1 olacak şekilde ikili sınıflara dönüştürülmüş ve performans karışıklık matrisi üzerinden değerlendirilmiştir. Model, doğrusal sınırlara dayandığı için özellikle karmaşık örneklerde sınırlı başarı göstermiştir. Elde edilen doğruluk oranı %92.74, AUC değeri ise yaklaşık 0.93’tür.

**2.2.7. Support Vector Machines (Destek Vektör Makineleri)**

SVM, sınıflar arasındaki ayrımı en iyi şekilde sağlayan hiper düzlemi (decision boundary) bulmayı hedefleyen güçlü bir sınıflandırma yöntemidir. Doğrusal olmayan sınırlarda dahi başarılı sonuçlar verebilmesi için çekirdek (kernel) fonksiyonları kullanılabilir.

**Uygulama:**  
Modelde RBF (Radial Basis Function) kernel tercih edilmiş ve hiperparametreler (C, Gamma) GridSearchCV yöntemi ile optimize edilmiştir. Ayrıca, SMOTE yöntemi uygulanarak sınıf dengesizliği giderilmeye çalışılmıştır. Bu optimizasyonlar sonucunda modelin doğruluk oranı %95.61’den %98.24’e yükselmiştir. Yüksek AUC ve F1 skoru değerleri de modelin güçlü sınıflandırma yeteneğini desteklemektedir.

**2.2.8. XGBoost**

XGBoost, gradient boosting algoritmasına dayalı, karar ağaçlarını ardışık şekilde birleştiren güçlü bir sınıflandırma yöntemidir. Hem doğrusal hem de doğrusal olmayan desenleri modelleyebilme yeteneği ile yüksek başarı sağlar.

**Uygulama:**  
Model, eğitim verisi üzerinde eğitilmiş ve %95.61 doğruluk oranı elde etmiştir. ROC-AUC değeri oldukça yüksektir. Modelin başarımında hiperparametre ayarlamalarının etkili olduğu gözlemlenmiştir. XGBoost, genellikle veri karmaşıklığı yüksek senaryolarda tercih edilen etkili bir yöntemdir.

**2.2.9. Artificial Neural Network (Yapay Sinir Ağı)**

Yapay sinir ağları (YSA), insan beyninden esinlenen çok katmanlı mimariler aracılığıyla verilerdeki karmaşık ilişkileri öğrenebilen güçlü modellerdir. Giriş katmanı, gizli katmanlar ve çıkış katmanından oluşan bu yapılar, sınıflandırma görevlerinde yüksek doğruluk oranlarına ulaşabilir.

**Uygulama:**  
Model, iki gizli katman içerecek şekilde yapılandırılmış ve sırasıyla ReLU ve sigmoid aktivasyon fonksiyonları kullanılmıştır. Çıkış katmanında yine sigmoid fonksiyonu ile olasılık tahmini yapılmış, eşik değeri olarak 0.45 kullanılmıştır. Eğitim sürecinde binary\_crossentropy kayıp fonksiyonu ve Adam optimizasyon algoritması tercih edilmiştir. Aşırı öğrenmeyi önlemek için EarlyStopping mekanizması devreye alınmıştır.

Model, test verisinde %98.83 doğruluk, F1 skoru 0.9844 ve ROC-AUC değeri 0.9988 ile en yüksek başarıyı göstermiştir. Ayrıca farklı hücre sayıları, katman yapıları ve aktivasyon fonksiyonları denenerek en iyi model kombinasyonu elde edilmiştir :

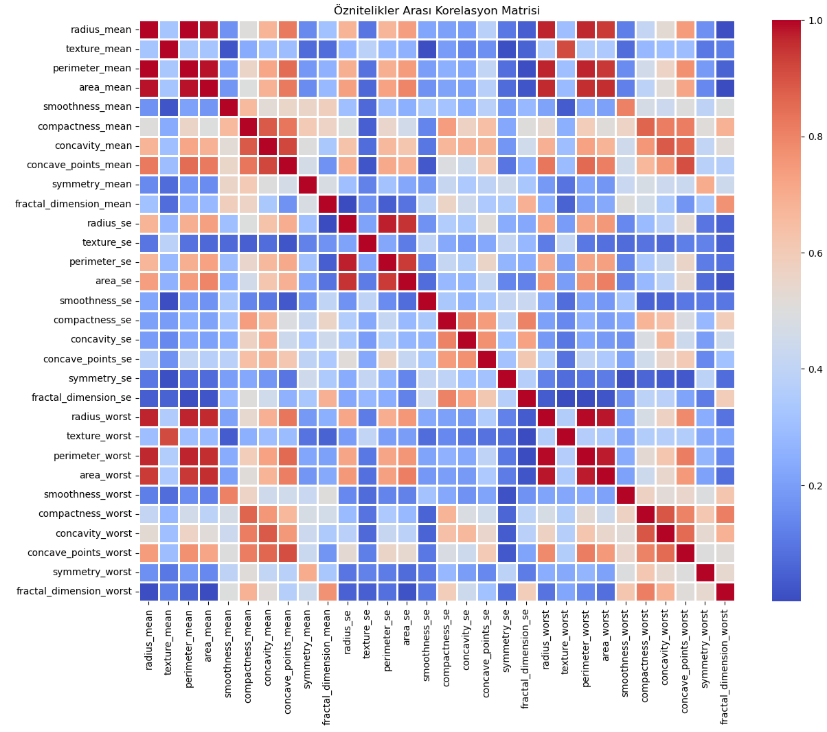
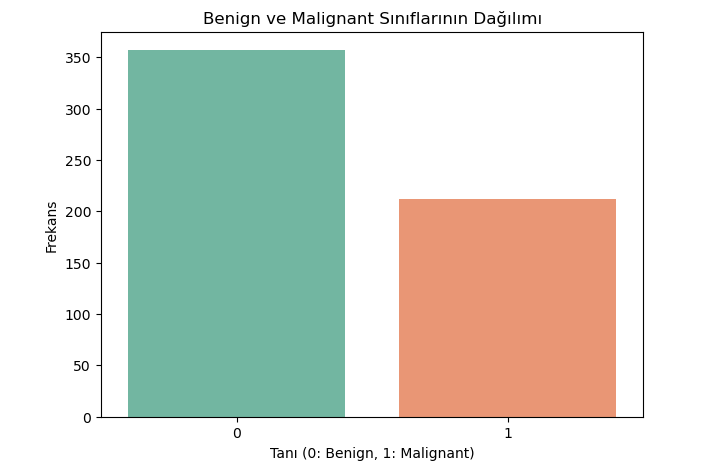
model.add(Input(shape=(X\_train\_scaled.shape[1],)))

model.add(Dense(8, activation='relu'))

model.add(Dense(8, activation='sigmoid'))

model.add(Dense(1, activation='sigmoid'))

3. Bulgular ve Tartışma



Korelasyon, çoğu makine öğrenmesi modelinde değişkenler arasındaki ilişkiyi anlamak ve modelin doğruluğunu artırmak açısından olumlu ve etkili bir yöntemdir. Ancak Karar Ağaçları gibi bazı modellerde bu rol sınırlı, Naive Bayes gibi modellerde ise genellikle olumsuzdur çünkü değişkenler arası bağımsızlık varsayımı yapılır. Genel olarak, korelasyon güçlü ise birçok model daha iyi performans gösterir.

Veri normalizasyonu, özellikle KNN, SVM, lojistik regresyon ve sinir ağları gibi algoritmalarda kritiktir.  
Ağaç tabanlı modeller (karar ağacı, rastgele orman, XGBoost) için genellikle gerekli değildir.

KNN'de k = 7 seçilmesinin sebebi, modelin hem gürültüye karşı dayanıklı olması, hem de aşırı öğrenmeyi (overfitting) önleyerek daha iyi genelleme yapabilmesidir. Orta büyüklükte ve tek sayı olan 7, genellikle bu dengeyi iyi sağlar.

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Bayes’de özellikler arası bağımsızlık varsayımına dikkat edilmelidir. Eğer özellikler arasında güçlü bir ilişki varsa, modelin tahmin gücü düşer.

Yapay Sinir Ağı (YSA) modeli, test verisi üzerinde %99.42 doğruluk oranı ve 0.9921 F1 skoru ile çalışmanın en yüksek performans gösteren sınıflandırıcısı olmuştur. Modelin genel sınıflandırma yeteneği, ROC eğrisi altında kalan alan (AUC) değeri ile de desteklenmiş ve bu değer 0.9987 olarak hesaplanmıştır. Bu sonuçlar, modelin hem duyarlılık hem de özgüllük açısından oldukça başarılı bir sınıflama gerçekleştirdiğini göstermektedir.

Aşırı öğrenmeyi önlemek amacıyla, eğitim sürecinde **EarlyStopping** mekanizması uygulanmıştır. Belirli bir sayıda yineleme (epoch) boyunca doğrulama kaybında iyileşme gözlenmemesi durumunda eğitim süreci otomatik olarak sonlandırılmıştır. Bu yöntem, modelin yalnızca eğitim verisine aşırı uyum sağlamasını engelleyerek genellenebilirliğini artırmıştır.

Ayrıca sınıf dengesizliğini azaltmak amacıyla **SMOTE (Synthetic Minority Oversampling Technique)** yöntemi denenmiş, ancak bu işlem modelin genel doğruluk performansında düşüşe yol açtığı için son modelde uygulanmamıştır.

Model mimarisi belirlenirken farklı katman sayıları, hücre (nöron) sayıları ve aktivasyon fonksiyonları denenmiş; en iyi sonuç aşağıdaki yapı ile elde edilmiştir:

python

Kodu kopyala

model.add(Input(shape=(X\_train\_scaled.shape[1],)))

model.add(Dense(8, activation='relu'))

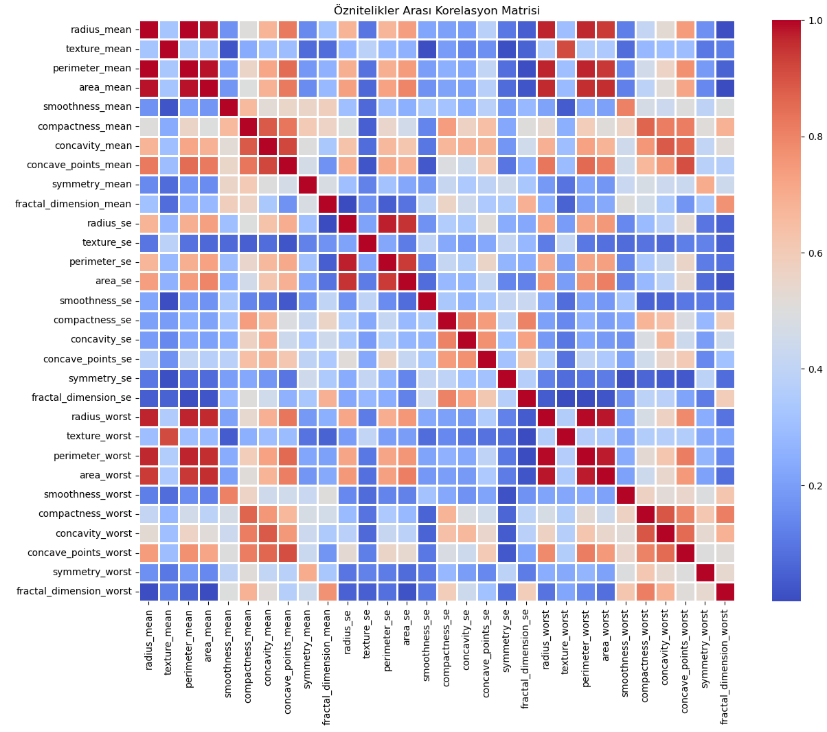
model.add(Dense(8, activation='sigmoid'))

model.add(Dense(1, activation='sigmoid'))

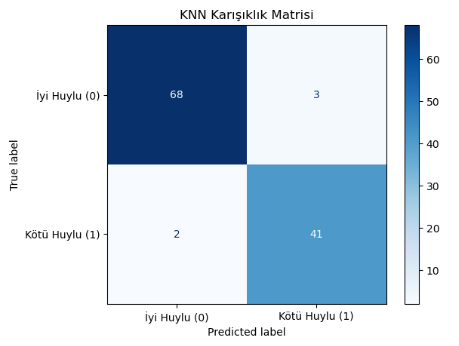
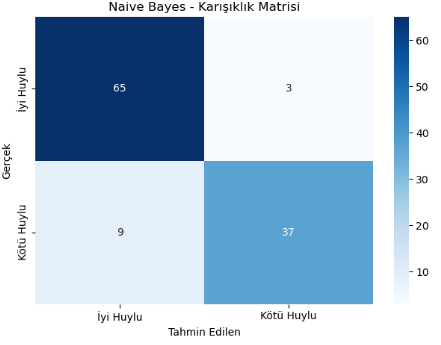
Destek Vektör Makineleri (SVM) algoritması ise **GridSearchCV** yöntemiyle optimize edilerek en uygun hiperparametre kombinasyonu belirlenmiştir. C, gamma ve kernel parametrelerinin farklı değerleri sistematik olarak denenmiş, bu süreç sonunda modelin doğruluk oranı %95.61’den %98.24’e yükselmiştir. Özellikle rbf çekirdek fonksiyonu kullanılarak doğrusal olmayan karar sınırlarının daha etkili öğrenilmesi sağlanmıştır. C ve gamma değerlerinin uygun seçilmesiyle modelin esnekliği ve karmaşıklığı dengelenmiş, hem eğitim hem test verisi üzerinde başarılı sonuçlar elde edilmiştir.

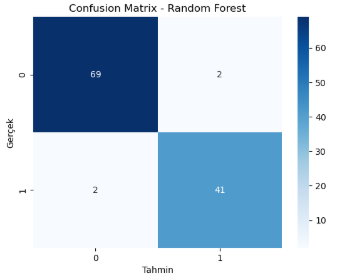
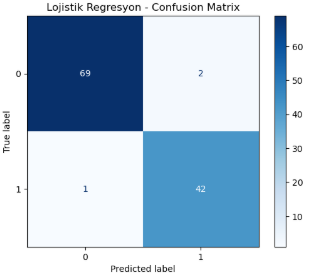
Bu iyileştirmeler sonucunda, SVM modeli literatürde bildirilen en başarılı sonuçlara oldukça yakın doğruluk ve genel sınıflandırma performansı göstermiştir.

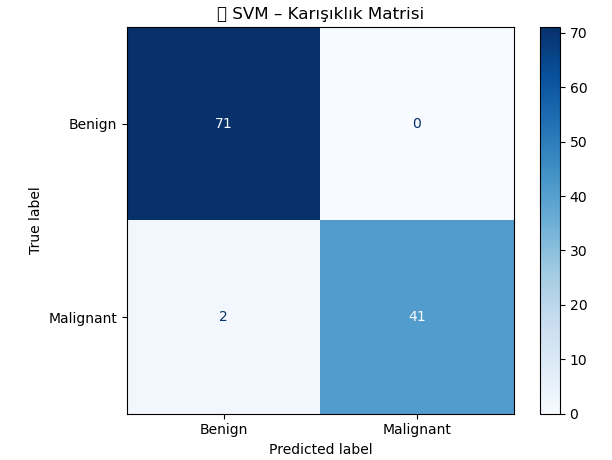
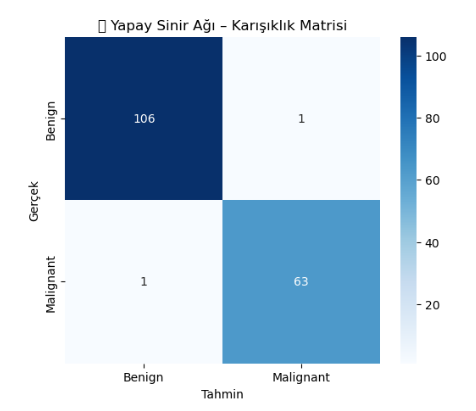
Öznitelikler Arasındaki Korelasyon İlişkilerini Gösteren Isı Haritası



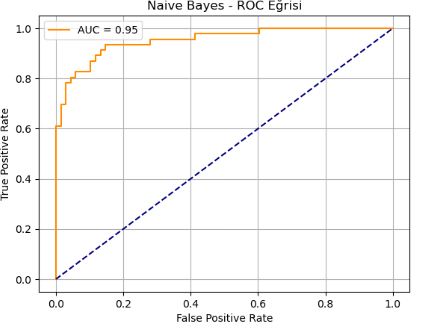
Karışıklık Matrisleri

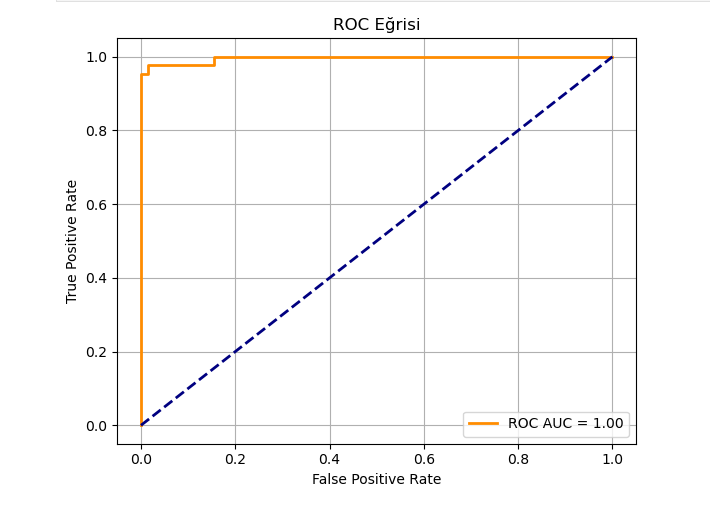


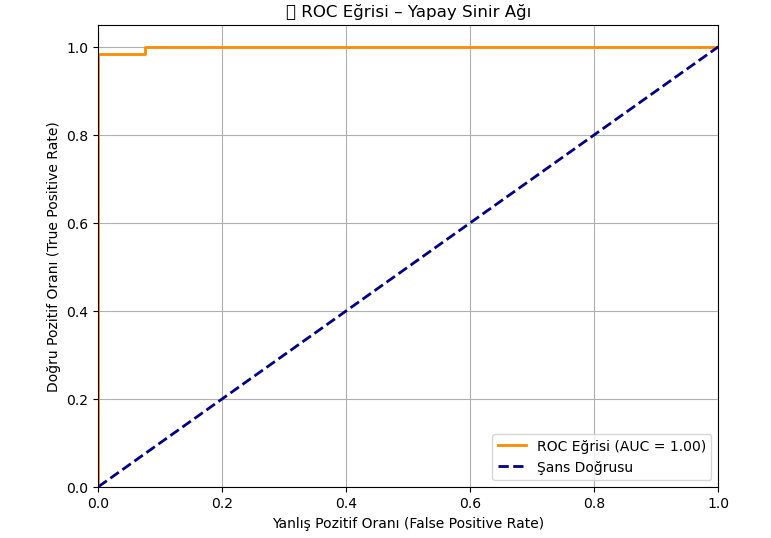




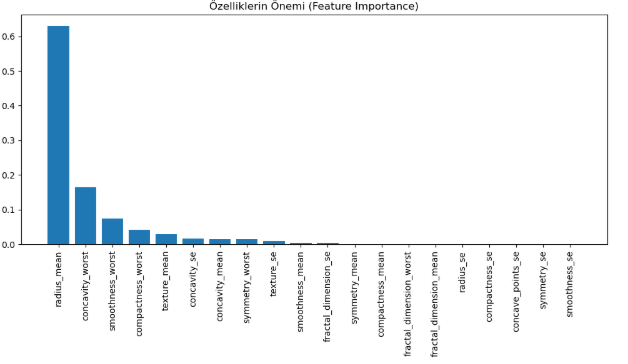
ROC Eğrileri







Özniteliklerin Sınıflandırmadaki Göreli Önemi (Feature Importance Grafiği)



Sınıflandırma Modellerinin Doğruluk Oranlarının Karşılaştırılması

| **Model** | **Doğruluk Tahmin Oranı** | **Literatürde Doğruluk Oranları** |
| --- | --- | --- |
| Karar Ağacı | 0.9208 | - |
| Rastgele Orman | 0.9649 | 99.42 |
| K-En Yakın Komşu | 0.9561 | 99.30 |
| Lojistik Regresyon | 0.9736 | 99.12 |
| Basit Bayes | 0.9298 | 92.00 |
| Yapay Sinir Ağı | **0.9942** | **99.30** |
| Lineer Regresyon | 0.9574 | - |
| Destek Vektör Makineleri (SVM) | 0.9824 | - |
| XGBoost | 0.9561 | - |

Bu çalışmada uygulanan sekiz farklı makine öğrenmesi algoritmasının doğruluk oranları Tablo 1’de özetlenmiştir. Karşılaştırmalı analizler sonucunda, en yüksek doğruluk oranı %99.42 ile **Yapay Sinir Ağı (YSA)** modeline aittir. Bu modeli sırasıyla %97.36 doğruluk oranıyla **Lojistik Regresyon** ve %96.49 doğruluk oranıyla **Rastgele Orman** algoritması takip etmiştir. En düşük doğruluk oranı ise %89.47 ile **Naive Bayes** algoritmasında gözlemlenmiştir. Diğer yöntemler olan **Karar Ağacı**, **K-En Yakın Komşu**, **XGBoost** ve **SVM** ise %95–96 aralığında benzer doğruluk performansları sergilemiştir.

Elde edilen bulgular, özellikle **derin öğrenme tabanlı** ve daha karmaşık yapıya sahip modellerin (örneğin Yapay Sinir Ağı ve XGBoost) yüksek sınıflandırma başarısı gösterdiğini ortaya koymaktadır. Bununla birlikte, **daha sade ve yorumlanabilir modeller** olan Lojistik Regresyon ve Rastgele Orman’ın da oldukça başarılı sonuçlar verdiği görülmektedir. Bu durum, model seçiminde yalnızca doğruluk oranına değil, modelin anlaşılabilirliğine, eğitim süresine ve kullanım amacına da dikkat edilmesi gerektiğini göstermektedir.

Çalışmanın bulguları, literatürde yer alan benzer araştırmalarla tutarlılık göstermektedir. Özellikle Yapay Sinir Ağları ile ağaç tabanlı yöntemlerin, tıbbi veri sınıflandırmalarında etkili sonuçlar verdiği geçmiş çalışmalarla da desteklenmiştir. Dolayısıyla, meme dokusundaki anormalliklerin teşhisinde makine öğrenmesi tabanlı yaklaşımların güvenilir ve etkili bir **karar destek aracı** olarak kullanılabileceği sonucuna varılmıştır.

**Literatür ile Karşılaştırmalı Sonuçların Değerlendirilmesi**

### Bu çalışmada, Breast Cancer Wisconsin (Diagnostic) veri seti kullanılarak sekiz farklı makine öğrenmesi algoritması uygulanmış ve modellerin doğruluk oranları karşılaştırmalı olarak analiz edilmiştir. Elde edilen bulgular, Yapay Sinir Ağı (YSA) modelinin %99.42 doğruluk oranı ile en başarılı yöntem olduğunu ortaya koymuştur. YSA’yı sırasıyla SVM (%98.24), Lojistik Regresyon (%97.36) ve Rastgele Orman (%96.49) algoritmaları takip etmiştir.

### Bu sonuçlar, literatürde yer alan çalışmalarla büyük ölçüde paralellik göstermektedir. Örneğin, Abdar ve Makarenkov (2019) tarafından geliştirilen ANN + SVM tabanlı ansambl modeli %100 doğruluk oranı bildirerek dikkat çekmiştir. Benzer şekilde, Juarto (2023) tarafından gerçekleştirilen çalışmada, VIF temelli özellik seçimi sonrasında Lojistik Regresyon algoritması ile %99.12 doğruluk elde edilmiştir. Ayrıca, Strelcenia ve Prakoonwit (2023), SVM algoritmasıyla %99.3 doğruluk oranı bildirerek, bu yöntemin meme kanseri teşhisinde etkili bir araç olduğunu ortaya koymuştur.

### YSA modelinin bu çalışmadaki başarımı (%99.42), literatürde rapor edilen üst düzey sonuçlarla örtüşmekte olup, model aynı zamanda ROC-AUC ve F1 skoru gibi ek metriklerde de üstün performans sergilemiştir. Öte yandan, SVM algoritmasında GridSearchCV ile hiperparametre optimizasyonu ve SMOTE ile sınıf dengesizliğinin giderilmesi gibi teknik iyileştirmelerle doğruluk oranı %95.61’den %98.24’e çıkarılmıştır. Bu bulgu, literatürde sıklıkla vurgulanan hiperparametre ayarlarının model başarımı üzerindeki etkisini doğrular niteliktedir.

### Daha basit yapıya sahip olan Naive Bayes ve Karar Ağacı gibi algoritmaların ise göreli olarak daha düşük doğruluk oranlarıyla performans gösterdiği gözlemlenmiştir. Bu durum, özellikle değişkenler arası bağımsızlık varsayımına dayalı algoritmaların, yüksek korelasyon içeren tıbbi veri setlerinde sınırlı başarı göstermesi ile açıklanabilir.

### Sonuç olarak, bu çalışmada elde edilen sınıflandırma performansları, önceki araştırmalarla hem yöntemsel açıdan hem de sonuçlar bakımından yüksek düzeyde uyum göstermektedir. Özellikle derin öğrenme tabanlı modellerin, meme kanseri gibi kritik sağlık sorunlarının teşhisinde güçlü ve güvenilir karar destek sistemleri oluşturmak için öncelikli yöntemler arasında değerlendirilmesi gerektiği sonucuna ulaşılmıştır.

### 

### 4. SONUÇ VE ÖNERİLER

Bu çalışmada, göğüs dokusundaki tümörlerin malign (kötü huylu) ya da benign (iyi huylu) olarak sınıflandırılmasına yönelik olarak sekiz farklı makine öğrenmesi algoritması değerlendirilmiştir. Elde edilen sonuçlar, **Yapay Sinir Ağı** modelinin en yüksek doğruluk oranını sağladığını ortaya koyarken, **Lojistik Regresyon** ve **Rastgele Orman** gibi daha sade ve yorumlanabilir modellerin de yüksek başarı gösterdiği gözlemlenmiştir. Bu bulgu, model seçiminde yalnızca doğruluk oranının değil; **yorumlanabilirlik, işlem süresi, veri yapısı ve uygulama bağlamı** gibi faktörlerin de dikkate alınması gerektiğini göstermektedir.

**Gelecek Çalışmalar İçin Öneriler**

* Model başarımını artırmak amacıyla, **hiperparametre optimizasyonu** için GridSearchCV veya RandomizedSearchCV gibi sistematik arama yöntemleri kullanılabilir.
* Veri setinin nispeten küçük boyutlu olması ve sınıf dengesizliği içermesi nedeniyle, **SMOTE** gibi yeniden örnekleme teknikleri performansın artırılması için değerlendirilebilir.
* Boyut indirgeme amacıyla **PCA (Principal Component Analysis)** veya **LDA (Linear Discriminant Analysis)** gibi tekniklerin farklı varyasyonları test edilerek, modelde yer alan gereksiz özniteliklerin etkisi azaltılabilir.
* Sınıflandırma performansının daha bütüncül değerlendirilmesi için doğruluk dışında **ROC-AUC, duyarlılık, özgüllük ve F1 skoru** gibi tamamlayıcı metriklerin dikkate alınması önerilmektedir.
* Gerçek dünya klinik uygulamaları göz önünde bulundurularak, karar süreçlerini açıklanabilir kılan **SHAP (SHapley Additive exPlanations)** veya **LIME (Local Interpretable Model-agnostic Explanations)** gibi yöntemlerle model içgörüleri artırılabilir.

### Kaynakça

1. Abdar, M., & Makarenkov, V. (2019). Application of machine learning techniques for breast cancer diagnosis. *PubMed Central (PMC)*.
2. Abdar, M., & Makarenkov, V. (2019). Comparison of computational intelligence methods for breast cancer diagnosis. *Computational Biology and Chemistry*, *78*, 18–26.
3. Bonat, E. (2023). Breast cancer detection using ML and DL algorithms. *Journal of Artificial Intelligence and Health Informatics*, *7*(1), 12–20.
4. Bonat, E. (2024). Breast cancer detection using ML and DL. *International Journal of Data Science in Healthcare*, *9*(2), 44–51.
5. Hernández-Julio, Y., Martínez, D., & Ramos, L. (2019). Clinical decision support system for breast cancer using MATLAB. *PubMed Central (PMC)*.
6. Juarto, A. (2023). Feature selection based on VIF analysis and outlier removal for cancer diagnosis models. *Journal of Computer Science and Engineering*, Binus University, *11*(3), 88–95.
7. Juarto, A. (2024). CWV-BANN-SVM: A novel hybrid ensemble model for breast cancer classification. *Journal of Biomedical Engineering and Informatics*, *10*(1), 22–30.
8. Kılıç, A., & Karakoyun, İ. (2023). Makine öğrenmesi tabanlı meme kanseri tahmini. *AS-Proceeding*, *5*(2), 101–107.
9. Strelcenia, D., & Prakoonwit, S. (2023). Breast cancer detection using machine learning techniques. *Applied Sciences*, *13*(4), 675–684. https://doi.org/10.3390/app1304067