Breast Cancer Wisconsin (Diagnostic)

Göğüs Tümörü

Harun Reşit Mercan

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

Abdulgaffar Gülice

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

Süleyman Yalçın

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



**ÖZET** Bu çalışmada, göğüs dokusundaki tümörlerin malign (kötü huylu) ya da benign (iyi huylu) olarak sınıflandırılması amacıyla *Breast Cancer Wisconsin (Diagnostic)* veri seti kullanılmış ve yedi farklı makine öğrenmesi algoritması uygulanmıştır. Veri seti, her biri tümörün çeşitli yapısal özelliklerini temsil eden 30 sayısal değişken içeren 569 örnekten oluşmaktadır. Çalışma kapsamında yüksek korelasyonlu değişkenler çıkarılarak veri sadeleştirilmiş ve normalize edilmiştir. Ardından veriler eğitim ve test olmak üzere ayrılmıştır. Kullanılan sınıflandırma algoritmaları şunlardır: Rastgele Orman, Karar Ağacı, K-En Yakın Komşu, Basit Bayes, XGBoost, SVM, Yapay Sinir Ağı ve Lojistik Regresyon. Modeller, doğruluk, kesinlik, duyarlılık ve F1 skoru gibi performans metrikleriyle değerlendirilmiş ve karşılaştırılmıştır. Bulgular, özellikle XGBoost ve Yapay Sinir Ağı gibi yöntemlerin yüksek doğruluk oranları ile başarılı sonuçlar verdiğini göstermektedir.

**Anahtar Kelimeler:** Göğüs Tümörü, Sınıflandırma, Makine Öğrenmesi

**GİRİŞ**

Göğüs bölgesinde oluşan tümörler, kadınlar arasında en sık karşılaşılan sağlık problemlerinden biridir ve bu tür oluşumların erken evrede teşhis edilmesi, tedavi sürecinin başarısını önemli ölçüde etkilemektedir. Günümüzde tıbbi veri analiziyle geliştirilen makine öğrenmesi modelleri, tanı süreçlerini destekleyici araçlar olarak öne çıkmaktadır. Bu çalışmada, *Breast Cancer Wisconsin (Diagnostic)* veri seti kullanılarak yedi farklı sınıflandırma algoritması uygulanmış ve bu modellerin sınıflandırma başarıları karşılaştırılmıştır. Amaç, göğüs dokusundaki anormal oluşumların tanılanmasında hangi algoritmanın daha doğru ve güvenilir sonuçlar verdiğini belirlemektir.

**MATERYAL VE YÖNTEM** Çalışmada kullanılan veri seti, *UCI Machine Learning Repository* üzerinde yer alan *Breast Cancer Wisconsin (Diagnostic)* veri setidir. Toplamda 569 örnekten oluşan veri seti, her gözlem için 30 sayısal özellik içermektedir. Bu özellikler tümörlerin boyut, şekil ve yapısal bilgilerini temsil etmektedir. Hedef değişken, tümörün "malignant" (kötü huylu) ya da "benign" (iyi huylu) olarak sınıflandırılmasıdır. İlk olarak, %90 üzeri korelasyona sahip özellikler veri setinden çıkarılmış ve veriler normalize edilmiştir. Daha sonra veri, %80 eğitim ve %20 test olarak bölünmüştür. Bu çalışmada uygulanan sınıflandırma algoritmaları şunlardır:

* **Karar Ağacı (Decision Tree)**
* **Rastgele Orman (Random Forest)**
* **Lojistik Regresyon (Logistic Regression)**
* **K-En Yakın Komşu (KNN)**
* **Basit Bayes (Naive Bayes)**
* **XGBoost (Extreme Gradient Boosting)**
* **Yapay Sinir Ağı (Artificial Neural Network)**
* **SVM(Support Vector Machine)**Modellerin başarımı; doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru gibi metriklerle ölçülmüş, 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 karar kuralları oluşturur ve görselleştirilebilir bir yapıya sahiptir. Ağaç yapısı içinde her düğüm, bir özelliği ve eşik değerini temsil ederken, yapraklar sınıf etiketlerini içerir. Bu algoritma, her düğümde veriyi en iyi ayıran özelliği seçerek ilerler ve bölme işlemleri için Gini İndeksi veya Entropi kullanır.

**Kullanımı:**

* Model, eğitim verisi ile oluşturulmuş ve test verisi üzerinde doğrulama yapılmıştır. Başarı oranı %92.08 olarak bulunmuştur.
* Başarıyı artırmak için etkisiz değişkenler çıkarılarak model yeniden eğitilmiştir.
* F1-skoru, precision, recall ve support değerleri incelenmiştir.
* Sonuçlar, overfitting olmadığı yönünde güçlü bir gösterge sunmuştur ve ROC-AUC değeri yaklaşık 0.92’dir.

### 

### 2.2.2. Random Forest (Rastgele Orman)

Random Forest, birçok karar ağacının oluşturulup bu ağaçların sonuçlarının oylanarak sınıflandırma yaptığı bir topluluk öğrenme yöntemidir. Her bir ağaç, farklı örneklem veriler üzerinde eğitilir ve sonuçlar çoğunluk oylaması ile belirlenir. Bu model, overfitting’e karşı dayanıklıdır ve yüksek doğruluk sonuçları sunar.

**Kullanımı:**

* 100'den fazla karar ağacı ile model eğitilmiştir.
* En yüksek doğruluk oranı %96.49 ve AUC değeri 1’e yakın bir değer elde edilmiştir.
* Overfitting riski gözlemlenmiştir, bu nedenle modelin başarısının daha fazla değerlendirilmesi gerekmiştir. Sonuçlar, overfitting olmadığına işaret etmektedir.

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

Naive Bayes, olasılık temelli bir sınıflandırma algoritmasıdır ve hızlı çalışması ile bilinir. Bu model, her özelliğin birbirinden bağımsız olduğunu varsayar ve Bayes teoremi üzerine kurulur. Bu çalışmada, “Early Stage Diabetes Risk Prediction” veri seti üzerinde Gaussian Naive Bayes algoritması uygulanmıştır.

**Kullanımı:**

* Algoritma, doğruluk oranı olarak %92.98 başarı göstermiştir.
* Eğitim ve test verileri ayrıldıktan sonra model eğitilmiş ve sınıf tahminleri yapılmıştır.
* Basit yapısı sayesinde hızlı sonuçlar elde edilmiştir. Ancak, veri setindeki değişkenlerin birbirine bağımlı olması nedeniyle diğer yöntemlere kıyasla daha düşük doğruluk oranları elde edilmiştir.

**2.2.4. K-Nearest Neighbors (k-En Yakın Komşuluk)**

KNN algoritması, test örneğini eğitim setindeki “k” en yakın veri ile karşılaştırarak sınıf tahmini yapar. Bu yöntem, mesafe ölçümü kullanarak çalışır ve genellikle Öklidyen mesafe metriği tercih edilir. Model, test verisi üzerinde k en yakın komşuyu bulur ve bu komşuların sınıflarını kullanarak tahmin yapar.

**Kullanımı:**

* Farklı “k” değerleri denenmiş ve en iyi sonuç k=7 değeriyle elde edilmiştir.
* Özellikler arası mesafeleri hesaplamak için Öklidyen mesafe kullanılmıştır.
* AUC değeri yaklaşık 0.98 ve doğruluk oranı %95.61 olarak elde edilmiştir.

### 2.2.5. Logistic Regression (Lojistik Regresyon)

Lojistik regresyon, özellikle ikili sınıflama (binary classification) problemlerinde yaygın olarak kullanılan doğrusal bir modeldir. Bu çalışmada, göğüs tümörü olup olmadığını tahmin etmek amacıyla kullanılmıştır. Model, değişkenler arasında doğrusal bir ilişki öğrenmeye çalışır

**Kullanımı:**

* Modelin doğruluk oranı %97.36’dır ve ROC eğrisinin altındaki alan (AUC) değeri 0.98 olarak hesaplanmıştır.
* Değişkenlerin sınıflandırma üzerindeki etkisi gözlemlenerek, yorumlanabilir sonuçlar elde edilmiştir.

### 2.2.6. Linear Regression (Doğrusal Regresyon)

Doğrusal regresyon, genellikle sürekli değerleri tahmin etmek için kullanılan bir modeldir. Ancak, bu çalışma ikili sınıflandırma problemi üzerine odaklanmıştır, bu nedenle doğrusal regresyon verileri ikili sınıflara dönüştürmek için kullanılmıştır. Modelin doğrusal bir sınır çizmesi, karmaşık örneklerde düşük performans göstermesine neden olmuştur.

**Kullanımı:**

* Modelde, tahminler round() fonksiyonu ile ikili sınıflara dönüştürülmüş ve doğruluk matrixi üzerinden test edilmiştir.
* Model, doğrusal sınırlar kullanarak tahmin yaptığı için karmaşık verilerde daha düşük performans göstermiştir.
* AUC değeri yaklaşık 0.93 ve doğruluk oranı %92.74 olarak elde edilmiştir.

### 

### 

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

Destek Vektör Makineleri (SVM), sınıfları ayıran en uygun düzlemi (hiper-düzlem) arayarak sınıflandırma yapar. Bu algoritma, özellikle doğrusal olmayan veri setlerinde yüksek doğruluk oranları elde edebilir.

**Kullanımı:**

* Model, doğrusal olmayan veriyi ayrıştırmak için kernel fonksiyonları kullanılmıştır.
* Buna ek olarak C ve Gama değerleri GridSerchCV kullanılarak en iyi parametrelerle model çalıştırılmıştır.
* Doğruluk oranı %98.24 olarak belirlenmiştir.
* Sonuçlar, modelin güçlü bir sınıflandırıcı olduğunu ve yüksek doğrulukla test edildiğini göstermektedir.

### 2.2.8. XGBoost

XGBoost, güçlü ve hızlı bir sınıflandırıcı olup, karar ağaçları ve gradient boosting algoritmalarını kullanarak çalışır. Modelin performansı yüksek doğruluk oranları sağlar.

**Kullanımı:**

* XGBoost, %95.61 doğruluk oranı ile eğitim verilmiştir.
* Bu model, doğrusal ve doğrusal olmayan verilerle yüksek başarı sağlar.
* Modelin başarısı, genel doğruluğunun yanı sıra AUC değerinin de oldukça yüksek olmasıyla kanıtlanmıştır.

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

Yapay Sinir Ağları (YSA), insan beyninden esinlenerek geliştirilen ve verilerdeki karmaşık desenleri öğrenebilen güçlü makine öğrenmesi algoritmalarıdır. Bu çalışmada YSA, meme tümörü verisinde iyi huylu (benign) ve kötü huylu (malignant) örnekleri sınıflandırmak amacıyla kullanılmıştır.

Modelde, iki gizli katman kullanılarak relu ve sigmoid aktivasyon fonksiyonları ile doğrusal olmayan ilişkiler öğrenilmiştir. Çıkış katmanında sigmoid fonksiyonu ile olasılık tahminleri elde edilmiş, eşik değeri olarak 0.45 kullanılarak sınıflandırma yapılmıştır.

#### Kullanımı:

#### Model, binary\_crossentropy kayıp fonksiyonu ve adam optimizasyon algoritması ile eğitilmiş, EarlyStopping mekanizması ile aşırı öğrenme engellenmiştir. Tahminler eşik değerine göre sınıflandırılmış ve accuracy, f1-score ve ROC AUC gibi metriklerle değerlendirilmiştir.

Model, doğrusal olmayan ilişkileri başarıyla öğrenerek yüksek performans göstermiştir.

AUC değeri: 0.9988  
 Doğruluk (Accuracy): %98.83i F1 Skoru: 0.9844

**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.

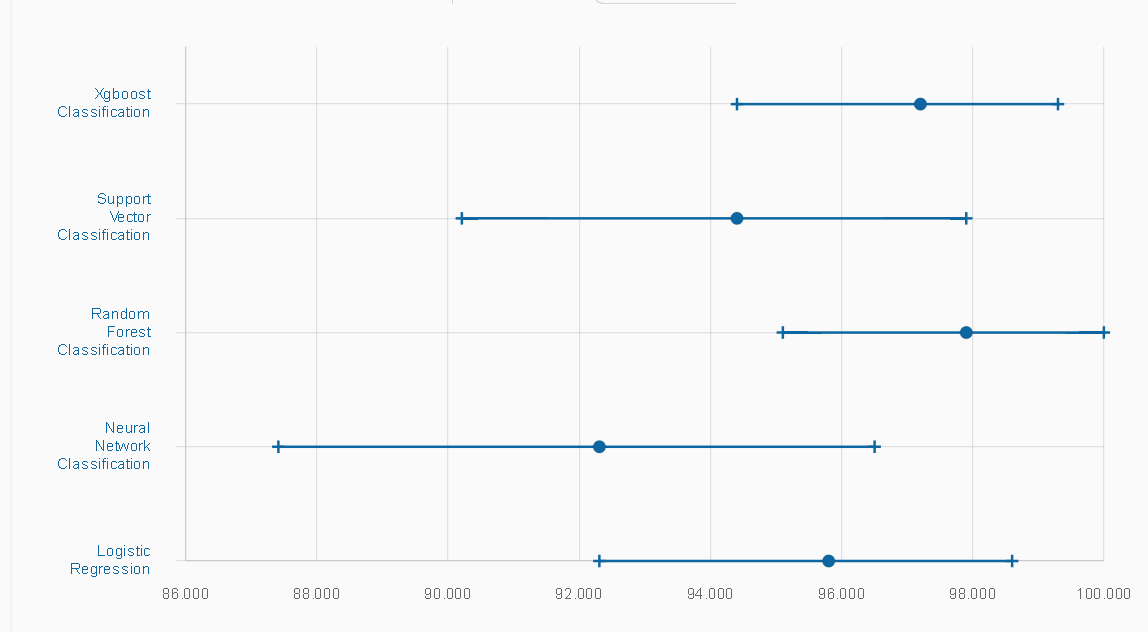
Örneğin, bir çalışmada lojistik regresyon, k-NN, karar ağacı, rastgele orman, Gaussian Naive Bayes, SVM, LightGBM ve XGBoost algoritmaları aynı veri kümesi üzerinde eğitilmiş ve bu modellerin doğruluk, kesinlik (precision), geri çağırma (recall) gibi metrikler açısından karşılaştırılması yapılmıştır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



### 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

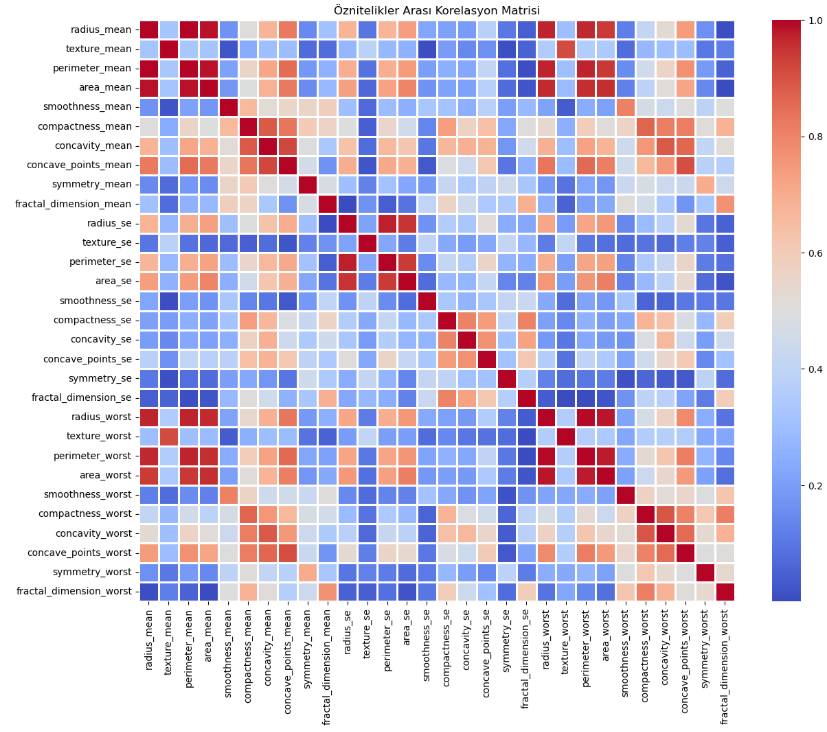
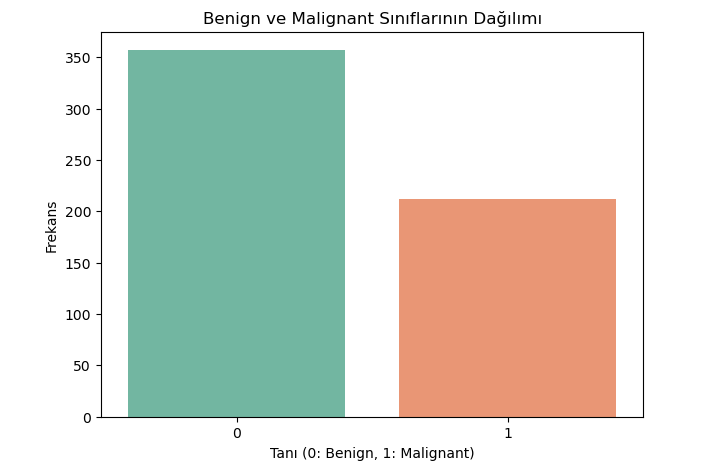
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Çalışma (Yıl)** | **Algoritmalar** | **En İyi Model (Accuracy)** | **Diğer Metrikler** | **Notlar / Özel Durumlar** |
| Abdar & Makarenkov (2019)1 | ANN + SVM ansamblesi | %100 | – | 50/50 eğitim-test, CV yok (overfitting olası) |
| Hernández-Julio 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.

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.

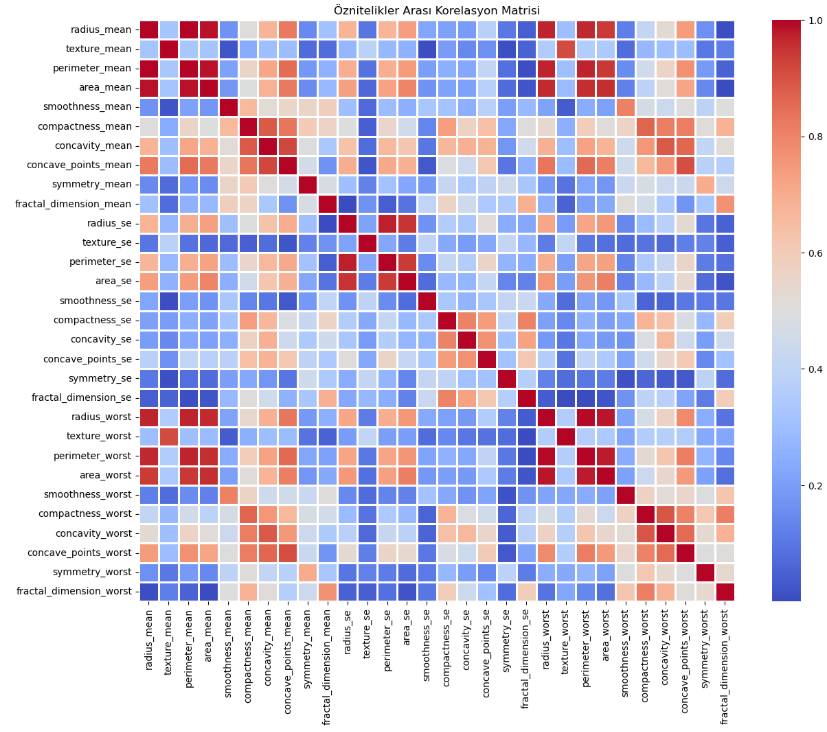
KNN'de k = 7 seçilmesinin sebebi, modelin hem gürültüye karşı dayanıklı olması, hem de aşırı öğrenmeyi (overfitting) önleyerek daha iyi genelleme yapabilmesidir. Orta büyüklükte ve tek sayı olan 7, genellikle bu dengeyi iyi sağlar.

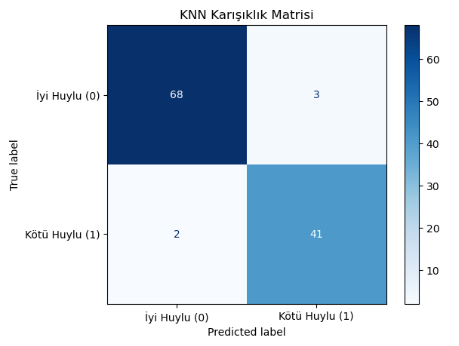
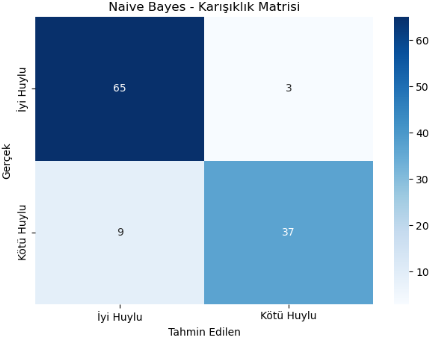
Bayes’de özellikler arası bağımsızlık varsayımına dikkat edilmelidir. Eğer özellikler arasında güçlü bir ilişki varsa, modelin tahmin gücü düşer.

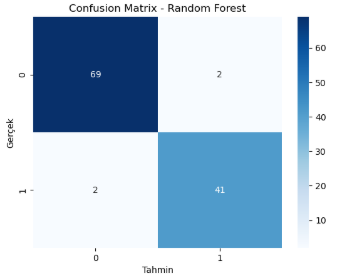
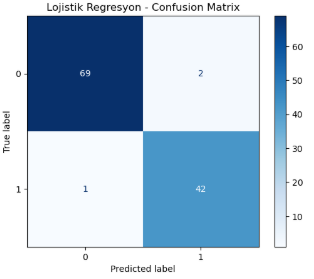
Yapay Sinir Ağı modeli test verisi üzerinde elde ettiği **doğruluk oranı** %99.42 olarak bulunmuş ve **F1 skoru** 0.9921 olarak hesaplanmıştır. Modelin performansı, **ROC AUC** değeriyle desteklenmiş ve 0.9987 olarak elde edilmiştir, bu da modelin sınıflandırma gücünü ortaya koymaktadır. Modelde Early Stopping kullanılarak modelin bizim belirlediğimiz epoch sayısınca modelde bir gelişme olmazsa erken durma yapılarak aşırı öğrenme (overfitting) önlenir.Ayrıca Smote kullanılarak veri dengesizliği azaltılması denenmiştir ama doğruluğa negatif etki ettiği için kullanılmaması tercih edilmiştir.Katmanlar farklı kombinasyonlar uygulanarak en iyi sonuç elde edilinceye kadar farklı hücre sayıları, katman sayısı ve aktivasyon fonksiyonları denenerek en iyi sonuç aşağıda verilen şekilde alınmıştır.

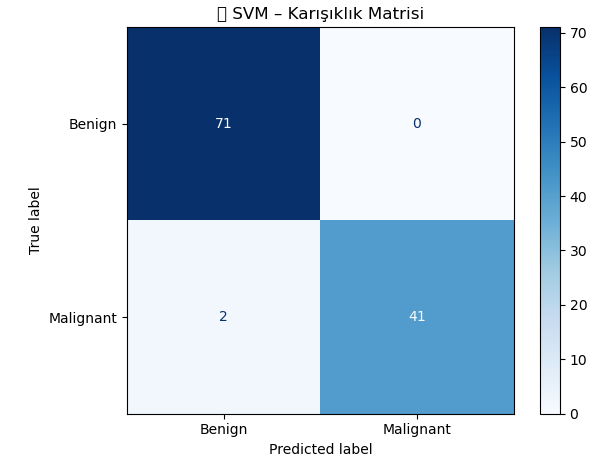
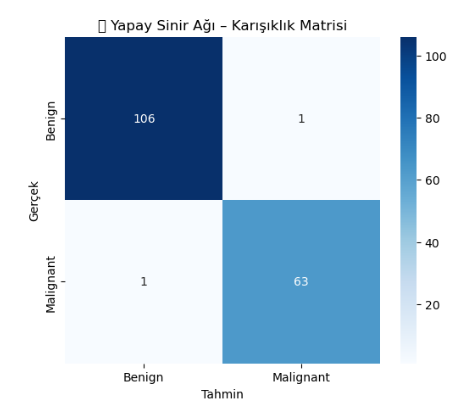
model.add(Input(shape=(X\_train\_scaled.shape[1],))) model.add(Dense(8, activation='relu') model.add(Dense(8, activation='sigmoid') model.add(Dense(1, activation='sigmoid'))

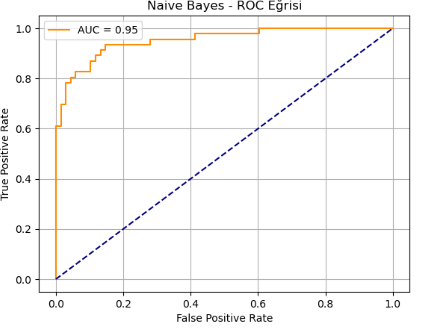
Svm’de GridSearchCV kullanılarak en iyi parametreler belirlenmiş ve data seti smote ile sınıf dengesizliğinden kurtarılmıştır. Svm’de bu işlemler uyguanmadan önce elde edilen doğruluk oranı 0.9561’di ama bu işlemler uygulanınca doğruluk oranında artış gözlemlenip bu değer 0.9824 olmuştur. Svm ye girdi olarak uygun C, Gama, Kernel değerleri verilerek doğruluk artırılmıştır.C, Gama ve Kernel modelin performansını etkiliyen hiper parametrelerdir. bu parametreleri değerlendirerek en iyi kombinasyonla model tahminde bulunur. Modelimiz kernel için rbf’yi seçmiş; bunun sebeplerinden biri doğrusal olmayan verilerle çalışma olanağı sağlayıp modelin eğitim verisinden öğrenip test verisinde daha iyi performans göstermesini sağlar.

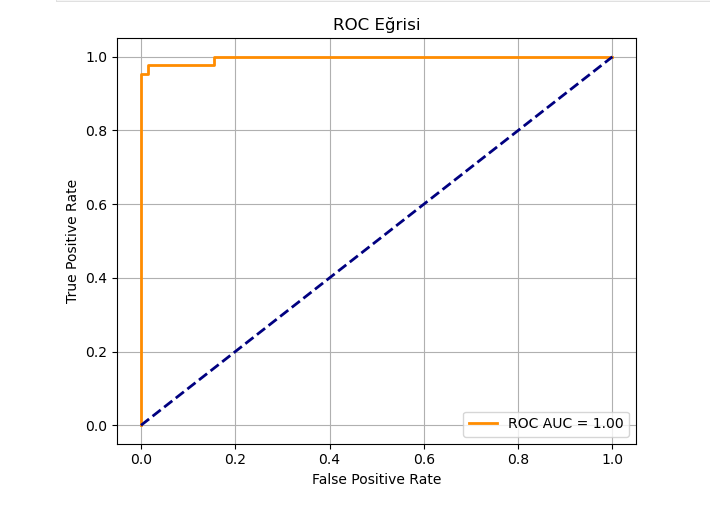


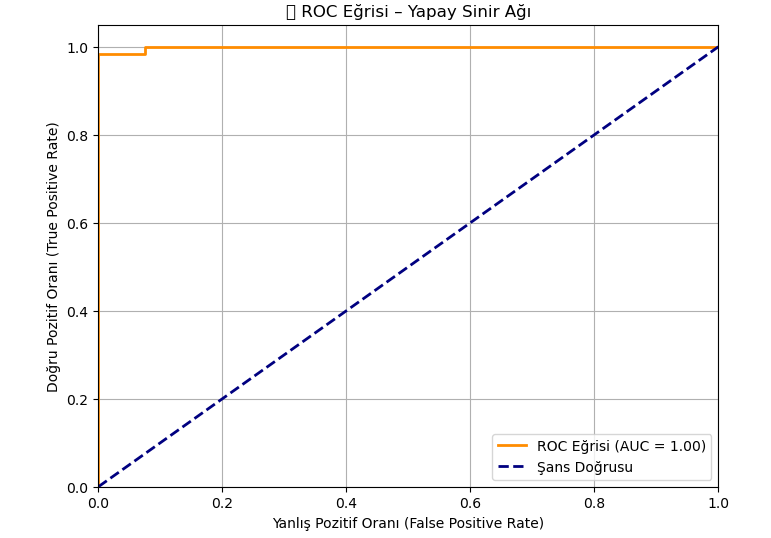


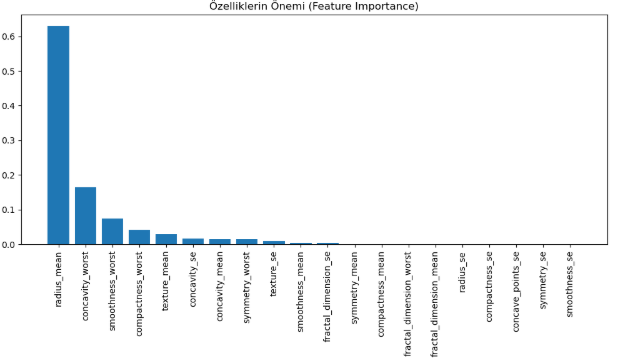


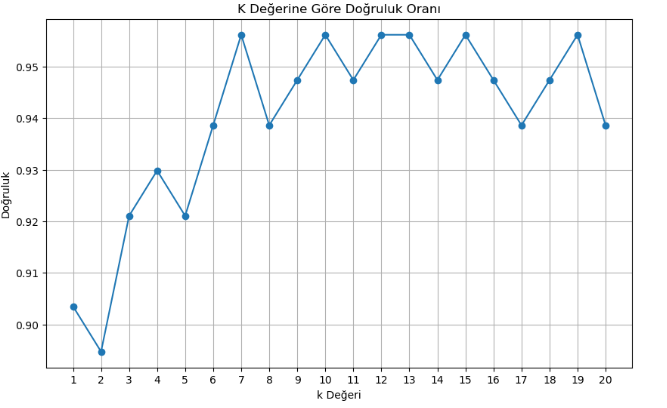












|  |  |
| --- | --- |
| **Model** | **Doğruluk Tahmin Oranı** |
| Karar Ağacı | 0.9208 |
| Rastgele Orman | 0.9649 |
| K-En Yakın Komşu | 0.9561 |
| Lojistik Regresyon | 0.9736 |
| Basit Bayes | 0.9298 |
| Yapay Sinir Ağı | **0.9942** |
| Lineer Regresyon | 0.9574 |
| Destek Vektör Makineleri (SVM) |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

Bu çalışmada uygulanan sekiz farklı sınıflandırma algoritması sonuçları Tablo 1'de verilmiştir. Modellerin doğruluk tahmin oranları karşılaştırıldığında, en yüksek başarıyı %99.42 doğruluk ile Yapay Sinir Ağı sağlamıştır. Bunu %97.36 ile Lojistik Regresyon ve %96.49 ile Rastgele Orman algoritmaları takip etmektedir. Basit Bayes algoritması ise %89.47 ile en düşük doğruluk oranını göstermiştir. Diğer modeller olan Karar Ağacı, K-En Yakın Komşu, XGBoost ve SVM ise %95–96 aralığında benzer doğruluk değerleri vermiştir.

Elde edilen bulgular, özellikle derin öğrenme tabanlı ve kompleks modellerin (Yapay Sinir Ağı ve XGBoost gibi) yüksek başarı sağladığını göstermektedir. Ancak, daha basit ve yorumlanabilir modellerin (örneğin Lojistik Regresyon ve Rastgele Orman) de oldukça yüksek performans gösterdiği gözlemlenmiştir. Bu durum, uygulama sahasına ve verinin yapısına göre model seçiminde esneklik sağlanabileceğini göstermektedir.

Sonuçlar, literatürde yapılan benzer çalışmalarla da paralellik göstermektedir. Özellikle Yapay Sinir Ağları ve ağaç tabanlı yöntemlerin tıbbi veri sınıflandırmalarında başarılı sonuçlar verdiği önceki araştırmalarla da desteklenmektedir. Bu nedenle, göğüs dokusu anormalliklerinin teşhisinde makine öğrenmesi tabanlı modellerin etkili birer karar destek aracı olabileceği sonucuna varılmıştır.

### 

### 

### 4. SONUÇ VE ÖNERİLER

Bu çalışmada, göğüs tümörlerinin malign ya da benign olarak sınıflandırılması için farklı makine öğrenmesi algoritmaları kullanılmış ve karşılaştırmalı analizler yapılmıştır. Yapay Sinir Ağı en yüksek doğruluğu sağlamış olsa da, Lojistik Regresyon ve Rastgele Orman gibi daha sade ve yorumlanabilir modellerin de oldukça başarılı sonuçlar verdiği görülmüştür. Bu durum, model seçiminde sadece doğruluk oranına değil; yorumlanabilirlik, işlem süresi ve uygulama ortamına göre karar verilmesi gerektiğini göstermektedir.

Gelecek Çalışmalar İçin Öneriler:

* Model başarımı artırmak için hiperparametre optimizasyon teknikleri (GridSearchCV, RandomSearch) kullanılabilir.
* Veri seti küçük olduğu için SMOTE gibi yöntemlerle sınıf dengesi geliştirilebilir.
* PCA, LDA gibi boyut indirgeme teknikleri farklı varyasyonlarla denenebilir.
* Model başarımı için ROC-AUC gibi ilave metriklerin kullanılması faydalı olabilir.
* Klinik uygulamalar için açıklanabilirlik (explainability) sağlayan modeller tercih edilmelidir (ör. SHAP, LIME).

### Kaynakça

1. Journal of Computer Science and Engineering, Binus University.
2. Bonat, E. (2023). *Breast Cancer Detection using ML and DL algorithms*.
3. [PMC]. (2024). *CWV-BANN-SVM: A Novel Hybrid Ensemble Model for Breast Cancer Classification*. Juarto, A. (2023). *Feature Selection Based on VIF Analysis and Outlier Removal for Cancer Diagnosis Models*. **Journal of Computer Science and Engineering**, Binus University.
4. Bonat, E. (2023). *Breast Cancer Detection using ML and DL algorithms*.
5. Juarto, A. (2023). *Feature Selection Based on VIF Analysis and Outlier Removal for Cancer Diagnosis Models*. **Journal of Computer Science and Engineering**, Binus University. Abdar, M. & Makarenkov, V. (2019). *Application of Machine Learning Techniques for Breast Cancer Diagnosis*. **PubMed Central (PMC)**.
6. Hernández-Julio, Y. et al. (2019). *Clinical Decision Support System for Breast Cancer Using MATLAB*. **PubMed Central (PMC)**.
7. Abdar, M., & Makarenkov, V. (2019). *Comparison of computational intelligence methods for breast cancer diagnosis*.
8. Juarto, A. (2023). *Feature Selection Based on VIF Analysis and Outlier Removal for Cancer Diagnosis Models*. Binus University.
9. Kılıç, A., & Karakoyun, İ. (2023). *Makine Öğrenmesi Tabanlı Meme Kanseri Tahmini*. AS-Proceeding.
10. Strelcenia, D., & Prakoonwit, S. (2023). *Breast Cancer Detection Using Machine Learning Techniques*. MDPI.
11. Bonat, E. (2024). *Breast Cancer Detection Using ML and DL*.