T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ SINIFLANDIRMA ALGORİTMALARININ VE PARAMETRELERİNİN GÖĞÜS KANSERİ TANI TAHMİNLERİ

Yunus Emre DEMİREL

YÜKSEK LİSANS TEZİ

Matematik Anabilim Dalı

Matematik Programı

Danışman

Doç. Dr. Mutlu AKAR

Temmuz, 2024

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ SINIFLANDIRMA ALGORİTMALARININ VE PARAMETRELERİNİN GÖĞÜS KANSERİ TANI TAHMİNLERİ

Yunus Emre DEMİREL tarafından hazırlanan tez çalışması 01.07.2024 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Matematik Anabilim Dalı, Matematik Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Mutlu AKAR
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Doç. Dr. Mutlu AKAR, Danışman
Yıldız Teknik Üniversitesi _____

Doç. Dr. Muttalip ÖZAVŞAR, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Fatih TEMİZ, Üye
Üsküdar Üniversitesi _____

Danışmanım Doç. Dr. Mutlu AKAR sorumluluğunda tarafımca hazırlanan "Makine Öğrenmesi Sınıflandırma Algoritmalarının ve Parametrelerinin Göğüs Kanseri Tanı Tahminleri" başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Yunus Emre DEMİREL

Aileme

ve

eşime

TEŞEKKÜR

Öncelikle bu tez çalışmasını tamamlamamda emeği geçen tez danışmanım Doç. Dr. Mutlu AKAR'a sonsuz teşekkürler. Profesyonelliği, bilgi birikimi, tecrübesi, güncel hayatı takip etmesi, rehberliği, yönlendirmeleri, geri bildirimleri, eleştirel analizleri, desteği, motivasyonu, iş birliği, katkıları, çalışma azmi, değerli zamanını ayırması ve bana güvenmesi sonucu bu çalışmayı bitirmiş bulunmaktayım.

Başta annem ve babam olmak üzere eşim ve tüm aileme her zaman bana karşı gösterdikleri sevgileri ve destekleri için çok teşekkür ederim.

Tezin yürütülmesindeki destekleri için Yıldız Teknik Üniversitesi Matematik Bölümü ve Fen Bilimleri Enstitüsü üyelerine de teşekkür ederim.

Ayrıca bu çalışmada kullandığım veri setini oluşturan tüm araştırmacılara da teşekkürü borç bilirim.

Yunus Emre DEMİREL

İÇİNDEKİLER

KISALTMA LİSTESİ vi
ŞEKİL LİSTESİ viii
TABLO LİSTESİ ix
ÖZET x
ABSTRACT xii
1 GİRİŞ 1
1.1 Tezin Amacı 1
1.2 Literatür Taraması 1
2 TANIM VE EPİDEMİYOLOJİ 4
2.1 Kanser 4
2.2 Göğüs Kanseri ve Türleri 4
2.3 Göğüs Kanseri Epidemiyolojisi 6
2.4 Göğüs Kanseri Tanı Yöntemleri ve Yanlış Tanılar 8
2.5 Göğüs Kanseri Tanısında Makine Öğrenmesi Avantajları 9
3 MAKİNE ÖĞRENMESİ 11
3.1 Makine Öğrenmesi Temelleri 11
3.2 Makine Öğrenmesi Yaklaşımları 11
3.2.1 Gözetimli/ Denetimli Öğrenme 12
3.2.2 Gözetimsiz/ Denetimsiz Öğrenme 15
3.2.3 Pekiştirmeli/ Takviyeli Öğrenme 15
4 MATERYAL VE METOT 16
4.1 Veri Seti Hikayesi 16
4.2 Veri Manipülasyonu 19
4.3 Göğüs Kanseri Tanısında Makine Öğrenmesi Kullanımı 29
4.4 Karışıklık Matrisi (Confusion Matrix) 30
4.5 Performans Metrikleri 31
4.6 Makine Öğrenmesi Algoritmalarının Performansları 35
4.7 Bazı Algoritma Parametrelerinin Performansa Etkileri 37
5 SONUÇ 43
KAYNAKÇA 46
TEZDEN ÜRETİLMİŞ YAYINLAR 49

KISALTMA LİSTESİ

AB	AdaBoost
AC	Accuracy (Doğruluk)
AGD	Adam Gradient Descent (Adam Gradyan İnişi)
ANN	Artificial Neural Network (Yapay Sinir Ağı)
AUC	Area Under Curve (Eğri Altındaki Alan)
B	Benign (İyi Huylu)
BNB	Bernoulli Naïve Bayes
CART	Classification and Regression Trees (Sınıflandırma ve Regresyon Ağaçları)
CB	CatBoost
CV	Cross Validation (Çapraz doğrulama)
DCIS	Ductal carcinoma in situ
DT	Decision Tree (Karar Ağacı)
F	F1 score (F1 skoru)
FDR	False Discovery Rate (Yanlış keşif oranı)
FN	False Negative (Yanlış Negatif)
FNA	Fine Needle Aspiration (İnce İğne Aspirasyonu)
FNR	False Negative Rate (Yanlış negatif oranı)
FP	False Positive (Yanlış Pozitif)
FPR	False Positive Rate (Yanlış pozitif oranı)
GB	Gradient Boosting
GPC	Gauss Process Classification (Gauss Süreç Sınıflandırması)
HGB	Hist Gradient Boosting
ICBC	Iranian Center for Breast Cancer (İran Göğüs Kanseri Merkezi)
ID	Identification (Kullanıcı Numarası)
KFCV	K Fold Cross Validation (k katlı çapraz doğrulama)
KNC	K Neighbors Classifier (K Komşu Sınıflandırıcı)
KNN	K Nearest Neighbors (K En Yakın Komşu)
LCIS	Lobüler carcinoma in situ
LDA	Linear Discriminant Analysis (Lineer Diskriminant Analizi)
LR	Logistic Regression (Lojistik Regresyon)
M	Malignant (Kötü Huylu)

MCC	Matthews Correlation Coefficient (Matthews Korelasyon Katsayısı)
MLP	Multi Layer Perceptron (Çok Katmanlı Algılayıcı)
NB	Naïve Bayes
NC	Nearest Centroid (En Yakın Merkez Nokta)
NN	Neural Network (Sinir Ağı)
NPV	Negative Predictive Value (Negatif tanı değeri)
PA	Passive Aggressive (Pasif Agresif)
poly	Polynomial (Polinom)
PR	Precision (Kesinlik)
PR-AUC	Precision-Recall Area Under Curve (Kesinlik-Duyarlılık Eğrisi Altında Kalan Alan)
QDA	Quadratic Discriminant Analysis (Kuadratik Diskriminant Analizi)
rbf	Radial Basis Function (Radyal Tabanlı Fonksiyon)
RE	Recall (Duyarlılık)
RF	Random Forest (Rastgele Orman)
ROC	Receiver Operating Characteristic (Alıcı İşlem Karakteristiği)
ROC-AUC	Receiver Operating Characteristic Area Under Curve (Alıcı İşlem Karakteristiği Eğrisi Altında Kalan Alan)
ROF	Rotation Forest (Rotasyon Orman)
SE	Sensitivity (Hassasiyet)
SGD	Stochastic Gradient Descent (Stokastik Gradyan İnişi)
SP	Specificity (Belirleyicilik)
SVC	Support Vector Classification (Destek Vektör Sınıflandırması)
SVM	Support Vector Machines (Destek Vektör Makineleri)
TN	True Negative (Doğru Negatif)
TP	True Positive (Doğru Pozitif)
UCI	University of California, Irvine (Kalifornia Üniversitesi Irvine Yerleşkesi)
WBC	Wisconsin Breast Cancer Dataset (Wisconsin Üniversitesi Göğüs Kanseri Veri Seti)
WDBC	Wisconsin Diagnostic Breast Cancer Dataset (Wisconsin Üniversitesi Tanısal Göğüs Kanseri Veri Seti)
XGB	XGBoost

ŞEKİL LİSTESİ

Şekil 2.1 Her iki cinsiyet ve her yaş grubu için kanser tanı sayıları **7**
Şekil 2.2 Her iki cinsiyet ve her yaş grubu için kanserden ölenlerin sayıları **7**
Şekil 2.3 Kadınlarda her yaş grubu için kanser tanı sayıları **8**
Şekil 4.1 Xcyt kullanımı (1990) **16**
Şekil 4.2 Veri setinin yapısal bilgileri **20**
Şekil 4.3 Tanı hedef değişkeninin sınıf frekansları **21**
Şekil 4.4 Veri setinin betimsel istatistikleri **22**
Şekil 4.5 Öznitelikler arası ikili ilişkiler 1 ..... **23**
Şekil 4.6 Öznitelikler arası ikili ilişkiler 2 ..... **24**
Şekil 4.7 Öznitelikler arası ikili ilişkiler 3 ..... **25**
Şekil 4.8 Öznitelikler arası ikili ilişkiler 4 ..... **26**
Şekil 4.9 symmetry2 ve concave_points3 arasındaki ilişki **27**
Şekil 4.10 symmetry2, concave_points3, radius1 arasındaki ilişki **28**
Şekil 4.11 radius1 dağılımı **28**
Şekil 4.12 random_state parametresi kontrolü **29**
Şekil 4.13 AC ile K değerleri arasındaki ilişki **38**

TABLO LİSTESİ

Tablo 3.1 Makine öğrenmesi yaklaşımları **12**
Tablo 4.1 Temel 10 öznitelik **17**
Tablo 4.2 Karışıklık matrisi genel görünümü **31**
Tablo 4.3 Metrikler **35**
Tablo 4.4 Algoritma performansları **36**
Tablo 4.5 Algoritma performansları (10 katlı çapraz doğrulama ile) **36**
Tablo 4.6 Parametre performansları **42**
Tablo 5.1 Performans kıyasları **43**

ÖZET

Makine Öğrenmesi Sınıflandırma Algoritmalarının ve Parametrelerinin Göğüs Kanseri Tanı Tahminleri

Yunus Emre DEMİREL

Matematik Anabilim Dalı

Matematik Programı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Mutlu AKAR

Makine öğrenmesi, veri analitiği ve bilişim alanlarında hızla ilerleyen ve geniş uygulama yelpazesiyle dikkat çeken bir disiplindir. Bu alandaki algoritmalar, veri temelli modeller oluşturarak etkili ve doğru kararlar alabilme yetenekleri sayesinde büyük ilgi görmektedir. Özellikle göğüs kanseri gibi karmaşık hastalıkların tanısında makine öğrenmesi algoritmaları kritik bir rol oynamaktadır. Dünya genelinde kadınlar arasında en yaygın kanser türü olan göğüs kanserinin erken tanısı, tedavi başarısını ve hasta yaşam kalitesini önemli ölçüde artırmaktadır. Bu çalışma, çeşitli makine öğrenmesi algoritmalarının göğüs kanseri tanısındaki performanslarını detaylı bir şekilde karşılaştırarak, klinik uygulamalara ışık tutmayı, daha doğru tanıların yapılmasına katkıda bulunmayı ve makine öğrenmesi teknolojilerinin tıbbi tanı süreçlerindeki potansiyelini ortaya koyarak, sağlık hizmetlerinde inovasyon ve gelişime katkı sağlamayı amaçlamaktadır. Çalışmamızda, göğüs kanseri araştırmalarında yaygın olarak kullanılan LR, SVM-SVC, KNN-KNC, NB-BNB, DT, RF, AB ve NN-MLP algoritmalarının yanı sıra SGD, Ridge, Perceptron, PA, LDA, NC, GPC, GB ve HGB olmak üzere 17 farklı makine öğrenmesi sınıflandırma algoritmasının göğüs kanseri tanısındaki başarıları

incelenmiştir. Her bir algoritma için modeller WDBC veri seti kullanılarak eğitilmiştir. Algoritmaların performansları AC, SE/RE, SP, PR, F, NPV, FPR, FDR ve FNR olmak üzere 9 farklı metrik ile değerlendirilmiştir. Ayrıca algoritmaların çeşitli parametre ayarlarının etkileri de incelenmiş olup, DT için min_samples_split, min_samples_leaf ve min_weight_fraction_leaf; RF için n_estimators parametrelerinin varsayılan değerlerinden farklı bazı değerleri için varsayılan değerlere göre algoritmaların daha yüksek tanı başarı oranları sağladıkları görülmüştür. Bu çalışma, göğüs kanseri tanısında makine öğrenmesi algoritmalarının ve parametrelerinin etkilerini inceleyerek, bu alandaki literatüre önemli bir katkı sunmayı hedeflemektedir. Kapsamlı analizimiz, farklı algoritmalar ve parametre ayarlarının göğüs kanseri tanısında nasıl bir performans sergilediğini ortaya koyarak, hangi yöntemlerin daha etkili olduğunu belirlemeye yardımcı olmaktadır. Sonuçlar, makine öğrenmesi algoritmalarının klinik uygulamalarda nasıl daha iyi kullanılabileceğine dair değerli bilgiler sağlamaktadır.

Anahtar Kelimeler: WDBC, göğüs kanseri tanısı, makine öğrenmesi, sınıflandırma algoritmaları.

ABSTRACT

Breast Cancer Diagnostic Predictions of Machine Learning Classification Algorithms and Its Parameters

Yunus Emre DEMİREL

Department of Mathematics

Master of Science Thesis

Supervisor: Assoc. Prof. Dr. Mutlu AKAR

Machine learning is a rapidly advancing discipline with numerous applications in data analytics and informatics. Algorithms in this field are valued for creating data-driven models that enable effective and accurate decision-making. Particularly, machine learning algorithms play a critical role in diagnosing complex diseases such as breast cancer. Early diagnosis of breast cancer, the most common cancer among women worldwide, significantly improves treatment success and patient quality of life. This study aims to enhance clinical practices by comparing the performance of various machine learning algorithms in diagnosing breast cancer, contributing to more accurate diagnoses, and showcasing the potential of these technologies in medical diagnosis, thus promoting innovation in healthcare services. In our study, the success of 17 different machine learning classification algorithms, including commonly used ones in breast cancer research such as LR, SVM-SVC, KNN-KNC, NB-BNB, DT, RF, AB, and NN-MLP, as well as SGD, Ridge, Perceptron, PA, LDA, NC, GPC, GB, and HGB, was examined in diagnosing breast cancer. Models for each algorithm were trained using the WDBC dataset. The performance of the algorithms was evaluated with nine metrics: AC, SE/RE, SP, PR, F, NPV, FPR, FDR, and FNR. Additionally, the effects of various parameter settings on the algorithms were examined, and it was observed that for

DT and RF, yielded higher diagnostic success rates than their default values. This study contributes to the literature by examining the effects of machine learning algorithms and parameters in breast cancer diagnosis. Our analysis reveals how different algorithms and parameter settings perform in breast cancer diagnosis, helping to determine which methods are more effective. The results provide valuable insights into how machine learning algorithms can be better utilized in clinical applications.

Keywords: WDBC, breast cancer diagnosis, machine learning, classification algorithms.

1
GİRİŞ

1.1 Tezin Amacı

Bu tez çalışmasının amacı, makine öğrenmesi gözetimli/ denetimli sınıflandırma algoritmalarının ve bu algoritmaların bazı parametrelerinin farklı değerlerinin, göğüs kanseri tanı tahminlerine olan etkilerini karşılaştırmaktır. Böylece çeşitli makine öğrenmesi gözetimli/ denetimli öğrenme sınıflandırma algoritmalarının ve parametrelerinin performanslarını değerlendirerek klinik uygulamalara rehberlik edecek ve daha doğru tanıların yapılmasına katkı sağlayacak değerli bir analiz sunmak amaçlanmıştır.

1.2 Literatür Taraması

S. Aruna ve çalışma arkadaşları (Aruna, Rajagopalan, & Nandakishore, 2011), WBC, WDBC ve UCI veri deposundan Göğüs Dokusu olmak üzere üç adet ikili ve çoklu sınıf veri seti kullanarak NB, DT, NN, SVM algoritmalarının performanslarını AC, PR, SE ve SP açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. SVM radyal tabanlı fonksiyon çekirdeği, AC, SE/RE ve SP metrikleri açısından diğer sınıflandırıcılara göre hem ikili hem de çoklu sınıf veri setleri için daha iyi performans göstermiştir.

Ahmad L. G. ve çalışma arkadaşları (Ahmad, Eshlaghy, Poorebrahimi, Ebrahimi, & Razavi, 2013), 1997 ile 2008 yılları arasında ICBC tarafından oluşturulan veri setini kullanarak DT, SVM, ANN algoritmalarının performanslarını SE, SP ve AC açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. SVM, en düşük hata oranı ve AC açısından diğer sınıflandırma yöntemlerine kıyasla daha üstün performans sergilemiştir.

D. Oyewola ve çalışma arkadaşları (Oyewola, Hakimi, Adeboye, & Shehu, 2016), WDBC veri setini kullanarak LR, LDA, QDA, RF ve SVM algoritmalarının performanslarını AC, SE ve SP açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. SVM, AC ve SE açısından diğer sınıflandırıcılarla karşılaştırıldığında daha başarılı sonuçlar elde etmiştir.

H. Asri ve çalışma arkadaşları (Asri, Mousannif, Al Moatassime, & Noel, 2016), WBC veri setini kullanarak SVM, DT, NB, KNN algoritmalarının performanslarını AC, PR, SE ve SP açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. SVM sınıflandırıcısının AC ve en düşük hata oranı açısından diğer sınıflandırıcılara göre daha etkili olduğu görülmüştür.

Y. Khourdifi ve M. Bahaj (Khourdifi & Bahaj, 2018), WBC veri setini kullanarak RF, NB, SVM ve KNN algoritmalarının performanslarını AC, RE ve PR açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. SVM, AC, RE ve PR açısından diğer sınıflandırıcılara göre daha yüksek performans göstermiştir.

M. Amrane ve çalışma arkadaşları (Amrane, Oukid, Gagaoua, & Ensari, 2018), WBC veri setini kullanarak NB ve KNN algoritmalarının performanslarını AC açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. KNN, AC açısından NB algoritmasına göre daha iyi sonuçlar vermiştir.

S. Sharma ve çalışma arkadaşları (Sharma, Aggarwal, & Choudhury, 2018), WDBC veri setini kullanarak RF, KNN ve NB algoritmalarının performanslarını AC, SE/RE, PR ve F açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. KNN, AC, PR, RE ve F açısından diğer sınıflandırıcılara göre daha etkili bir başarı göstermiştir.

L. Westerdijk ve S. Bhulai (Westerdijk & Bhulai, 2018), WDBC veri setini kullanarak LR, RF, SVM ve NN algoritmalarının performanslarını AC, SE, SP ve ROC-AUC açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. RF, AC açısından diğer sınıflandırıcılara göre daha üstün bir verimlilik sunmuştur. Diğer sonuçlar ilgili çalışmada tablolarla verilmiştir.

M. F. Ak (Ak, 2020), WDBC veri setini kullanarak LR, KNN, SVM, NB, DT, RF ve ROF algoritmalarının performanslarını AC açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemiştir. LR, AC açısından diğer sınıflandırıcılara göre daha iyi sonuçlar elde etmiştir.

M. M. Islam ve çalışma arkadaşları (Islam, ve diğerleri, 2020), WBC veri setini kullanarak SVM, KNN, RF, ANN ve LR algoritmalarının performanslarını AC, SE, SP, PR, NPV, FNR, FPR, F, MCC, PR-AUC ve ROC-AUC açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. ANN, AC, PR ve F açısından diğer sınıflandırıcılara göre daha başarılı olmuştur.

P. Gupta ve S. Garg (Gupta & Garg, 2020), WDBC veri setini kullanarak KNN, LR, DT, RF, SVM ve AGD algoritmalarının performanslarını AC, PR, RE ve F açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. AGD, AC, PR, RE ve F açısından diğer sınıflandırıcıların önüne geçmiştir.

M. A. Naji ve çalışma arkadaşları (Naji, ve diğerleri, 2021), WDBC veri setini kullanarak SVM, RF, LR, DT ve KNN algoritmalarının performanslarını AC, PR, F ve AUC açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. SVM, AC, PR, SE/RE ve F açısından diğer sınıflandırıcılara göre daha üstün başarı göstermiştir.

S. Sakib ve çalışma arkadaşları (Sakib, ve diğerleri, 2022), WDBC veri setini kullanarak SVM, DT, LR, RF, KNN ve NN algoritmalarının performanslarını AC, RE, SP, PR, F, FNR, FPR ve MCC açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. RF, AC ve F açısından diğer sınıflandırıcılara göre daha yüksek etkinlik göstermiştir.

I. Chhillar ve A. Singh (Chhillar & Singh, 2024), WDBC veri setini kullanarak SVM, DT, RF, NB, KNN, XGB, AB ve CB algoritmalarının performanslarını AC, PR, RE, AUC ve ROC açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. SVM ve RF diğer sınıflandırıcılara göre daha yüksek başarı sağlamıştır.

O. Sevli (Sevli, 2019), WDBC veri setini kullanarak SVM, NB, RF, KNN ve LR algoritmalarının performanslarını AC, SE/RE, SP, PR açısından karşılaştırarak en iyi sınıflandırıcıyı bulmak istemişlerdir. LR sınıflandırıcısının diğer sınıflandırıcılara göre daha verimli olduğu gözlemlenmiştir.

Tüm bunların yanı sıra, farklı makine öğrenmesi yöntemleri, erken tanının kritik olduğu çeşitli hastalıklar ve kanser türlerinin tespitinde kullanılmış ve başarılı sonuçlar elde edilmiştir (Akar & Sirakov, 2019), (Akar, Sirakov, & Mete, 2022), (Badik & Akar, 2024). Bu durum, erken tanı için makine öğrenmesi ve derin öğrenme modellerinin son derece etkili ve tercih edilebilir olduğunu ortaya koymaktadır.

2
TANIM VE EPİDEMİYOLOJİ

2.1 Kanser

Hayatımız boyunca vücudumuzda bulunan milyarlarca hücre; yeni hücreler üretmek için büyüme, gelişme, yenilenme, onarım ve üreme gibi birçok sebeple büyür ve bölünür. Hücreler yaşlanınca veya hasar görünce ölürler. Ölen hücrelerin yerine bölünerek oluşan yeni hücreler geçer. Fakat bazen hücreler yaşlandıklarında veya genetik ya da çevresel faktörler sebebi ile hasar gördüklerinde ölmeleri gerekirken hayatta kalıp gereksiz şekilde bölünerek yeni hücreler oluşturmaya devam ederler. Bu şekilde vücudumuzda normalin dışında hareket eden hücrelerin kontrol dışı, düzene aykırı bir şekilde çoğalması kanser olarak tanımlanır. Kanser hücreleri çoğalarak tümör adı verilen doku kütleleri oluşturlar. Temelde iki tür tümör sınıfı mevcuttur: İyi huylu (benign) ve kötü huylu (malignant). Kötü huylu tümörler büyüdükçe oldukları yerden kopup kan veya lenf sistemi aracılığı ile vücudun farklı bölgelerine ulaşabilir (metastaz) ve ulaştıkları bölgelerde yeni tümörler oluşturabilir. Aksine iyi huylu tümörler, çevre dokulara yayılmaz veya çevre dokuları istila etmezler. Bununla birlikte, iyi huylu tümörler bazen oldukça büyük olabilmektedirler. İyi huylu tümörler tedavi edildikten sonra genellikle tekrar meydana gelmezler, fakat kötü huylu tümörler tedavi edildikten sonra tekrar meydana gelebilirler (National Cancer Institute, 2021).

2.2 Göğüs Kanseri ve Türleri

Göğüs kanseri, göğüs içinde başlayan bir kanser türüdür. Tek veya her iki göğüste de başlayabilmektedir. Göğüs kanseri, göğsün farklı bölgelerinde başlayabilir (American Cancer Society, 2024). Göğüs kanserinin birçok çeşidi mevcuttur. Göğüs kanseri çeşitleri, çevre dokulara yayılıp yayılmadığına, oluştuğu dokuya, hormonlara, kanserin üretebileceği proteinlere veya genlere göre farklı şekillerde isimlendirilebilmektedir. Temel olarak çevre dokulara yayılıp yayılmama durumuna göre göğüs kanseri, invaziv olmayan (noninvaziv/ in situ) ve invaziv olmak üzere iki ana gruba ayrılmaktadır. Göğüs kanseri, çevre dokulara yayılma durumu söz konusu değilse invaziv olmayan, çevre dokulara yayılma durumu söz

konusu ise invaziv olarak isimlendirilmektedir. İnvaziv olmayan göğüs kanserleri de kendi içinde oluştuğu dokuya göre farklı şekillerde isimlendirilmektedir (Türk Tıbbi Onkoloji Derneği, 2024). Örneğin; duktal carcinoma (karsinom) in situ (DCIS) ve lobüler carcinoma (karsinom) in situ (LCIS).

Kadınlarda süt üretim sistemi oldukça karmaşık bir yapıdadır. Bu sistem, ince tüp şeklinde kanallardan oluşur. Bu kanallara süt kanalları (duktuslar) denmektedir. Süt kanalları, süt üreten bezlerden (lobüller) sütü toplar ve göğüs ucuna doğru taşır. Her bir süt kanalı göğüs ucuna doğru ilerlerken, göğüs dokusunda dallanmalar yapar ve giderek daha küçük kanallara ayrılır. Bu kanallar göğüs dokusunun her tarafına yayılmaktadır ve göğüs ucuna kadar sütü taşımak için karmaşık bir ağ yapısı oluşturmaktadır. Göğüs ucunda süt kanalları birleşerek daha büyük bir kanal oluştururlar ve göğüs ucunda bir açıklıktan süt dışarıya verilir (Memorial Sloan Kettering Cancer Center, 2024). İşte bu süt kanallarında oluşan ve invaziv olmayan tümöre DCIS, süt bezlerinde oluşan ve invaziv olmayan tümöre ise LCIS adı verilmektedir (Türk Tıbbi Onkoloji Derneği, 2024). Karsinom, göğüs kanalları da dahil olmak üzere dokuların iç veya dış yüzeylerini oluşturan epitelyal doku hücrelerinden başlayan kanser anlamına gelmektedir. In situ (in SY-too) ise Latince "orijinal yerinde" yani kanserin başladığı bölge ile sınırlı kaldığı, çevre dokulara yayılmadığı anlamına gelmektedir (Memorial Sloan Kettering Cancer Center, 2024). Örneğin; DCIS terimi, göğüs kanserinin göğüs kanallarının yüzeyindeki hücrelerde başladığını ve kanser hücrelerinin yerinde bulunduğunu yani çevre dokulara yayılmadığını ifade etmektedir.

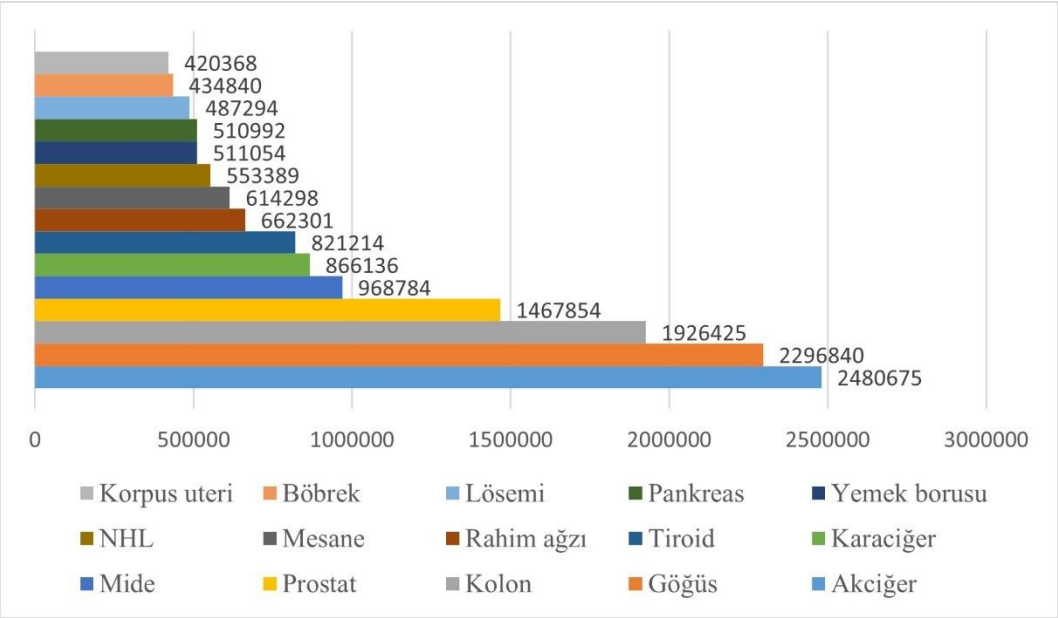
İnvaziv göğüs kanserinin ise birçok çeşidi mevcuttur. Örneğin; süt kanallarında başlayan ve invaziv olan göğüs kanser tipine invaziv duktal karsinom, süt bezlerinde başlayan ve invaziv olan göğüs kanser tipine invaziv lobüler karsinom denmektedir. Diğer bazı invaziv kanser türleri de şunlardır; invaziv sarkom, invaziv papiller karsinom, invaziv medüller karsinom, invaziv mukozal karsinom, metaplazik invaziv karsinom, invaziv tübüler karsinom, invaziv mikst karsinom, skuamöz hücreli karsinom, invaziv paget hastalığı, invaziv papillotoksik karsinom, kribriform invaziv karsinom, invaziv kistoadenokarsinom, invaziv adeno-kistik karsinom, sınırlı mikst karsinom, invaziv mikropapiller karsinom, invaziv sürüngen hücreli karsinom (bazal tip göğüs kanseri), invaziv hemanjioperitsitoma, invaziv anjiyosarkom, müsinöz göğüs kanseri (invaziv müsinöz karsinom veya kolloidal

karsinom), enflamatuar göğüs kanseri, invaziv dalgalı hücreli karsinom, metaplastik göğüs kanseri, filloides tümörleri, göğüs anjiyosarkomu, göğüs lenfoması.

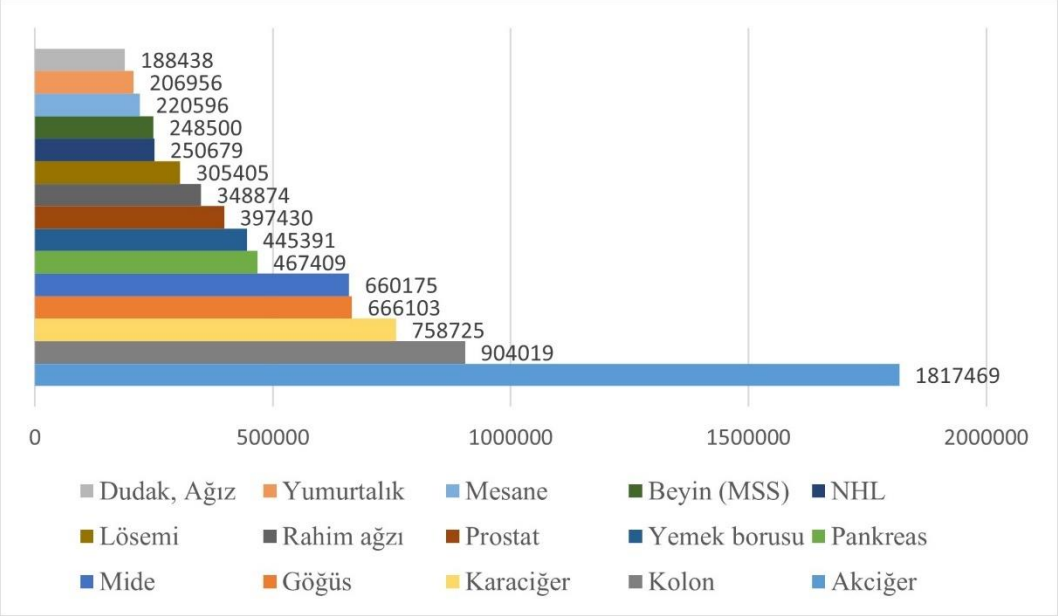
Hormonlara göre sınıflandırılan göğüs kanser çeşitlerinden bazıları ise şunlardır (Türk Tıbbi Onkoloji Derneği, 2024); hormon reseptörü pozitif (örneğin, östrojen reseptörü pozitif veya ER+, yani östrojen için reseptörlerin mevcut olduğu durum), hormon reseptörü negatif (örneğin, östrojen reseptörü negatif veya ER-, yani östrojen için reseptörlerin mevcut olmadığı durum), progesteron reseptörü pozitif veya PR+ (yani progesteron için reseptörlerin mevcut olduğu durum), progesteron reseptörü negatif veya PR- (yani progesterone için reseptörlerin mevcut olmadığı durum), HER2 (insan epidermal büyüme faktörü reseptörü 2) pozitif (HER2+), HER2 (insan epidermal büyüme faktörü reseptörü 2) negatif (HER2-). HER2 sağlıklı göğüs hücrelerinin büyüme kontrolünde yardımcı bir proteindir. Büyümeyi teşvik eden bir proteindir. HER2 pozitif / HER2+ göğüs kanseri, kanser hücrelerinin normal değerinden daha yüksek HER2 proteini içerdiği durumdur. HER2 pozitif / HER2+, diğer göğüs kanseri tiplerinden daha çabuk büyüme ve yayılma eğilimindedir (Tenekeci, 2024).

2.3 Göğüs Kanseri Epidemiyolojisi

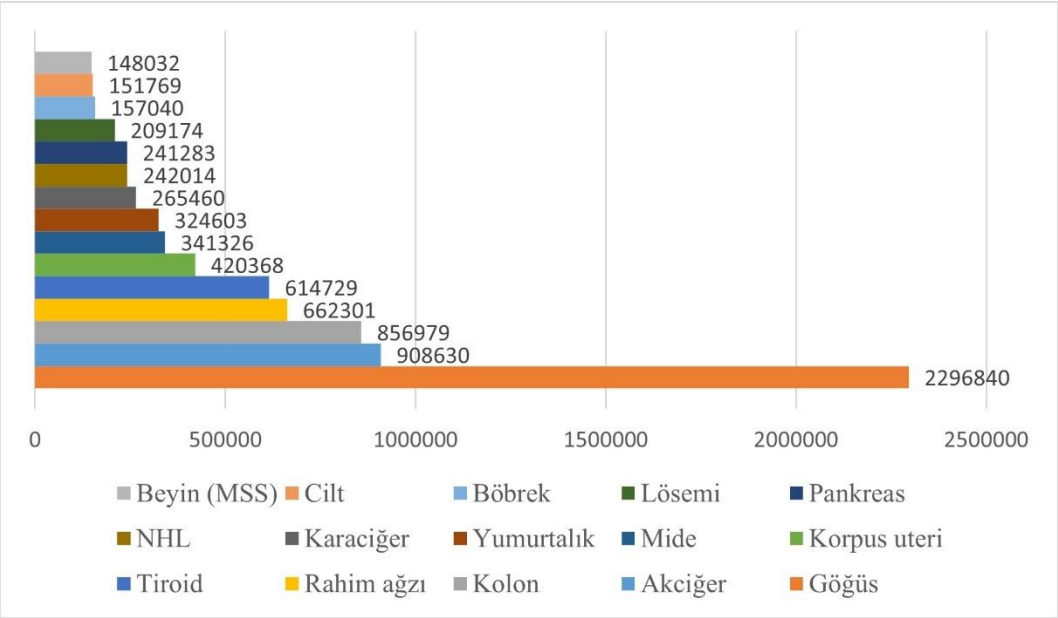
2022 yılında dünya genelinde en sık görülen akciğer kanserinden sonra en sık görülen ikinci kanser türü 2 296 840 yeni vaka ile göğüs kanseri (breast cancer) olmuştur (Şekil 2.1). Bu vakaların 666 103'ü ölümle sonuçlanmıştır. Ölüme sebep olan diğer kanser türleri arasında göğüs kanseri 4. sırada yer almaktadır (Şekil 2.2). 2 296 840 yeni vakanın tamamını kadınlar oluşturmaktadır. Bunun yanı sıra 2022 yılında kadınlar arasında en sık görülen kanser türü göğüs kanseri olmuştur (Şekil 2.3). Nadir de olsa göğüs kanseri, kadınlara nazaran erkeklerde de görülebilmektedir (Sung, ve diğerleri, 2022).



Şekil 2.1 Her iki cinsiyet ve her yaş grubu için kanser tanı sayıları



Şekil 2.2 Her iki cinsiyet ve her yaş grubu için kanserden ölenlerin sayıları



Şekil 2.3 Kadınlarda her yaş grubu için kanser tanı sayıları

2.4 Göğüs Kanseri Tanı Yöntemleri ve Yanlış Tanılar

Göğüs kanserinin özellikle kadınlarda çok sık görülmesi, vaka sayılarının gün geçtikçe artıyor olması, ayrıca makine öğrenmesi gibi günümüz teknolojileri ile erken tanının daha kolay konulması, erken tanı durumunda tedavinin olumlu karşılık verebilmesi gibi sebepler ile göğüs kanseri tarama testleri daha da önemli hale gelmektedir.

Göğüs kanserinin erken tanısı veya kanserin varlığının araştırılması için kadınlara bazı görüntüleme ve tarama testleri önerilmektedir. El ile kendi kendine tarama, hekim tarafından yapılan fizik ve klinik muayene, manyetik rezonans (MR), mamografi (film mamografisi, dijital mamografi, tomosentez mamografi, 2 boyutlu mamografi), ultrasonografi, positron emisyon tomografi (PET), biyopsi gibi tarama testleri göğüs kanseri tanısında kullanılan tarama testlerinin başlıcalarıdır.

Göğüs kanseri tanısı ve kanserin varlığının araştırılması için kullanılan bu görüntüleme ve tarama testlerinin sonuçları her zaman doğru çıkmayabilir. Tarama testi sonuçları, kişide kanser olmamasına rağmen anormal görünebilir. Buna yanlış pozitif test sonucu denir. Kanser olmamasına rağmen anormal tarama test sonuçları sebebi ile genellikle riskleri de olan daha fazla tarama testi veya cerrahi işlem (biyopsi gibi) hekimler tarafından talep edilebilir. Anormal test sonuçlarının çoğunun kanser olmadığı ortaya çıkmaktadır. Kişinin kanser olmadığı halde test

sonuçlarının anormal çıktığı grup genellikle; 50 yaş altı kadınlardan, daha önce biyopsi yapılmış kadınlardan, ailesinde göğüs kanseri öyküsü olan kadınlardan ve menopoz için hormon alan kadınlardan oluşmaktadır. Bununla birlikte yanlış pozitif sonuçlar (kişi aslında kanser olmadığı halde kişinin kanser olduğunu gösteren sonuçlar) birkaç günden yıllar sonraya kadar dahi kaygıya neden olabilir.

Tüm bunların aksine kişide göğüs kanseri mevcut olsa bile tarama testi sonuçları normal görünebilir. Buna yanlış negatif test sonucu denir. Yanlış negatif (aslında kişi kanser olduğu halde kişinin kanser olmadığını gösteren sonuçlar) sonuca sahip olan kadınlar, semptomları olsa bile tedavi edilmeyi erteleyebilmektedir. Mamografi testleri kullanılarak elde edilen yanlış negatif test sonuçları sebebi ile yaklaşık 5 kanserli kişiden 1 kişi gözden kaçmaktadır. Kişi kanser olduğu halde test sonuçlarının normal çıktığı grup genellikle; genç kadınlardan, yoğun göğüs dokusuna sahip olan kadınlardan ve hızla büyüyen kansere sahip kadınlardan oluşmaktadır (National Cancer Institute, 2023).

Tüm bu sebeplerden ötürü, göğüs kanseri tanısının konulmasında, kanserin varlığının araştırılmasında tanıyı koyacak ve araştırmayı yapacak olan uzman/ uzmanlara daha doğru tanılar ve sonuçlar elde ederek yardımcı olmak amacı ile makine öğrenmesi sınıflandırma algoritmalarından faydalanılabilir.

2.5 Göğüs Kanseri Tanısında Makine Öğrenmesi Avantajları

Göğüs kanseri tanısında makine öğrenmesi kullanımının pek çok avantajı mevcuttur. Makine öğrenmesi algoritmaları, insanlara nazaran büyük veri setlerini çok daha hızlı bir şekilde analiz edebilir ve karmaşık desenleri daha kolay tespit edebilir. Örneğin, mamografi görüntülerini değerlendirmek için kullanılan makine öğrenmesi algoritmaları, potansiyel anormallikleri belirlemek için milyonlarca pikseli çok hızlı bir şekilde tarayabilir. Tarama sonuçlarını insan gözünden kaçabilecek ince detayları yüksek hassasiyet ve doğruluk oranı ile analiz eden makine öğrenmesi algoritmaları, insan önyargılarından arınmış objektif bir yaklaşım sunar. Bu sayede göğüs kanseri tanı süreci oldukça hızlandırılabilir ve doğru tanı oranı ve verimlilik arttırılabilir. Ayrıca makine öğrenmesi algoritmaları, hastaların tıbbi geçmişlerini, genetik profillerini ve diğer faktörleri de analiz ederek kişiye özel tedavi planları önerme konusunda uzmanlara yardımcı olabilir. Bu, her hasta için en etkili tedavi stratejisinin belirlenmesine katkıda bulunabilir. Öte

yandan makine öğrenmesi algoritmaları, tıbbi uzmanların eğitimini desteklemek ve uzmanlık alanlarını geliştirmek için kullanılabilir. Ayrıca büyük veri setleri üzerinde yapılan analizler, yeni bilgilerin keşfedilmesine ve sağlık profesyonellerinin bilgi birikimlerinin artmasına olanak tanıyabilir. Bununla birlikte uzmanların tıbbi kaynakları daha verimli kullanmalarına yardımcı olarak uzmanların zamanlarını daha verimli bir şekilde kullanmalarını sağlayabilir ve böylece uzmanlar daha kısa sürede daha fazla hastaya hizmet verebilir.

3
MAKİNE ÖĞRENMESİ

3.1 Makine Öğrenmesi Temelleri

Bilgisayarların insanlardan daha hızlı işlem yapabilme yeteneklerini kullanarak insanlığa faydalı çalışmalar sunmak ve başarıya daha hızlı bir şekilde ulaşmak için son zamanlarda çalışılan en popüler alanlardan birisi de bir yapay zekâ alt dalı olan makine öğrenmesi alanıdır. Makine öğrenmesi terimi, ilk kez 1959'da yayınlanan "Dama Oyununu Kullanarak Makine Öğrenmesi Üzerine Bazı Çalışmalar (Samuel, 1959)" adlı eserin sahibi Arthur L. Samuel ile popüler hale gelmiştir (Samuel, 1959). Makine öğrenmesi teriminin birçok tanımı yapılabilir. Makine öğrenmesi, bilgisayarlara açıkça programlanmadan öğrenme yeteneği kazandıran bir çalışma alanıdır, makine öğrenmesi, bilgisayarları veriden öğrenebilecek şekilde programlamanın bilimi ve sanatıdır (Géron, 2019). Makine öğrenmesi, veri ve algoritmalar sayesinde bilgisayarların veri analizi yaparak ve modeller kurarak öğrenme yeteneği kazanmasını sağlayan bir yapay zekâ alt dalıdır. Makine öğrenmesi, algoritmalar kullanarak veri seti üzerinde matematiksel ve istatistiksel modellemeler kurar, bu modeller ile veri analizi yapar ve gelecekteki kararlar için bilgisayarların tahminlerde bulunmasını sağlar. Günümüzde makine öğrenmesi çalışmalarının temelde iki amacı mevcuttur. Bu amaçlar, algoritmalar ile oluşturulan modellere göre verileri sınıflandırmak ve geliştirilen bu modellere göre sonuçları henüz belli olmayan yeni verilerin sonuçlarını tahmin etmektir.

3.2 Makine Öğrenmesi Yaklaşımları

Makine öğrenmesi genellikle üç ana kategori altında incelenir: gözetimli/ denetimli, gözetimsiz/ denetimsiz ve pekiştirmeli/ takviyeli öğrenme (Tablo 3.1). Bu farklı yaklaşımlar, veri analizi ve öngörü yeteneklerini geliştirmek için kullanılan temel araçlardır. Bu yaklaşımlar, veri tabanlı problemlerin çözümünde çeşitli teknikler sunar. Makine öğrenmesi teknikleri, günlük yaşamda ve iş dünyasında kullanılarak çeşitli problemlerin çözümünde yardımcı olabilir.

Tablo 3.1 Makine öğrenmesi yaklaşımları

	Makine öğrenmesi yaklaşımları
1	Gözetimli/ Denetimli Öğrenme (Supervised Learning) i. Sınıflandırma (Classification) Problemleri ii. Regresyon (Regression) Problemleri
2	Gözetimsiz/ Denetimsiz Öğrenme (Unsupervised Learning)
3	Pekiştirmeli/ Takviyeli Öğrenme (Reinforcement Learning)

3.2.1 Gözetimli/ Denetimli Öğrenme

Gözetimli/ Denetimli öğrenme yaygın olarak kullanılan bir makine öğrenmesi yaklaşımıdır. Bu öğrenme yaklaşımı ile halihazırda girdileri (öznitelikler-features) ve çıktıları (hedef değişkenler) belli olan bir veri seti ve makine öğrenmesi algoritmaları kullanılarak model eğitilir. Eğitilen bu model ile girdiler ve çıktılar arasındaki ilişki bilgisayar tarafından öğrenilir. Böylece eğitilen model, yeni girdi verilerini aldığında çıktı verilerini doğru tahmin etme yeteneği kazanır. Gözetimli/ Denetimli öğrenme genel olarak veri toplama ile başlar. İlk adım, problem ile ilgili yeterli miktarda girdileri ve çıktıları olan veri seti oluşturmaktır. Veri seti oluşturulduktan sonra veri seti hikayesinin bilinmesi bir sonraki aşama olan veri ön işleme adımı için kritiktir. Veri ön işleme, veri setinin kalitesini arttırmak, veri analizinin ve makine öğrenmesi modellerinin daha doğru sonuçlar vermesini sağlamak için önem arz etmektedir. Genel olarak veri ön işleme aşamaları; öznitelik seçimi, veri temizliği, veri dönüşümü, veri bölme işlemlerinden oluşur.

Öznitelik seçimi, makine öğrenmesi model performansı için önemli bir aşamadır. Veri seti oluşturulurken özniteliklerin doğru seçilmesi ve uygun bir şekilde öznitelik mühendisliği yapılması, makine öğrenmesi modelinin performansını doğrudan etkileyebilir. Bu nedenle, öznitelik seçimi ve öznitelik mühendisliği, makine öğrenmesi projelerinde önemli bir rol oynar.

Oluşturulan veri setinde eksik, hatalı, bozuk, anlamsız, tutarsız veya aykırı veriler mevcut olabilir. Bu nedenle veri setindeki eksiklikler, hatalar, bozukluklar, anlamsızlıklar, tutarsızlıklar ve aykırılıklar giderilerek düzeltmeler ve bütünleştirmeler ile veri temizliği yapılır. Varsa tekrarlanan ya da kullanıcı

numarası gibi makine öğrenmesi için gerekli olmayan veriler, veri setinden çıkartılır.

Bazı makine öğrenmesi algoritmaları, yapısı gereği girdi verilerinin yani özniteliklerin aynı aralıkta olmasını ve mümkünse küçük sayılardan oluşmasını tercih ederler. Bu nedenle veri dönüşümü yapılmalıdır. Veri dönüşümü aşamasında girdi verileri yani öznitelikler ölçeklendirilebilir, dönüştürülebilir, standardizasyon ve normalizasyon yapılabilir. Öte yandan hedef değişkenler veya öznitelikler kategorik değişkenlerden oluşabilir. Bu kategorik hedef değişkenleri veya öznitelikler genellikle metin (string) tipinde olduğundan makine öğrenmesi algoritmaları bu kategorik hedef değişkenler veya öznitelikler ile düzgün çalışamayabilir. Bu nedenle makine öğrenmesi sınıflandırma algoritmalarının sağlıklı çalışması için varsa kategorik değişkenler sayısal değişkenlere dönüştürülür.

Veri ön işlemenin son aşaması olan veri bölme aşamasında ise veri seti rastgele iki parçaya ayrılır. Makine öğrenmesi projelerinde genellikle veri seti, %80 eğitim seti ve %20 test veri seti olacak şekilde rastgele iki parçaya ayrılır. Her defasında farklı farklı rastgele veri bölme işlemi yapılarak elde edilen eğitim ve test veri setleri, makine öğrenmesi algoritmalarının performans kıyası için adil olmayacağından makine öğrenmesi algoritmalarının performans kıyası için her defasında rastgele bölünmüş ancak aynı eğitim ve test veri setlerini kullanmak algoritmaların performans ölçümlerinin adil olması için tercih edilir. Python programlama dilinde random_state parametresine bir pozitif tam sayı değer atanarak rastgele bölünmüş aynı eğitim ve test veri setleri kullanılabilir. Bu tez çalışması boyunca Python programlama dili kullanılmıştır.

Gözetimli/ Denetimli makine öğrenme yaklaşımı, sınıflandırma (classification) ve regresyon (regression) problemlerini çözmek için kullanılır. Çıktılar; eğer yaş, zaman, fiyat, maaş, derece gibi sürekli değerlerden oluşuyorsa bu regresyon problemidir. Eğer çıktılar; evet ya da hayır, doğru ya da yanlış, iyi huylu tümör veya kötü huylu tümör, cinsiyet, tür, cins, grup, sınıf etiketi gibi ayrık veya kategorik değerlerden oluşuyorsa bu sınıflandırma problemidir. Bu tez çalışmasında bir sınıflandırma problemi üzerinde çalışılmıştır.

3.2.1.1 Sınıflandırma Problemleri

Sınıflandırma problemleri bir nesnenin özelliklerine göre hangi sınıfa ait olduğunu belirleme problemidir. Sınıflandırma problemlerinin çözümünde kullanılan veri setinin yapısında öznitelikler ve bu özniteliklere karşılık gelen nesnelerin sınıflarını belirten kategorik hedef değişkenleri olan sınıf etiketleri yer almaktadır. Özniteliklerin benzerlikleri sayesinde nesnelerin sınıfları belirlenebilmektedir. Bu öznitelikler, çap, boy, uzunluk, renk, cinsiyet, konum, şekil, boyut, çevre gibi nesnelerin özelliklerinden oluşmaktadır. Hedef değişkeni ise, bir fotoğrafta yer alan nesnenin ne olduğu, el yazısı ile yazılan bir harf ya da rakamın hangi harf ya da hangi rakam olduğu, bir elektronik postanın spam olup olmadığı, bir hücrenin kanser hücresi olup olmadığı, abonelik sisteminde yer alan bir müşterinin geçmiş davranışlarına göre bir dahaki yıl abone olup olmayacağı, bir bankadan kredi alan bir kişinin krediyi düzenli bir şekilde krediyi aldığı bankaya geri ödeyip ödemeyeceği, yazılı bir ifadede yer alan kelimelerden yazının içerisindeki duygunun hangi duygu türüne ait olduğu, bir pirinç tanesinin hangi türe ait olduğu, bir çiçeğin türünün ne olduğu gibi birçok sorunun cevabı olan kategorik/ süreksiz/ kesikli/ ayrık sınıf etiketlerinden oluşur.

Veri seti kullanılarak makine öğrenmesi sınıflandırma algoritmaları ile model oluşturulup eğitildikten sonra test veri seti ile test edilir ve modelin performansı ölçülür. Model performansı başarılı kabul edilirse sınıflandırma problemleri için bu model kullanılabilir. Sınıflandırma Problemi 4 tip problemden oluşmaktadır (Asif, 2024). İkili sınıflandırma problemleri (Binary Classification): Hedef değişkenin 2 sınıf etiketi içerdiği sınıflandırma problemleridir. Çok sınıflı sınıflandırma problemleri (Multi Class Classification): Hedef değişkenin 2'den fazla sınıf etiketi içerdiği sınıflandırma problemleridir. Çoklu etiket sınıflandırma problemleri (Multi Label Classification): Veri setinde yer alan nesnenin aynı anda birden çok hedef değişken sınıf etiketine sahip olduğu sınıflandırma problemleridir. Dengesiz sınıflandırma problemleri (Imbalanced Classification): Veri setinde yer alan nesne sayısının dengesiz olduğu sınıflandırma problemleridir.

3.2.1.2 Regresyon Problemleri

Regresyon problemleri de sınıflandırma problemleri gibi bir nesnenin özelliklerine göre hangi sınıfa ait olduğunu belirleme problemidir. Sınıflandırma

problemlerinden farkı hedef değişkenin sürekli değerlerden oluşmasıdır. Örneğin; daha önceki verilere dayanarak ev fiyat tahmini, gelir tahmini, hava durumu tahmini, hastalık prognozu (bir hastalığın ilerlemesi veya bir tedavi yönteminin etkinliği gibi tıbbi durumların tahmin edilmesi. Örneğin, kanser tedavisi sırasında hastanın sağ kalma süresini tahmin etmek), hisse senedi fiyat tahmini, trafik yoğunluğu tahmini vb.

3.2.2 Gözetimsiz/ Denetimsiz Öğrenme

Gözetimsiz/ denetimsiz öğrenme, makine öğrenmesinde önemli bir paradigmadır. Bu yöntem, veri kümesindeki desenleri veya yapıları otomatik olarak tespit etmeye çalışır, ancak bu süreçte herhangi bir etiketlenmiş veriye ihtiyaç duymaz. Gözetimsiz öğrenme, veri setlerindeki desenleri keşfetmek ve veri hakkında anlam çıkarmak için güçlü bir araçtır. Bu yaklaşım, özellikle büyük miktarlarda etiketlenmemiş veriye sahip olduğumuzda, bilgi çıkarma ve karar alma süreçlerini hızlandırabilir ve genelleştirme yeteneğini artırabilir.

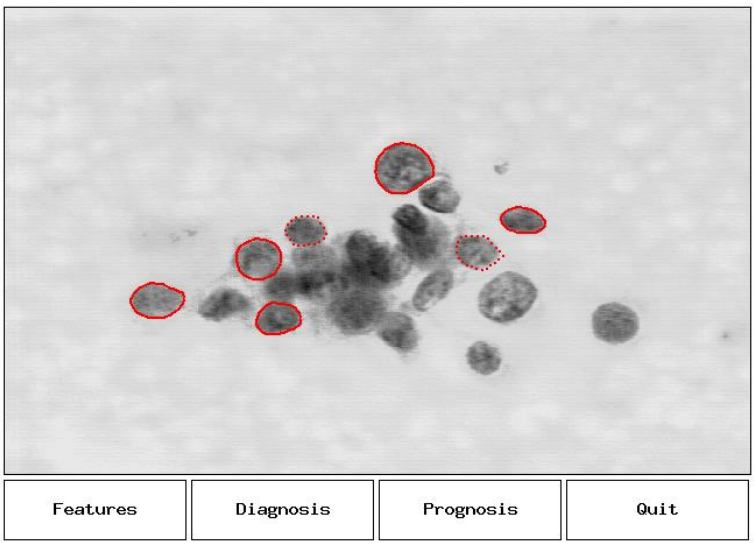
3.2.3 Pekiştirmeli/ Takviyeli Öğrenme

Pekiştirmeli / takviyeli öğrenme, bir ajanın (algoritma veya yapay zekâ sistemi) belirli bir çevrede hedefe ulaşmak için eylemler gerçekleştirerek öğrenmesini sağlar. Bu öğrenme yönteminde, ajan çevresini algılar, belirli bir durumda bir eylem seçer, bu eylemin sonucunda bir ödül alır ve bu ödül ile ilgili geri bildirim alır. Amaç, zamanla daha fazla ödül toplamak için en iyi eylem stratejisini öğrenmektir. Pekiştirmeli öğrenme, gözetimsiz/ denetimsiz öğrenme ve gözetimli/denetimli öğrenme arasında bir yerde bulunur. Gözetimli/ denetimli öğrenmede olduğu gibi etiketlenmiş veriye dayanmaz ve gözetimsiz/ denetimsiz öğrenme gibi desenleri belirli bir şekilde keşfetmez. Bunun yerine, ajanın deneyimlerinden öğrenir ve çevresine etkisi sonucu aldığı geri bildirimi kullanarak davranışını şekillendirir.

4
MATERYAL VE METOT

4.1 Veri Seti Hikayesi

Bu tez çalışmasında Wisconsin üniversitesi hastanesinden Dr. William H. Wolberg ve çalışma arkadaşları tarafından oluşturulan ve Kaliforniya Üniversitesi, Irvine (UCI) makine öğrenmesi veri deposunda bulunan Wisconsin Üniversitesi Tanısal Göğüs Kanseri Veri Seti (WDBC) kullanılmıştır (Wolberg, Street, & Mangasarian, 1995). Göğüs kanseri tanısı ve devamında veri setini oluşturmak için, incelenmek istenen bölgelerdeki cilt üzerinden ince iğne aspirasyon biyopsisi işlemi (FNA) ile hücre örnekleri alınmıştır. FNA, mikroskop altında hücre çekirdeklerinin incelenmesi için boyanmıştır. Hücrelerin iyice farklılaştıkları anda lam ve lamel dijital bir kamera veya bir kare yakalama kartı (capture card) yardımıyla taranmıştır. Daha sonra Xcyt programı kullanılarak hücrelerin sınırları fare yardımıyla yaklaşık olarak çizilmiştir ve sonrasında "snakes (yılanlar)" olarak bilinen bir bilgisayar görüntüleme yaklaşımı kullanılarak tam çekirdek sınırları çizilmiştir (Şekil 4.1).



Şekil 4.1 Xcyt kullanımı (1990)

Her bir hücre (veya çoğu) Şekil 4.1'deki gibi ayrı ayrı işaretlendikten sonra, program her bir tümör hücresi çekirdeği için 10 ana öznitelik hesaplamıştır

(Wolberg, Mangasarian, & Street, 1990). Her bir tümör hücresi çekirdeği kullanılarak elde edilen 10 temel öznitelik Tablo 4.1'de verilmiştir.

Tablo 4.1 Temel 10 öznitelik

	Öznitelikler
1	Yarıçap1
2	Doku1
3	Çevre1
4	Alan1
5	Pürüzsüzlük1
6	Yoğunluk1
7	İçbükeylik1
8	İçbükey nokta sayısı1
9	Simetri1
10	Fraktal boyut1

Yarıçap1 (radius1): Çekirdek merkezinden sınır noktalarına olan ortalama uzaklığı gösterir.

Doku1 (texture1): Görüntünün her pikselindeki gri tonların yoğunluğunu ifade eden gri tonlamalı değerlerin standart sapmasını gösterir.

Çevre1 (perimeter1): Hücre çekirdeğinin sınırının toplam uzunluğunu gösterir.

Alan1 (area1): Sınırın iç kısmındaki piksel sayısı ile dijitalleştirmenin neden olduğu hata için sınırın çevresindeki piksellerin yarısını ekleyerek hesaplanan piksel sayısını gösterir. Bu yöntem, sınırı oluşturan piksellerin yanlış hesaplanmasından kaynaklanan hata miktarını telafi etmeyi amaçlar.

Pürüzsüzlük1 (smoothness1): Bir yarıçap uzunluğu ile aynı merkezli ve birbirlerine yakın olan iki yarıçap hattının ortalama uzunluğu arasındaki farkı gösterir. Yani yarıçap uzunluklarının yerel olarak nasıl değiştiğini gösterir.

Yoğunluk1 (compactness1): Hücre çekirdeklerinin sıkılığını ölçmek için çevre ve alan değerleri kullanılarak elde edilen bir değeri gösterir.

İçbükeylik1 (concavity1): Konturun (çekirdeğin dış çizgisi) içbükey kısımlarının şiddetini gösterir. Hücre çekirdeğinin sınırının içbükey bölgelerinin ne kadar belirgin olduğunu belirtir. Yüksek bir içbükeylik, sınırın oyuklarla dolu olduğunu ve bu nedenle daha pürüzlü bir yapıya sahip olduğunu gösterir.

İçbükey nokta sayısı1 (concave points1): Hücre çekirdeğinin konturundaki içbükey kısımların sayısını gösterir.

Simetri1 (symmetry1): Bir çekirdek için simetriyi belirlemek için, öncelikle çekirdeğin merkezinden başlayarak sınır noktalarına ulaşan en uzun çizgi bulunur. Daha sonra, bu en uzun çizgiye dik olan hatlar çizilir ve bu hatlar boyunca sınırın uzunluk farkları ölçülür. Başka bir deyişle, en uzun çizgi boyunca, sınıra göre dikey olarak çizilen hatlar arasındaki uzunluk farkları hesaplanır. Bu uzunluk farkları, çekirdeğin simetrisi hakkında bilgi sağlar.

Fraktal boyut1 (fractal dimension1): Çevrenin Ölçümü: Öncelikle, hücre çekirdeğinin çevresi farklı uzunluklardaki ölçüm çubuklarıyla ölçülür. Bu ölçüm çubukları, çekirdeğin çevresini belirli bir hassasiyetle ölçmeye yöneliktir. Bu çubuklar, sınırlı bir hassasiyet nedeniyle, gerçek çevre uzunluğunu tam olarak temsil etmeyebilir.

L(s) Hesaplanması: Ölçüm çubuklarının uzunluğuna bağlı olarak, ölçülen çevrenin toplam uzunluğu L(s) belirlenir. Burada, "s" ölçüm çubuğunun uzunluğunu temsil eder.

Logaritmik İşlemler: Sonrasında, L(s) değeri üzerinde logaritmik işlemler uygulanır. Bu işlem, ölçülen çevrenin uzunluğunun ölçüm çubuğu uzunluğuna bağlı olarak nasıl değiştiğini anlamamıza yardımcı olur. Logaritmik işlem, özellikle fraktal yapıları analiz etmek için sıkça kullanılır çünkü fraktal yapılar logaritmik özelliklere sahiptir.

Fraktal Boyutun Belirlenmesi: Son olarak, log(L(s)) değerini log(s) değerine karşı çizerek bir grafik elde edilir. Bu grafikte, logaritmik ölçekler üzerinde bir ilişki görülebilir. Bu ilişkiyi inceleyerek, fraktal boyutun yaklaşık bir tahmini belirlenir (Bourke, 2003).

Bu tahmin, genellikle grafikteki eğimin negatif değeri olarak elde edilir ve bu, fraktal boyutun yaklaşık bir tahminini verir. Bu adımlar, bir nesnenin fraktal boyutunu belirlemede yaygın olarak kullanılır ve karmaşık yapıları anlamak için önemlidir (Westerdijk & Bhulai, 2018).

Geriye kalan 20 öznitelik ise Tablo 4.1'de verilen 10 temel özniteliğin ortalaması, standart hatası, en küçük ve en büyük değeri gibi sayısal değerlerden elde edilmiştir. Sonuç olarak veri seti 568 örnek ihtiva etmekte olup veri setindeki her bir örnek 30 adet öznitelik ile kavuşturulmuş şekilde verilmektedir. Tüm bu özniteliklerin yanı sıra her bir örnek için bir kullanıcı numarası (ID) tanımlanmış ve her bir örneğin özniteliklerine bağlı tanılarından oluşan bir hedef değişken; tanı (Diagnosis) belirtilmiştir. Hedef değişkenler, B ve M tanı sınıflarından oluşmaktadır. Veri seti 357 adet B ve 211 adet M hedef değişkeni ihtiva etmektedir.

4.2 Veri Manipülasyonu

Bu tez boyunca Anaconda Navigator arayüzündeki Jupyter Lab geliştirme ortamı kullanıldı. WDBC Pandas kütüphanesi yardımıyla okunarak Jupyter Lab geliştirme ortamına aktarıldı. UCI makine öğrenmesi veri deposunda bulunan WDBC özniteliklerin isimlerini içeren satıra sahip olmadığından okumada kolaylık sağlaması amacı ile özniteliklerin isimlerini içeren satır "columns" fonksiyonu kullanılarak veri setine eklendi.

Makine öğrenmesi sınıflandırma algoritmaları için ID verilerine ihtiyaç olmadığından ID sütunu, "drop" fonksiyonu kullanılarak veri setinden çıkartıldı. Daha sonra veri setinin yapısal bilgileri "info" fonksiyonu kullanılarak ekrana yazdırıldı (Şekil 4.2).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 568 entries, 0 to 567
Data columns (total 31 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Diagnosis          568 non-null    object
 1   radius1            568 non-null    float64
 2   texture1           568 non-null    float64
 3   perimeter1         568 non-null    float64
 4   area1              568 non-null    float64
 5   smoothness1        568 non-null    float64
 6   compactness1       568 non-null    float64
 7   concavity1         568 non-null    float64
 8   concave_points1    568 non-null    float64
 9   symmetry1          568 non-null    float64
 10  fractal_dimension1  568 non-null    float64
 11  radius2            568 non-null    float64
 12  texture2           568 non-null    float64
 13  perimeter2         568 non-null    float64
 14  area2              568 non-null    float64
 15  smoothness2        568 non-null    float64
 16  compactness2       568 non-null    float64
 17  concavity2         568 non-null    float64
 18  concave_points2    568 non-null    float64
 19  symmetry2          568 non-null    float64
 20  fractal_dimension2  568 non-null    float64
 21  radius3            568 non-null    float64
 22  texture3           568 non-null    float64
 23  perimeter3         568 non-null    float64
 24  area3              568 non-null    float64
 25  smoothness3        568 non-null    float64
 26  compactness3       568 non-null    float64
 27  concavity3         568 non-null    float64
 28  concave_points3    568 non-null    float64
 29  symmetry3          568 non-null    float64
 30  fractal_dimension3  568 non-null    float64
dtypes: float64(30), object(1)
memory usage: 137.7+ KB
```

Şekil 4.2 Veri setinin yapısal bilgileri

Tanı (Diagnosis) hedef değişkeni "nesne (object)" tipinde olduğundan bazı fonksiyonlar tanı özniteliğine "metin (string)" muamelesi yapabilir ve bu sorunlara yol açabilir. Bundan dolayı tanı özniteliğinin tipi "Categorical" fonksiyonu kullanılarak "kategorik" olarak değiştirildi.

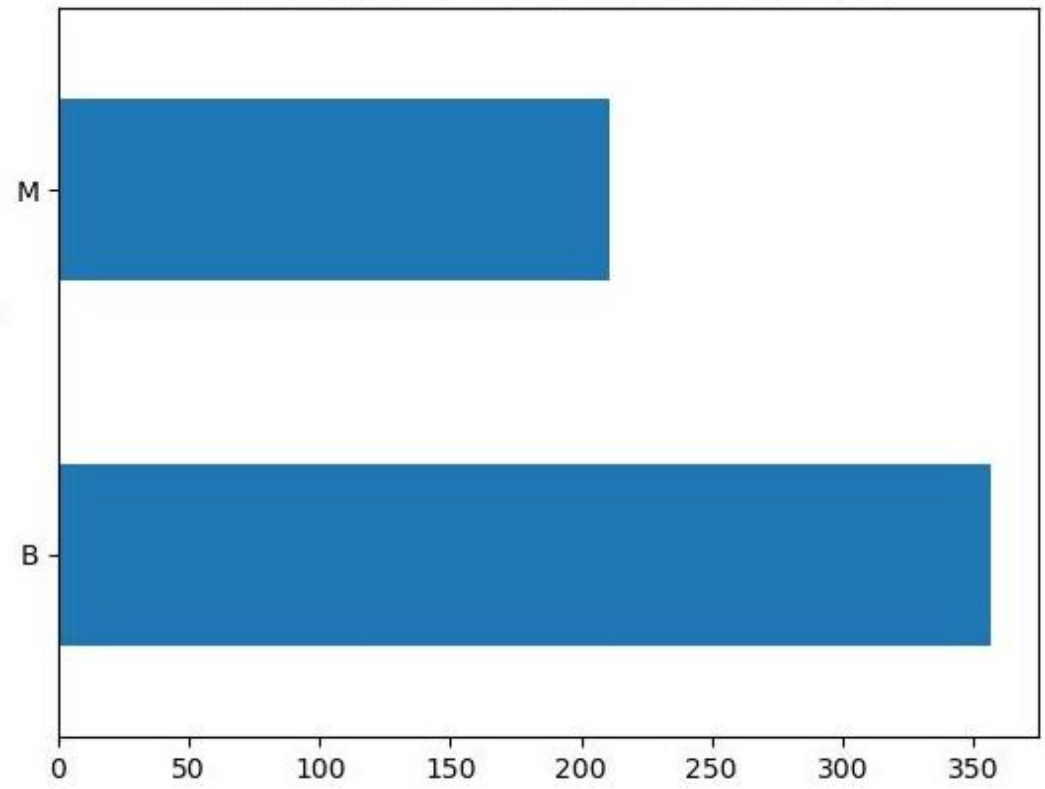
0 Diagnosis 568 non-null category

Bu işlemler sonucu "shape" fonksiyonu kullanılarak veri setinin boyutunun (568,31) olduğu gözlemlendi.

Veri setinden eksik verinin varlığının araştırılması için "isnull" fonksiyonu kullanıldı. Veri setinde eksik bir verinin olmadığı görüldü.

Veri setinin ikili sınıflandırma (Binary Classification) yapısında olduğunu ve sınıf etiket sayılarını görmek için "value_counts" fonksiyonu kullanıldı. Tanı sınıfında 357 adet B ve 211 adet M olmak üzere tanı hedef değişken sütununda toplam 568 sınıf etiketi yer aldığı görüldü.

Sonra "plot" fonksiyonu kullanılarak tanı hedef değişkeninin sınıf frekansları yazdırıldı (Şekil 4.3).



Şekil 4.3 Tanı hedef değişkeninin sınıf frekansları

Sonra "describe" fonksiyonu kullanılarak veri setinin bazı betimsel istatistiksel değerleri gözlemlendi (Şekil 4.4).

	mean	std	min	25%	50%	75%	max
radius1	14,12049	3,523416	6,981	11,6975	13,355	15,78	28,11
texture1	19,30533	4,288506	9,71	16,1775	18,855	21,8025	39,28
perimeter1	91,91475	24,28585	43,79	75,135	86,21	103,875	188,5
area1	654,2798	351,9238	143,5	420,175	548,75	782,625	2501
smoothness1	0,096321	0,014046	0,05263	0,08629	0,095865	0,1053	0,1634
compactness1	0,104036	0,052355	0,01938	0,064815	0,092525	0,1304	0,3454
concavity1	0,088427	0,079294	0	0,02954	0,0614	0,12965	0,4268
concave_points1	0,048746	0,038617	0	0,02031	0,033455	0,07373	0,2012
symmetry1	0,181055	0,027319	0,106	0,1619	0,1792	0,195625	0,304
fractal_dimension1	0,06277	0,007035	0,04996	0,057698	0,061515	0,06612	0,09744
radius2	0,403958	0,276038	0,1115	0,232375	0,32395	0,477325	2,873
texture2	1,217402	0,551979	0,3602	0,83315	1,1095	1,47425	4,885
perimeter2	2,855984	2,009288	0,757	1,605	2,2855	3,33675	21,98
area2	40,13802	45,28241	6,802	17,85	24,485	45,0175	542,2
smoothness2	0,007042	0,003005	0,001713	0,005166	0,006375	0,008151	0,03113
compactness2	0,025437	0,017897	0,002252	0,013048	0,020435	0,032218	0,1354
concavity2	0,031855	0,030199	0	0,015063	0,025875	0,041765	0,396
concave_points2	0,011789	0,006173	0	0,007635	0,01092	0,01471	0,05279
symmetry2	0,020526	0,008264	0,007882	0,015128	0,018725	0,023398	0,07895
fractal_dimension2	0,003791	0,002646	0,000895	0,002245	0,003162	0,004526	0,02984
radius3	16,25315	4,82232	7,93	13,01	14,965	18,7675	36,04
texture3	25,69192	6,141662	12,02	21,095	25,425	29,7575	49,54
perimeter3	107,1251	33,47469	50,41	84,1025	97,655	125,175	251,2
area3	878,5789	567,8463	185,2	514,975	685,55	1073,5	4254
smoothness3	0,132316	0,022818	0,07117	0,1166	0,1313	0,146	0,2226
compactness3	0,253541	0,156523	0,02729	0,1469	0,21185	0,3376	1,058
concavity3	0,271414	0,207989	0	0,114475	0,22655	0,3814	1,252
concave_points3	0,114341	0,065484	0	0,06473	0,09984	0,161325	0,291
symmetry3	0,289776	0,061508	0,1565	0,25035	0,28205	0,317675	0,6638
fractal_dimension3	0,083884	0,018017	0,05504	0,071413	0,080015	0,092065	0,2075

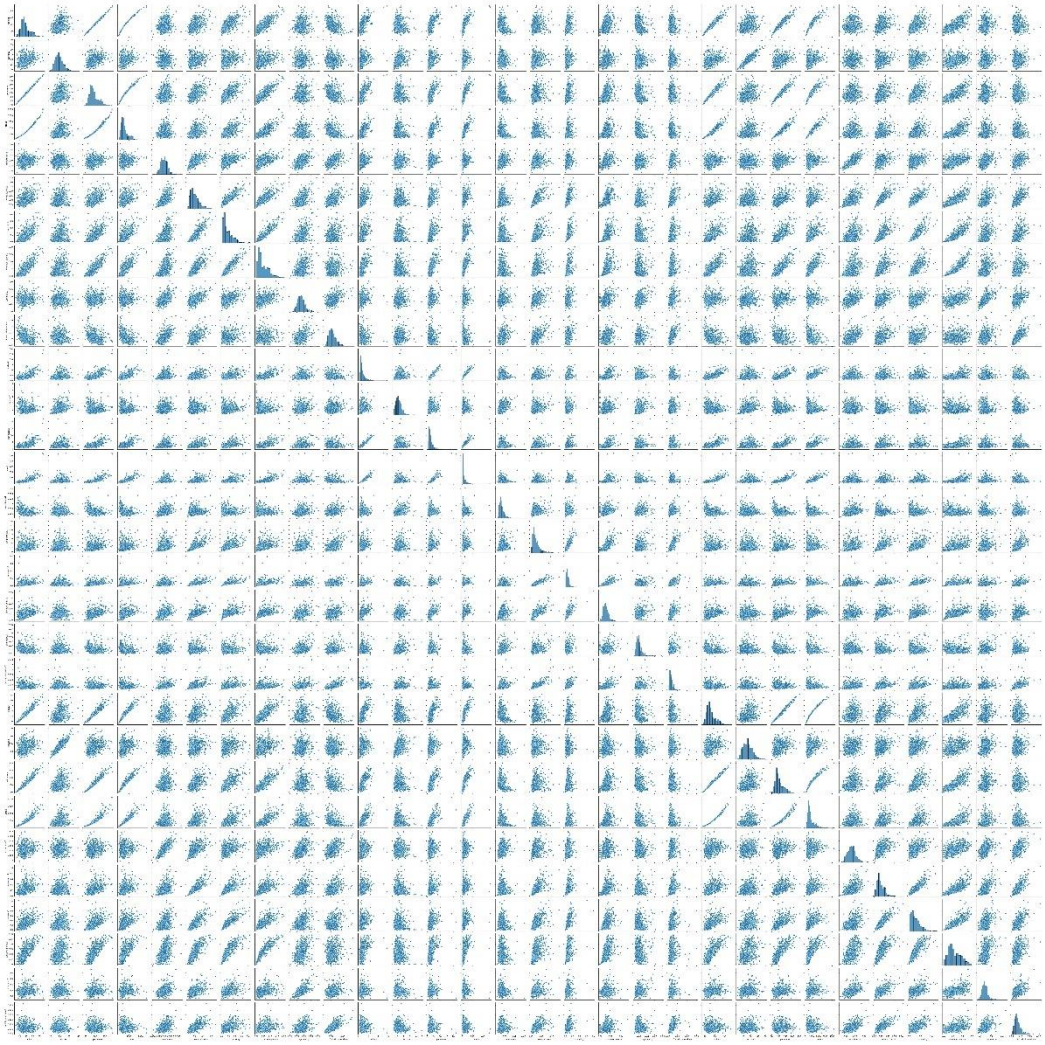
Şekil 4.4 Veri setinin betimsel istatistikleri

Şekil 4.4'teki "mean" ifadesi ortalama değeri, "std" ifadesi standart sapmayı, "min" ifadesi en küçük değeri, %25, %50, %75 ifadeleri verilen %25, %50 ve %75'inci yüzdelik dilimleri (çeyreklikler), "max" ifadesi ise en büyük değeri ifade etmektedir.

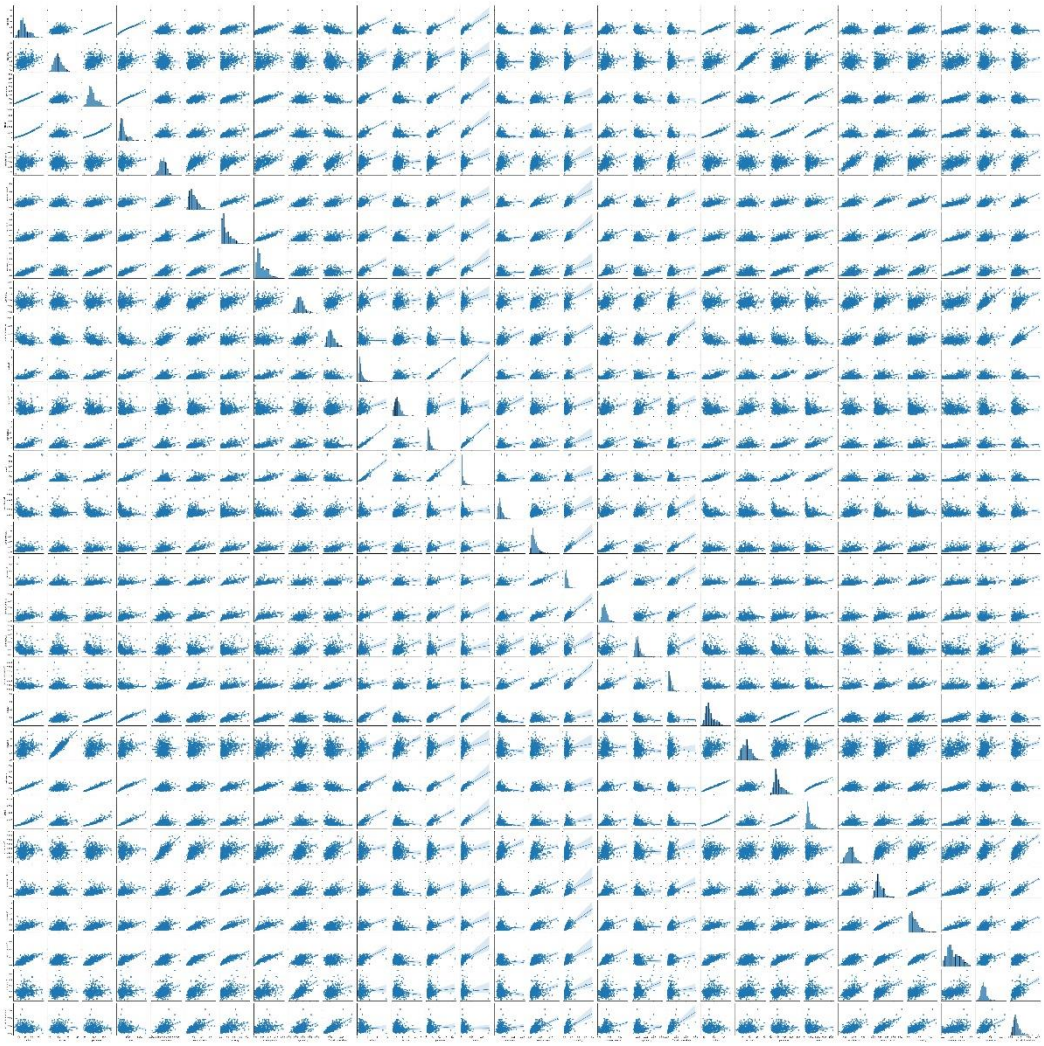
Daha sonra "pairplot" fonksiyonu kullanılarak özniteliklerin birbirleri ile olan ikili ilişkileri grafikler halinde yazdırıldı (Şekil 4.5). Öznitelikler arasındaki ikili ilişkileri daha da iyi anlamak adına kind = "reg" parametresi kullanılarak grafiklere regresyon çizgileri eklendi. Regresyon çizgileri, öznitelikler arasındaki doğrusal ilişkiyi gösterir ve bu ilişkinin ne kadar güçlü olduğunu anlamak için kullanılabilir (Şekil 4.6). Bir sonraki aşamada ise "pairplot" fonksiyonunun "hue" parametresi kullanılarak tanı hedef değişken tiplerine farklı renkler atanarak ve ayrıca kind =

“reg” parametresi ile grafiklere regresyon çizgileri de eklenerek grafikler tekrar yazdırıldı (Şekil 4.7, Şekil 4.8).

Seaborn kütüphanesinde bulunan bir fonksiyon olan “pairplot”, veri setindeki tüm sayısal değişkenler arasındaki ilişkiyi görselleştirmek için kullanılır. Bu fonksiyon, veri setindeki her bir sayısal değişkenin diğer sayısal değişkenlerle nasıl ilişkili olduğunu gösteren dağılım grafiklerini çizer. Eğer veri setinde kategorik değişkenler bulunuyorsa, bu değişkenlere göre alt gruplara ayrılmış dağılım grafiklerini çizme yeteneğine de sahiptir.



Şekil 4.5 Öznitelikler arası ikili ilişkiler 1



Şekil 4.6 Öznitelikler arası ikili ilişkiler 2



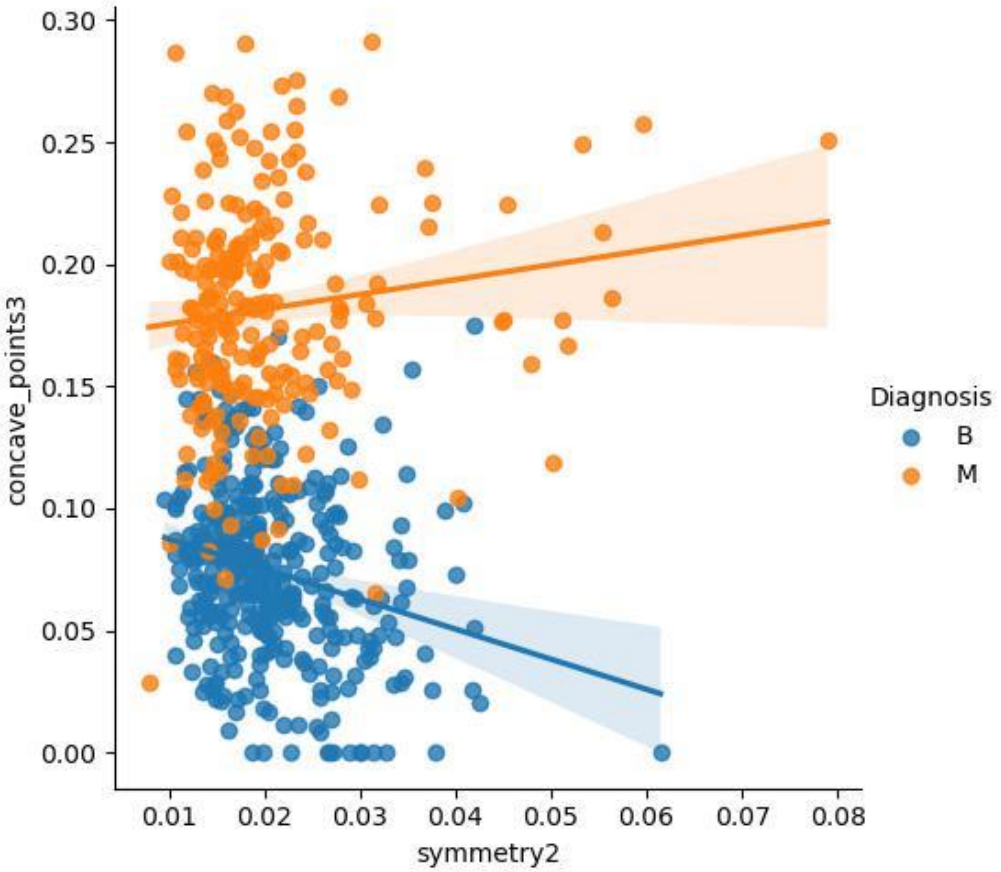
Şekil 4.7 Öznitelikler arası ikili ilişkiler 3



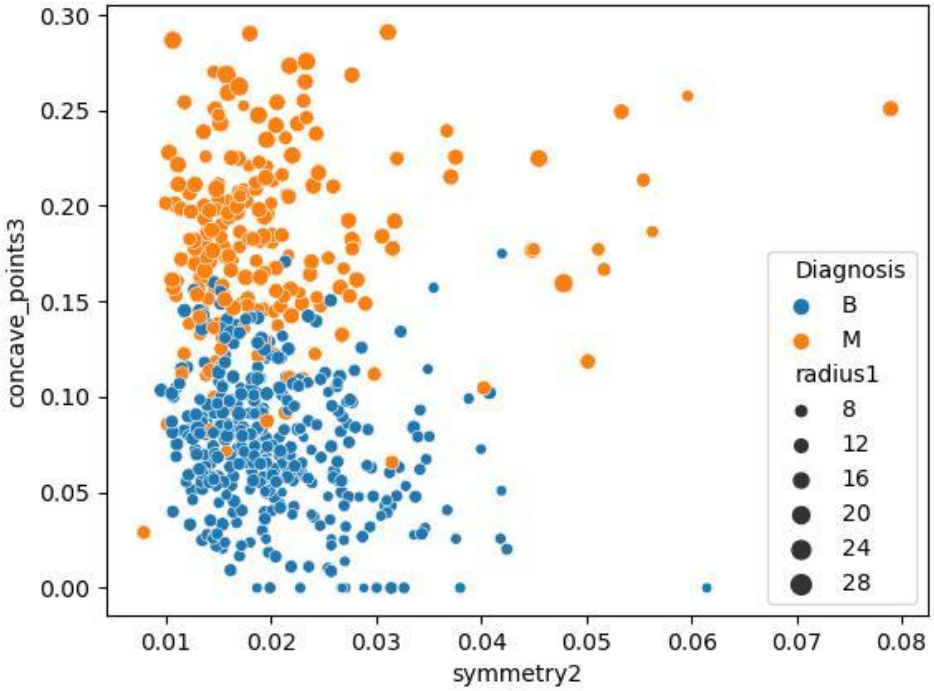
Şekil 4.8 Öznitelikler arası ikili ilişkiler 4

İstenildiği taktirde bu ikili ilişkiler, ikişer ikişer, hatta daha fazla öznitelik de kıyaslamaya dahil edilerek daha fazla boyutlu ve daha detaylı grafikler halinde yazdırılabilir. Örnek olarak, symmetry2 ve concave_points3 öznitelikleri arasındaki ilişkiyi gösteren grafik regresyon çizgileri de eklenerek çizdirildi (Şekil 4.9). Benzer şekilde symmetry2, concave_points3 ve radius1 özniteliklerin arasındaki ilişkiyi anlamak adına üç öznitelik arasındaki ilişkiyi gösteren grafik çizildi (Şekil 4.10). Bunun gibi diğer tüm öznitelikler arasındaki ikili ve üçlü ilişkiler daha detaylı grafikler halinde çizdirilerek gözlemlenebilir. Böylece öznitelikler arasındaki ilişkiler çok daha iyi bir şekilde anlaşılabilir. Tüm bunların yanı sıra veri setindeki aykırı değerlerin varlığı, merkezi eğilimleri, yayılımları ve dağılımları gibi kavramların hızlıca anlaşılması amacı ile Seaborn kütüphanesinde bulunan "displot" ve "distplot" fonksiyonları ve bu fonksiyonların parametreleri kullanıldı. Bu fonksiyonlar ve parametreleri kullanılarak makine öğrenmesi

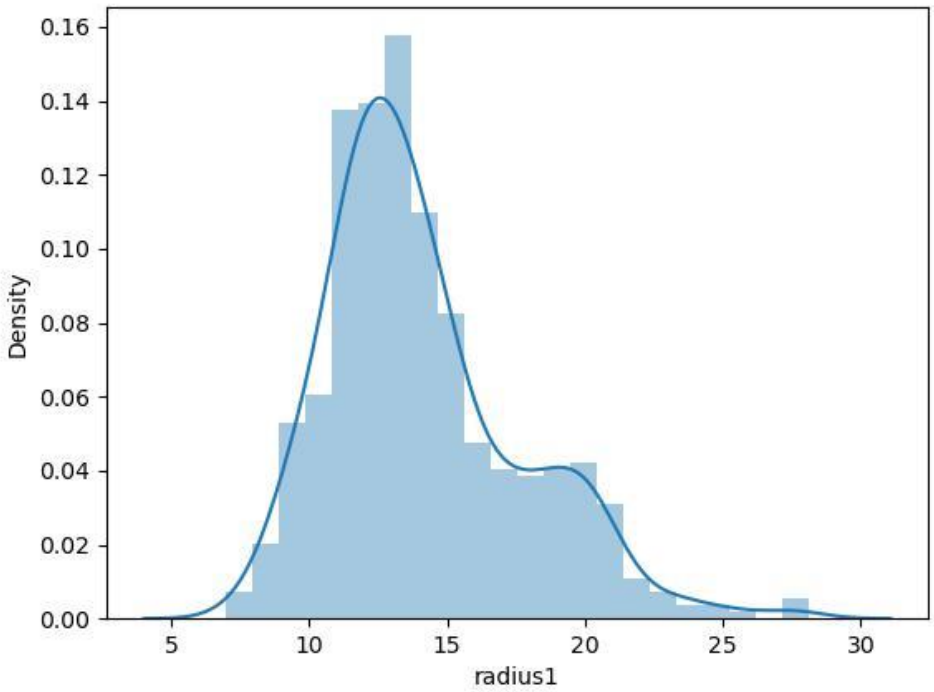
algoritmaları kullanılmadan önce veri seti hakkında ilk gözlemler yapılabilir. Örneğin Şekil 4.11'de verilen grafik radius1 özniteliğinin dağılımını göstermektedir. Bunun gibi daha pek çok grafik çizdirmek ve öznitelikler hakkında daha detaylı bilgiler edinmek için Seaborn kütüphanesindeki fonksiyonlar ve bu fonksiyonların parametreleri oldukça faydalı olmaktadır.



Şekil 4.9 symmetry2 ve concave_points3 arasındaki ilişki



Şekil 4.10 symmetry2, concave_points3, radius1 arasındaki ilişki



Şekil 4.11 radius1 dağılımı

4.3 Göğüs Kanseri Tanısında Makine Öğrenmesi Kullanımı

Veri manipülasyonu adımından sonra gözetimli/ denetimli makine öğrenmesi sınıflandırma algoritmalarının göğüs kanseri tanı başarılarının karşılaştırılması ve model eğitimi amacı ile önce, öznitelikler ile tanı hedef değişkenleri birbirlerinden ayrıştırılmıştır. Daha sonra özniteliklerin standardizasyonu gerçekleştirilmiştir. Bunun yanı sıra tanı hedef değişkenlerinin tipleri, M=1, B=0 olacak şekilde metin (string) tipinden tam sayı (integer) tipine dönüştürülmüştür. Bu işlemlerden hemen sonra WDBC, %80 eğitim veri seti ve %20 test veri seti olacak şekilde rastgele iki parçaya ayrılmıştır. Her defasında farklı farklı rastgele veri bölme işlemi yapılarak elde edilen eğitim ve test veri setleri, gözetimli/ denetimli makine öğrenmesi sınıflandırma algoritmalarının performans kıyası için adil olmayacağından her defasında rastgele bölünmüş ancak aynı eğitim ve test veri setleri kullanılarak gözetimli/ denetimli makine öğrenmesi sınıflandırma algoritmalarının performans ölçümlerinin adil olması sağlanmıştır. Python programlama dilinde "random_state" parametresine her defasında aynı pozitif tam sayı değeri atanarak rastgele bölünmüş aynı eğitim ve test veri setleri kullanılabilir. Aynı eğitim ve test veri setlerinin kullanıldığının kontrolü amacı ile her adımda örnek olarak eğitim veri setinin bağımsız değişkenleri olan özniteliklerin 50. satırı, 25. sütunu, bağımlı değişken olan tanı hedef değişkenlerin 453. sü, test veri setlerinin bağımsız değişkenleri olan özniteliklerin 103. satırı, 14. sütunu ve bağımlı değişken olan tanı hedef değişkenlerin 112. si ekrana yazdırılarak aynı değere sahip oldukları kontrol edilmiştir (Şekil 4.12). Bu değerler tamamen rastgele, keyfi olarak seçilmiştir.

```
print([x_train[50,25],y_train[453],x_test[103,14],y_test[112]])
[-0.8526431448570426, 1, 0.3053862884401457, 0]
```

Şekil 4.12 random_state parametresi kontrolü

Rastgele ayrılmış bu eğitim seti kullanılarak Python programlama dilinde açık kaynaklı bir makine öğrenmesi kütüphanesi olan Scikit Learn kütüphanesinin içerisinde (Supervised learning, 2024) yer alan gözetimli/ denetimli makine öğrenmesi sınıflandırma algoritmalarından SVM-SVC, DT, LR, RF, KNN-KNC, NN-MLP (Sakib, ve diğerleri, 2022), LDA (Oyewola, Hakimi, Adeboye, & Shehu, 2016), NB-BNB (Chhillar & Singh, 2024), GB, HGB, AB (Chhillar & Singh, 2024), Ridge, SGD, Perceptron, PA, NC ve GPC (Supervised learning, 2024) olmak

üzere toplam 17 adet gözetimli/ denetimli makine öğrenmesi sınıflandırma algoritmaları ile modeller eğitilmiştir. Buna ek olarak M=1, B=0 olacak şekilde metin (string) tipinden tam sayı (integer) tipine dönüştürülme adımından hemen sonra 10 katlı çapraz doğrulama ile de modeller eğitilmiştir.

4.4 Karışıklık Matrisi (Confusion Matrix)

Karışıklık matrisi, sınıflandırma problemlerinde modelin performansını detaylı bir şekilde değerlendirmek için kullanılan etkili bir analiz aracıdır. Bu matris, modelin tahmin ettiği sınıflar ile gerçek sınıflar arasındaki ilişkiyi sistematik bir şekilde gösterir. Karışıklık matrisi, dört ana bileşeni ile (doğru pozitifler, yanlış pozitifler, doğru negatifler ve yanlış negatifler) modelin ne kadar doğru çalıştığını, hangi tür hataların yapıldığını ve bu hataların modelin genel performansını nasıl etkilediğini gösterir. Böylece modelin belirli sınıflarda ne kadar başarılı veya başarısız olduğu belirlenebilir ve bu doğrultuda iyileştirmeler yapılabilir. Karışıklık matrisi, özellikle dengesiz veri setlerinde ve farklı hata türlerinin kritik olduğu uygulamalarda, model performansının kapsamlı bir şekilde anlaşılmasına yardımcı olmaktadır.

Doğru pozitif, modelin gerçekte pozitif olan örnekleri doğru bir şekilde pozitif olarak sınıflandırdığı durumu temsil eder. Örneğin, kanser tarama modeli için doğru pozitif, hastanın gerçekte kansere sahip olduğu durumda modelin de o hastanın kanserli olduğunu tespit ettiği durumdur.

Doğru negatif, modelin gerçekte negatif olan örnekleri doğru bir şekilde negatif olarak sınıflandırdığı durumu temsil eder. Örneğin, kanser tarama modeli için doğru negatif, hastanın gerçekte kansere sahip olmadığı durumda modelin de o hastanın kanserli olmadığını tespit ettiği durumdur.

Yanlış negatif, modelin gerçekte pozitif olan örnekleri yanlışlıkla negatif olarak sınıflandırdığı durumu temsil eder. Örneğin, kanser tarama modeli için yanlış negatif, hastanın gerçekte kansere sahip olduğu halde modelin yanlış bir şekilde o hastanın kanserli olmadığını belirten durumdur.

Yanlış pozitif ise modelin gerçekte negatif olan örnekleri yanlışlıkla pozitif olarak sınıflandırdığı durumu temsil eder. Örneğin, kanser tarama modeli için yanlış

pozitif, hastanın gerçekte kansere sahip olmadığı halde modelin yanlış bir şekilde o hastanın kanserli olduğunu belirten durumdur.

Karışıklık matrisi, bu dört durumu detaylı bir şekilde görselleştirerek modelin performanslarını değerlendirmeye olanak sağlamaktadır.

Bu tezde kullanılan her bir gözetimli/ denetimli makine öğrenmesi sınıflandırma algoritması için doğru pozitif (true pozitive), doğru negatif (true negative), yanlış pozitif (false positive) ve yanlış negatif (false negative) tanı sonuçları kullanılarak karışıklık matrisleri oluşturulmuştur (Tablo 4.2).

Karışıklık matrisleri, her bir model için ayrı ayrı oluşturulduktan sonra bu tezde kullanılan her bir gözetimli/ denetimli makine öğrenmesi sınıflandırma algoritması için bu karışıklık matrisleri yardımı ile Tablo 4.3'te verilen metrikler hesaplamıştır.

Tablo 4.2 Karışıklık matrisi genel görünümü

Karışıklık matrisi		Tahmin edilen durum	
		Pozitif	Negatif
Gerçek durum	Pozitif	TP	FN
	Negatif	FP	TN

4.5 Performans Metrikleri

Karışıklık matrisleri, modellerin hangi sınıfları ne kadar doğru bir şekilde tahmin ettiğini gösterirken, metrikler (Tablo 4.3) bu tahminlerin kalitesini nicel olarak değerlendirmektedir. Bu metrikler (Tablo 4.3), modelin performansını tarafsız bir şekilde değerlendirmemizi sağlar. Örneğin, modelin genel doğruluğu yüksek olsa bile, belirli bir sınıfın kötü tahmin edilmesi, gerçek dünya uygulamalarında ciddi sonuçlara yol açabilir. Bu metrikler, böyle dengesizlikleri ve hataların farklı sonuçlarını göz önünde bulundurarak, modelin gerçek performansını daha doğru bir şekilde değerlendirir. Böylece modelin güvenilirliği ve uygulanabilirliği arttırılabilir.

AC (4.1), sınıflandırma modellerinin doğru tahmin ettiği örneklerin toplam örneklere oranını ifade eden bir performans metriğidir. Bu metrik, modelin ne kadarının doğru sonuçlar ürettiğini değerlendirirken, doğru tahminlerin toplam örneklere oranını belirleyerek modelin genel etkinliğini ölçer. AC, genel olarak model performanslarını ölçmek için kullanışlıdır fakat bazı sınıflandırma problemlerinde tek başına yetersiz kalabilmektedir. Bu nedenle AC skorunun Tablo 4.3'te verilen SE/ RE, SP, PR veya F metrikleri ile değerlendirilmesi tavsiye edilmektedir.

$$\mathrm{AC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4.1}$$

SE/RE (4.2), bir modelin gerçek pozitifleri doğru bir şekilde tahmin etme oranını ifade eden bir performans metriğidir. Bu metrik, modelin pozitif sınıfı ne kadar doğru tanımlayabildiğini değerlendirirken, doğru pozitif tahminlerin gerçek pozitif örneklere oranını belirleyerek modelin etkinliğini ölçer. SE/RE skoru da bazı sınıflandırma problemlerinde tek başına yetersiz kalabilmektedir. Bu nedenle benzer şekilde SE skorunun da Tablo 4.3'te verilen AC, SP, PR veya F metrikleri ile değerlendirilmesi tavsiye edilmektedir.

$$\mathrm{SE/RE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4.2}$$

SP (4.3), bir modelin gerçek negatifleri doğru bir şekilde tahmin etme oranını ifade eden bir performans metriğidir. Bu metrik, modelin negatif sınıfı ne kadar doğru tanımlayabildiğini değerlendirirken, doğru negatif tahminlerin gerçek negatif örneklere oranını belirleyerek modelin etkinliğini ölçer. SP skoru, özellikle negatif sınıfın doğru tanımlanmasının kritik olduğu alanlarda model performansını ölçmek için kullanışlıdır, fakat bazı sınıflandırma problemlerinde tek başına yetersiz kalabilmektedir. Bu nedenle benzer şekilde SP skorunun da Tablo 4.3'te verilen AC, SE/RE, PR veya F metrikleri ile değerlendirilmesi tavsiye edilmektedir.

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4.3}$$

PR (4.4), bir modelin pozitif tahminlerinin ne kadarının gerçekten doğru olduğunu ifade eden bir performans metriğidir. Bu metrik, modelin pozitif tahminlerde ne kadar isabetli olduğunu değerlendirirken, doğru pozitif tahminlerin toplam pozitif tahminlere oranını belirleyerek modelin etkinliğini ölçer. PR skoru, özellikle yanlış pozitiflerin maliyetinin yüksek olduğu alanlarda model performansını ölçmek için kullanışlıdır, fakat bazı sınıflandırma problemlerinde tek başına yetersiz kalabilmektedir. Bu nedenle benzer şekilde PR skorunun da Tablo 4.3'te verilen AC, SE/RE, SP veya F metrikleri ile değerlendirilmesi tavsiye edilmektedir.

$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4.4}$$

F (4.5), bir modelin PR ve SE/RE metriklerinin harmonik ortalamasını ifade eden bir performans metriğidir. Bu metrik, modelin pozitif sınıfları ne kadar iyi tespit edip doğru bir şekilde sınıflandırdığını değerlendirirken, hem doğru pozitif tahminlerin gerçek pozitif örneklere oranını hem de modelin toplam doğru tahminler içerisindeki doğruluğunu dikkate alarak modelin genel performansını ölçer. F skoru, özellikle hem pozitif örneklerin kaçırılmasının hem de yanlış pozitiflerin maliyetli olduğu alanlarda model performansını ölçmek için kullanışlıdır, fakat bazı sınıflandırma problemlerinde tek başına yetersiz kalabilmektedir. Bu nedenle benzer şekilde F skorunun da Tablo 4.3'te verilen AC, SE/RE, SP veya PR metrikleri ile değerlendirilmesi tavsiye edilmektedir.

$$\mathrm{F} = \frac{2}{\frac{1}{\mathrm{PR}} + \frac{1}{\mathrm{RE}}} = \frac{2 \times \mathrm{PR} \times \mathrm{RE}}{\mathrm{PR} + \mathrm{RE}}, (\mathrm{RE} = \mathrm{SE}) \tag{4.5}$$

NPV (4.6), bir modelin negatif tahminlerinin ne kadar doğru olduğunu gösteren bir performans metriğidir. Bu metrik, modelin negatif olarak sınıflandırdığı örneklerin gerçekten negatif olma olasılığını değerlendirir ve doğru negatif tahminlerin toplam negatif tahminlere oranını belirleyerek modelin etkinliğini ölçer. NPV, özellikle yanlış negatiflerin maliyetli olduğu alanlarda model performansını ölçmek için kullanışlıdır, fakat bazı sınıflandırma problemlerinde tek başına yetersiz kalabilmektedir. Bu nedenle benzer şekilde NPV'nin de Tablo 4.3'te verilen diğer metrikler ile değerlendirilmesi tavsiye edilmektedir.

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \tag{4.6}$$

FPR (4.7), bir modelin yanlış pozitif tahminlerde bulunma oranını ifade eden bir performans metriğidir. Bu metrik, modelin negatif sınıfları yanlış bir şekilde pozitif olarak tanımlama olasılığını değerlendirir ve yanlış pozitif tahminlerin toplam gerçek negatif örneklere oranını belirleyerek modelin etkinliğini ölçer. FPR, özellikle yanlış pozitiflerin maliyetli olduğu alanlarda model performansını ölçmek için kullanışlıdır, fakat bazı sınıflandırma problemlerinde tek başına yetersiz kalabilmektedir. Bu nedenle benzer şekilde FPR'nin de Tablo 4.3'te verilen diğer metrikler ile değerlendirilmesi tavsiye edilmektedir.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{4.7}$$

FDR (4.8), bir modelin pozitif tahminlerinin ne kadarının yanlış olduğunu gösteren bir performans metriğidir. Bu metrik, modelin pozitif olarak sınıflandırdığı örnekler arasında ne kadarının aslında negatif olduğunu değerlendirir ve yanlış pozitif tahminlerin toplam pozitif tahminlere oranını belirleyerek modelin etkinliğini ölçer. FDR, özellikle yanlış pozitiflerin maliyetli olduğu alanlarda model performansını ölçmek için kullanışlıdır, fakat bazı sınıflandırma problemlerinde tek başına yetersiz kalabilmektedir. Bu nedenle benzer şekilde FDR'nin de Tablo 4.3'te verilen diğer metrikler ile değerlendirilmesi tavsiye edilmektedir.

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} \tag{4.8}$$

FNR (4.9), bir modelin gerçek pozitifleri yanlış bir şekilde negatif olarak sınıflandırma oranını ifade eden bir performans metriğidir. Bu metrik, modelin pozitif sınıfları kaçırma olasılığını değerlendirir ve yanlış negatif tahminlerin toplam gerçek pozitif örneklere oranını belirleyerek modelin etkinliğini ölçer. FNR, özellikle pozitif örneklerin kaçırılmasının maliyetli olduğu alanlarda model performansını ölçmek için kullanışlıdır, fakat bazı sınıflandırma problemlerinde tek başına yetersiz kalabilmektedir. Bu nedenle benzer şekilde FNR'nin de Tablo 4.3'te verilen diğer metrikler ile değerlendirilmesi tavsiye edilmektedir.

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}} \tag{4.9}$$

Tablo 4.3 Metrikler

	1	2	3	4	5	6	7	8	9
Metrikler	AC	SE/RE	SP	PR	F	NPV	FPR	FDR	FNR

4.6 Makine Öğrenmesi Algoritmalarının Performansları

Bu bölümde, WDBC veri seti kullanılarak, Bölüm 4.3'te verilen gözetimli/ denetimli makine öğrenmesi sınıflandırma algoritmaları ile eğitilen modellerin göğüs kanseri tanısındaki performansları kapsamlı bir şekilde değerlendirilmektedir.

Modellerin Tablo 4.3'te verilen performans metrikleri incelenmiş, her bir algoritmanın tanı koyma yetkinliği titizlikle analiz edilmiştir. Analiz sonucunda, belirli algoritmaların diğerlerine kıyasla daha iyi performans gösterdiği tespit edilmiştir. Bu bulgular, hangi algoritmanın spesifik tıbbi tanı senaryolarında daha etkili kullanılabileceği konusunda önemli ipuçları sağlamaktadır. Bu analizler, algoritmaların göğüs kanseri tanısında ne kadar etkili ve güvenilir olduklarını karşılaştırmalı olarak ortaya koymayı amaçlamaktadır.

Elde edilen sonuçlar, farklı algoritmaların güçlü ve zayıf yönlerini belirleyerek, en uygun tanı modelinin seçimine rehberlik etmeyi amaçlamaktadır.

Ridge, LR, SGD, Perceptron, PA, LDA, SVM-SVC, KNN-KNC, NC, GPC, NB-BNB, DT, GB, HGB, RF, AB ve NN-MLP algoritmalarının performansları AC, SE/RE, SP, PR, F, NPV, FPR, FDR, FNR metrikleri ve %80 eğitim veri seti, %20 test veri seti ayrımı ile Tablo 4.4'te, 10 katlı çapraz doğrulama ile Tablo 4.5'te verilmektedir.

LR, Perceptron ve SVM-SVC algoritmalarının tüm metrik sonuçlarının aynı değerlerden oluştuğu gözlemlenmiştir. Bu algoritmaların sonuçları aynı olsa da aralarında bazı farklılıklar mevcuttur. Örneğin, SVM-SVC algoritması ile

Perceptron algoritmalarının birbirlerinden farkı; SVM-SVC algoritması her iki sınıf arasındaki mesafenin en büyük olduğu düzlemi bulduktan sonra sonlanırken, Perceptron algoritması verileri doğru bir şekilde sınıflandırdıktan sonra sona erer (Kharel, 2020).

Tablo 4.4 Algoritma performansları

Algoritmalar	**Metrikler**								
	AC	SE/RE	SP	PR	F	NPV	FPR	FDR	FNR
Ridge	94,73	86,95	100,0	100,0	93,02	91,89	0,00	0,00	13,04
LR	98,24	95,65	100,0	100,0	97,77	97,14	0,00	0,00	4,34
SGD	95,61	91,30	98,52	97,67	94,38	94,36	1,47	2,32	8,69
Perceptron	98,24	95,65	100,0	100,0	97,77	97,14	0,00	0,00	4,34
PA	97,36	93,47	100,0	100,0	96,62	95,77	0,00	0,00	6,52
LDA	94,73	86,95	100,0	100,0	93,02	91,89	0,00	0,00	13,04
SVM-SVC	98,24	95,65	100,0	100,0	97,77	97,14	0,00	0,00	4,34
KNN-KNC	96,49	91,30	100,0	100,0	95,45	94,44	0,00	0,00	8,69
NC	90,35	80,43	97,05	94,87	87,05	88,00	2,94	5,12	19,56
GPC	96,49	93,47	98,52	97,72	95,55	95,71	1,47	2,27	6,52
NB-BNB	94,73	91,30	97,05	95,45	93,33	94,28	2,94	4,54	8,69
DT	92,10	89,13	94,11	91,11	90,10	92,75	5,88	8,88	10,86
GB	92,98	86,95	97,05	95,23	90,90	91,66	2,94	4,76	13,04
HGB	93,85	86,95	98,52	97,56	91,95	91,78	1,47	2,43	13,04
RF	92,98	84,78	98,52	97,50	90,69	90,54	1,47	2,5	15,21
AB	95,61	89,13	100,0	100,0	94,25	93,15	0,00	0,00	10,86
NN-MLP	97,36	95,65	98,52	97,77	96,70	97,10	1,47	2,22	4,34

Tablo 4.5 Algoritma performansları (10 katlı çapraz doğrulama ile)

Algoritmalar	**Metrikler**								
	AC	SE/RE	SP	PR	F	NPV	FPR	FDR	FNR
Ridge	94,64	90,47	97,14	95,00	92,68	94,44	2,85	5,00	9,52
LR	98,21	100,0	97,14	95,45	97,67	100,0	2,85	4,54	0,00
SGD	98,21	100,0	97,14	95,45	97,67	100,0	2,85	4,54	0,00
Perceptron	98,21	100,0	97,14	95,45	97,67	100,0	2,85	4,54	0,00
PA	98,21	100,0	97,14	95,45	97,67	100,0	2,85	4,54	0,00
LDA	94,64	90,47	97,14	95,00	92,68	94,44	2,85	5,00	9,52
SVM-SVC	98,21	100,0	97,14	95,45	97,67	100,0	2,85	4,54	0,00
KNN-KNC	98,21	95,23	100,0	100,0	97,56	97,22	0,00	0,00	4,76

NC	96,42	95,23	97,14	95,23	95,23	97,14	2,85	4,76	4,76
GPC	98,21	100,0	97,14	95,45	97,67	100,0	2,85	4,54	0,00
NB-BNB	96,42	95,23	97,14	95,23	95,23	97,14	2,85	4,76	4,76
DT	96,42	100,0	94,28	91,30	95,45	100,0	5,71	8,69	0,00
GB	98,21	95,23	100,0	100,0	97,56	97,22	0,00	0,00	4,76
HGB	98,21	95,23	100,0	100,0	97,56	97,22	0,00	0,00	4,76
RF	98,21	95,23	100,0	100,0	97,56	97,22	0,00	0,00	4,76
AB	98,21	95,23	100,0	100,0	97,56	97,22	0,00	0,00	4,76
NN-MLP	98,21	100,0	97,14	95,45	97,67	100,0	2,85	4,54	0,00

Tablo 4.5 Algoritma performansları (10 katlı çapraz doğrulama ile) (devamı)

4.7 Bazı Algoritma Parametrelerinin Performansa Etkileri

Bu bölümde, WDBC veri seti kullanılarak, Bölüm 4.3'te verilen SVM-SVC, KNN-KNC, DT ve RF gözetimli/ denetimli makine öğrenmesi sınıflandırma algoritmalarının bazı parametrelerinin göğüs kanseri tanı tahmin performanslarına olan etkileri incelenmiştir. Modellerin Tablo 4.3'te verilen performans metrikleri incelenmiş, parametrelerin tanı koyma yetkinliğine etkileri titizlikle analiz edilmiştir. Bu analizler, algoritma parametrelerinin kanser tanısında ne kadar etkili ve güvenilir olduklarını karşılaştırmalı olarak ortaya koymayı amaçlamaktadır. %80 eğitim veri seti, %20 test veri seti ayrımı ile elde edilen sonuçlar, farklı algoritma parametrelerinin güçlü ve zayıf yönlerini belirleyerek, en uygun tanı modelinin seçimine rehberlik etmeyi amaçlamaktadır.

SVM-SVC parametrelerinden "C", "kernel", "degree" ve "gamma" parametrelerinin performansa etkileri incelenmiştir.

"C" parametresinin tipi ondalıklı sayı olmakla birlikte bu parametrenin varsayılan değeri 1.0'dır. Varsayılan değer olan 1.0'dan farklı değerlerin de performansa etkileri incelenmiş olup en iyi performansın "C" parametresinin 1.0 varsayılan değerinde olduğu durumda elde edildiği gözlemlenmiştir.

"kernel" parametresi, 'lineer', 'polinom', 'radyal tabanlı fonksiyon', 'sigmoid' ve 'precomputed' çekirdek türlerinden oluşmaktadır. Bu parametrenin varsayılan çekirdek türü 'radyal tabanlı fonksiyon' olmakla birlikte diğer çekirdek türlerinin de performansa etkileri incelenmiştir.

Çekirdek türü ‘lineer’ olduğunda AC 95,61, SE/RE 91,30, SP 98,52, PR 97,67, F 94,38, NPV 94,36, FPR 1,47, FDR 2,32, FNR 8,69 olarak hesaplanmıştır.

Çekirdek türü ‘polinom’ olduğunda AC 85,96, SE/RE 65,21, SP 100,0, PR 100,0, F 78,94, NPV 80,95, FPR 0,0, FDR 0,0, FNR 34,78 olarak hesaplanmıştır.

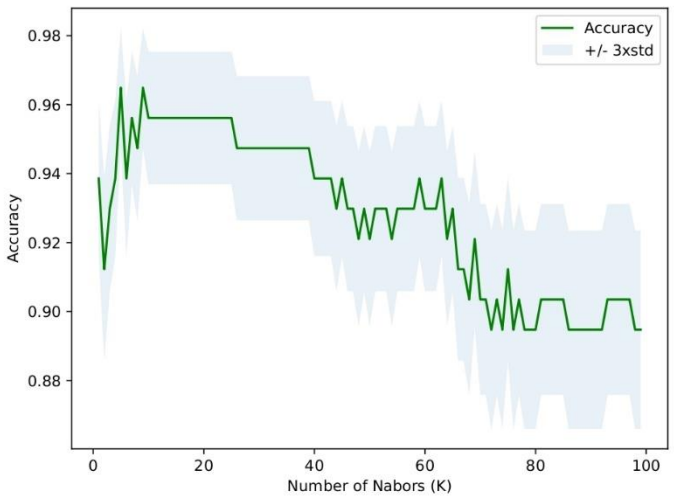
Çekirdek türü ‘sigmoid’ olduğunda AC 97,36, SE/RE 95,65, SP 98,52, PR 97,77, F 96,70, NPV 97,10, FPR 1,47, FDR 2,22, FNR 4,34 olarak hesaplanmıştır.

“degree” parametresinin tipi tam sayı olmakla birlikte bu parametrenin varsayılan değeri 3’tür. Varsayılan değer olan 3’ten farklı değerlerin de performansa etkileri incelenmiş olup en iyi performansın “degree” parametresinin 3 varsayılan değerinde olduğu durumda elde edildiği gözlemlenmiştir.

“gamma” parametresi, ‘scale’, ‘auto’ veya bir ondalıklı sayı olabilir. “gamma” parametresinin varsayılan tipi ‘scale’ olmakla birlikte diğer tiplerin de performansa etkileri incelenmiştir. “gamma” parametresi ‘auto’ seçildiğinde varsayılan tip olan ‘scale’ ile aynı performansın elde edildiği görülmüştür.

KNN-KNC parametrelerinden “n_neighbors”, “weights”, “algorithm”, “leaf_size”, “p”, “metric” ve “n_jobs” parametrelerinin performansa etkileri incelenmiştir.

“n_neighbors” parametresi, kullanılacak komşu sayısını ifade etmektedir. Bu parametre doğal olarak tam sayı değerleri almaktadır. En yüksek AC değerini veren komşu sayısının bulunması için bu parametre bir döngüye sokulmuştur. En yüksek AC değerini veren komşu sayısının varsayılan değer olan 5 olduğu hesaplanmıştır (Şekil 4.13).



Şekil 4.13 AC ile K değerleri arasındaki ilişki

“weights” parametresi, komşulara ne kadar ağırlık verileceğini ifade eden parametredir. “weights” parametresi, ‘uniform’ ve ‘distance’ tipinde olabilir. ‘uniform’ parametresi her komşuya eşit ağırlık verirken ‘distance’ parametresi uzaklık ile ters orantılı olacak şekilde komşulara ağırlık vermektedir. Bu parametrenin varsayılan tipi ‘uniform’ olmakla birlikte parametre ‘distance’ seçildiğinde de performansta bir değişikliğin olmadığı hesaplanarak gözlemlenmiştir.

“algorithm” parametresi, en yakın komşuları hesaplamak için kullanılan algoritmalardır. ‘auto’ değeri veya ‘ball_tree’, ‘kd_tree’ ve ‘brute’ algoritmalarından biri seçilebilir. Bunula beraber “algorithm” parametresinin varsayılan tipi ‘auto’ olarak tanımlanmıştır. Bu üç algoritma da incelenmiş olup performansta bir değişikliğin olmadığı hesaplanarak gözlemlenmiştir.

“leaf_size” parametresinin tipi tam sayı olmakla birlikte bu parametrenin varsayılan değeri 30’dur. Varsayılan değer olan 30’dan farklı değerlerin de performansa etkileri incelenmiş olup 500’e kadar olan değerler için performansta bir değişikliğin olmadığı hesaplanarak gözlemlenmiştir.

“p” parametresinin tipi ondalıklı sayı olmakla birlikte bu parametrenin varsayılan değeri 2’dir. “p” parametresi 1 seçilirse Manhattan uzaklığı, “p” parametresi 2 seçilirse Euclid uzaklığı kullanılmış olur. Rastgele bir “p” değeri için ise Minkowski metriği kullanılmış olmaktadır. Varsayılan değer olan 2’den farklı değerlerin de performansa etkileri incelenmiş olup 34’e kadar olan değerler için performansta bir değişikliğin olmadığı hesaplanarak gözlemlenmiştir.

“metric” parametresi, mesafe hesaplamada kullanılacak metriği belirtir. Bu parametre bir tam sayı değeri almakla beraber varsayılan tipi, “p” değeri 2 olduğunda elde edilen standart Euclid uzaklığının ifade edildiği Minkowski metriğidir. Varsayılan tip Minkowski’den farklı değerlerin de performansa etkileri incelenmiş olup en yüksek AC değerini veren değerin varsayılan tip olan Minkowski olduğu hesaplanarak gözlemlenmiştir.

“n_jobs” parametresini 1 veya -1 değerlerini alabilmektedir. Bu parametrenin varsayılan değeri 1’dir. Varsayılan değer olan 1’den farklı olarak -1 değeri için de performansa etkisi incelenmiş olup -1 değerleri için performansta bir değişikliğin olmadığı hesaplanarak gözlemlenmiştir.

DT parametrelerinden “criterion”, “splitter”, “max_depth”, “min_samples_split”, “min_samples_leaf” ve “min_weight_fraction_leaf” parametrelerinin performansa etkileri incelenmiştir (Tablo 4.6).

“criterion” parametresi, ‘gini’, ‘entropy’ ve ‘log_loss’ kriterlerinden birisi olabilmektedir. Bu parametrenin varsayılan tipi ‘gini’ kriteri olmakla birlikte diğer kriterlerin de performansa etkileri incelenmiştir.

“criterion” parametresi ‘entropy’ ve ‘log_loss’ olduğunda AC 91,22, SE/RE 86,95, SP 94,11, PR 90,90, F 88,88, NPV 91,42, FPR 5,88, FDR 9,09, FNR 13,04 olarak hesaplanmıştır.

“splitter” parametresi, ‘best’ ve ‘random’ değerlerinden birisi olabilmektedir. Varsayılan değer ‘best’ olmakla birlikte ‘random’ değerinin de performansa etkisi incelenmiştir. “splitter” parametresi ‘random’ olarak seçildiğinde AC 89,47, SE/RE 82,60, SP 94,11, PR 90,47, F 86,36, NPV 88,88, FPR 5,88, FDR 9,52, FNR 17,39 olarak hesaplanmıştır.

“max_depth” parametresi, tam sayı değerlerden oluşmaktadır. 1 ile 10 arasında farklı değerler seçilerek “max_depth” parametresinin performansa etkisi araştırılmıştır. En başarılı performans “max_depth” parametresi 2 olduğunda gerçekleşmiştir. “max_depth” parametresi 2 seçildiğinde AC 91,22, SE/RE 86,95, SP 94,11, PR 90,90, F 88,88, NPV 91,42, FPR 5,88, FDR 9,09, FNR 13,04 olarak hesaplanmıştır.

“min_samples_split” parametresi, tam sayı veya ondalık değerlerden oluşmaktadır. 1 ile 10 arasında farklı değerler seçilerek “min_samples_split” parametresinin performansa etkisi araştırılmıştır. Varsayılan değer 2 olmakla birlikte en başarılı performans “min_samples_split” parametresi 2,7 ve 8 olduğunda gerçekleşmiştir. “min_samples_split” parametresi 8 olarak seçildiğinde en yüksek AC 92,10, en yüksek SP 95,58, en yüksek PR 93,02, en yüksek F 89,88, en yüksek NPV 91,54 olarak hesaplanmıştır. “min_samples_split” parametresi 7 olarak seçildiğinde en yüksek SE/RE 86,95 olarak hesaplanmıştır. “min_samples_split” parametresi 2 olarak seçildiğinde ise en düşük FPR 4,41, en düşük FDR 6,97 ve en düşük FNR 13,04 olarak hesaplanmıştır.

“min_samples_leaf” parametresi, tam sayı veya ondalık değerlerden oluşmaktadır. “min_samples_leaf” parametresinin 1 ile 10 arasındaki değerlerinin performansa

etkileri hesaplanarak incelenmiştir. Varsayılan değer 2 olmakla birlikte en başarılı performans "min_samples_leaf" parametresi 2,3 ve 5 olduğunda gerçekleşmiştir. "min_samples_leaf" 2 olarak seçildiğinde en yüksek SP 97,05 ve en düşük FNR 13,04 olarak hesaplanmıştır. "min_samples_leaf" 3 olarak seçildiğinde en yüksek AC 92,98, en yüksek SE/RE 86,96, en yüksek PR 95,23, en yüksek F 90,90 ve en yüksek NPV 91,66 olarak hesaplanmıştır.

"min_samples_leaf" 5 olarak seçildiğinde ise en düşük FPR 2,94, en düşük FDR 4,76 ve en düşük FNR 13,04 olarak hesaplanmıştır.

"min_weight_fraction_leaf" parametresi, ondalık değerlerden oluşmaktadır. Varsayılan değer 0,0 olmakla birlikte "min_weight_fraction_leaf" parametresi 0,4 seçildiğinde diğer bazı değerlere göre daha iyi performans sergilemektedir. "min_weight_fraction_leaf" parametresi 0,4 seçildiğinde AC 94,73, SE/RE 93,47, SP 95,58, PR 93,47, F 93,47, NPV 95,58, FPR 4,41, FDR 6,52, FNR 6,52 olarak hesaplanmıştır.

RF parametrelerinden "n_estimators", "criterion", "max_depth", "min_samples_split" ve "min_samples_leaf" parametrelerinin performansa etkileri incelenmiştir (Tablo 4.6).

"n_estimators" parametresi, ormandaki ağaç sayısını ifade etmektedir. Bu parametrenin varsayılan tipi tam sayı olmakla birlikte bu parametrenin varsayılan değeri 100'dür. Varsayılan değer olan 100'den farklı değerlerin de performansa etkileri bir döngü yardımı ile incelenmiş olup en iyi performansın "n_estimators" parametresinin 10 olduğu durumda elde edildiği gözlemlenmiştir.

"n_estimators" parametresi 10 iken en yüksek AC 92,98, en yüksek SE/RE 84,78, en yüksek SP 98,52, en yüksek PR 97,50, en yüksek F 90,69, en yüksek NPV 90,54, en düşük FPR 0,0, en düşük FDR 0,0 ve en düşük FNR 0,0 olarak hesaplanmıştır.

"criterion" parametresi, 'gini', 'entropy' ve 'log_loss' kriterlerinden birisi olabilmektedir. Bu parametrenin varsayılan tipi 'gini' kriteri olmakla birlikte diğer kriterlerin de performansa etkileri incelenmiştir. Bu üç kriter tipinin de eşit performans gösterdiği hesaplanarak gözlemlenmiştir.

"max_depth" parametresi, bir tam sayı değerine eşittir. Varsayılan değer tanımlanmamış durumdadır. Farklı değerlerin performansa etkileri bir döngü

yardımı ile incelenmiş olup en iyi performansın "max_depth" parametresinin 2 olduğu durumda elde edildiği gözlemlenmiştir.

"max_depth" parametresi 2 iken en yüksek AC 92,98, en yüksek SE/RE 84,78, en yüksek SP 98,52, en yüksek PR 97,50, en yüksek F 90,69, en yüksek NPT 90,54, en düşük FPR 1,47, en düşük FDR 2,5 ve en düşük FNR 15,21 olarak hesaplanmıştır.

"min_samples_split" parametresi, tam sayı veya ondalıklı sayı olabilmektedir. Bu parametrenin varsayılan değeri 2 olmakla birlikte diğer değerlerinde performansa etkileri incelenmiştir. En yüksek AC, en yüksek SE/RE, en yüksek SP, en yüksek PR, en yüksek F, en yüksek NPV, en düşük FPR, en düşük FDR ve en düşük FNR değerlerini veren sayının varsayılan değer olan 2 olduğu hesaplanarak gözlemlenmiştir.

min_samples_leaf parametresi, tam sayı veya ondalıklı sayı olabilmektedir. Bu parametrenin varsayılan değeri 1 olmakla birlikte diğer değerlerinde performansa etkileri incelenmiştir. En yüksek AC, en yüksek SE/RE, en yüksek SP, en yüksek PR, en yüksek F, en yüksek NPV, en düşük FPR, en düşük FDR ve en düşük FNR değerlerini veren sayının 2 olduğu hesaplanarak gözlemlenmiştir.

Tablo 4.6 Parametre performansları

Algoritma ve parametreleri			**Metrikler**								
		Değer	AC	SE/RE	SP	PR	F	NPV	FPR	FDR	FNR
DT	min_samples_split	8	-	-	95,58	93,02	-	-	-	-	-
	min_samples_leaf	2	-	-	97,05						
		3	92,98	-	-	95,23	90,90	-	-	-	-
		5	-	-	-	-	-	-	2,94	4,76	-
	min_weight_fraction_leaf	0,4	94,73	93,47	95,58	93,47	93,47	95,58	4,41	6,52	6,52
RF	n_estimators	10	-	-	-	-	-	-	0,00	0,00	0,00

5
SONUÇ

Bu bölümde Bölüm 1.2'de verilen çalışmalar arasından WDBC veri setini kullanarak makine öğrenmesi sınıflandırma algoritmalarının performanslarını kıyaslayan çalışmaların sonuçları ile bu tez çalışmasından elde edilmiş makine öğrenmesi sınıflandırma algoritmalarının performanslarının sonuçları Tablo 5.1'de verilmektedir. Kaynakçada yer alan çalışmaların sonuçları kısaca K1, K2, …, K10 ile ve bu tez çalışmasında elde edilen sonuçlar ise kısaca S1 ve S2 ile ifade edilmiştir.

- K1: (Aruna, Rajagopalan, & Nandakishore, 2011)
- K2: (Oyewola, Hakimi, Adeboye, & Shehu, 2016)
- *K3: (Sharma, Aggarwal, & Choudhury, 2018)*
- *K4: (Westerdijk & Bhulai, 2018)*
- K5: (Ak, 2020)
- K6: (Gupta & Garg, 2020)
- K7: (Naji, ve diğerleri, 2021)
- *K8: (Sakib, ve diğerleri, 2022)*
- K9: (Chhillar & Singh, 2024)
- K10: (Sevli, 2019)
    - S1: %80 eğitim veri seti, %20 test veri seti ayrımı ile
    - *S2: 10 katmanlı çapraz doğrulama ile*

Tablo 5.1 Performans kıyasları

Algoritmalar	**Çalışmalar**	**Metrikler**						
		AC	SE/RE	SP	PR	F	FPR	FNR
LR	K2	94,5	94,6	94,5	-	-	-	-
	K4	*97,35*	*95,24*	*98,59*	-	-	-	-
	K5	98,06	-	-	-	-	-	-
	K6	95,80	95,00	-	96,50	95,50	-	-
	K7	95,80	91,00	-	98,00	94,00	-	-
	K8	*95,08*	*85,70*	*98,60*	*97,30*	*91,10*	*0,14*	*14,3*
	K10	**98,24**	95,35	**100,0**	**100,0**	-	-	-
	S1	**98,24**	95,65	**100,0**	**100,0**	**97,77**	**0,00**	4,34
	S2	*98,21*	***100,0***	*97,14*	*95,45*	*97,67*	*2,85*	***0,00***

LDA	K2	93,50	**92,30**	93,5	-	-	-	-
	S1	**94,73**	86,95	**100,0**	**100,0**	**93,02**	**0,00**	13,04
	S2	*94,64*	*90,47*	*97,14*	*95,00*	*92,68*	*2,85*	***9,52***
SVM	K1	98,06	97,20	97,80	-	-	-	-
	K2	95,80	98,4	94,4	-	-	-	-
	K4	*98,23*	*95,24*	***100,0***	-	-	-	-
	K5	96,49	-	-	-	-	-	-
	K6	97,20	97,00	-	97,50	97,00	-	-
	K7	97,20	94,40	-	98,00	96,00	-	-
	K8	*94,73*	*85,70*	*97,20*	*94,70*	*90,00*	*0,28*	*14,30*
	K9	97,90	96,30	-	98,10	-	-	-
	K10	93,86	86,05	98,59	97,37	-	-	-
	S1	**98,24**	95,65	**100,0**	**100,0**	**97,77**	**0,00**	4,34
	S2	*98,21*	***100,0***	*97,14*	*95,45*	*97,67*	*2,85*	***0,00***
KNN	*K3*	*95,90*	*90,47*	-	*98,27*	*94,20*	-	-
	K5	96,49	-	-	-	-	-	-
	K6	95,80	93,50	93,50	93,00	-	-	-
	K7	93,70	91,00	-	92,00	91,00	-	-
	K8	*93,32*	*78,60*	***100,0***	***100,0***	*88,00*	***0,00***	*21,40*
	K9	96,50	**96,30**	-	94,50	-	-	-
	K10	91,23	81,39	97,18	94,59	-	-	-
	S1	96,49	91,30	**100,0**	**100,0**	95,45	**0,00**	8,69
	S2	***98,21***	*95,23*	***100,0***	***100,0***	***97,56***	***0,00***	***4,76***
NB	K1	92,61	**95,70**	97,80	-	-	-	-
	K3	*94,47*	*85,71*	-	*88,52*	*87,09*	-	-
	K5	94,73	-	-	-	-	-	-
	K9	95,10	94,40	-	92,70	-	-	-
	K10	94,74	86,05	**98,59**	**97,37**	-	-	-
	S1	94,73	91,30	97,05	95,45	93,33	**2,94**	8,69
	S2	***96,42***	*95,23*	*97,14*	*95,23*	***95,23***	*2,85*	***4,76***
DT	K1	92,97	94,4	93,90	-	-	-	-
	K5	95,61	-	-	-	-	-	-
	K6	95,80	95,00	-	96,50	**95,50**	-	-
	K7	95,10	92,00	-	94,00	93,00	-	-
	K8	*94,20*	*88,10*	***97,20***	***94,90***	*91,40*	***0,28***	*11,90*
	K9	94,40	96,30	-	89,70	-	-	-
	S1	92,10	89,13	94,11	91,11	90,10	5,88	10,86
	S2	***96,42***	***100,0***	*94,28*	*91,30*	*95,45*	*5,71*	***0,00***

RF	K2	94,2	95,30	95,30	-	-	-	-
	K3	*94,74*	*93,65*	-	*92,18*	*92,90*	-	-
	K4	*97,35*	*92,86*	***100,0***	-	-	-	-
	K5	95,61	-	-	-	-	-	-
	K6	97,20	**97,50**	-	97,00	97,00	-	-
	K7	96,50	94,00	-	94,00	95,00	-	-
	K8	*96,66*	*92,90*	***100,0***	***100,0***	*96,30*	***0,00***	***0,71***
	K9	97,90	96,30	-	98,10	-	-	-
	K10	96,69	95,35	97,18	95,35	-	-	-
	S1	92,98	84,78	98,52	97,50	90,69	1,47	15,21
	S2	***98,21***	*95,23*	***100,0***	***100,0***	***97,56***	***0,00***	*4,76*
AB	K9	95,10	94,40	-	92,70	-	-	-
	S1	95,61	89,13	**100,0**	**100,0**	94,25	**0,00**	10,86
	S2	***98,21***	***95,23***	***100,0***	***100,0***	***97,56***	***0,00***	***4,76***
NN	K1	93,67	95,90	97,80	-	-	-	-
	K4	*97,35*	*92,86*	***100,0***	-	-	-	-
	K8	*90,35*	*76,20*	*98,60*	*97,00*	*85,40*	***0,14***	*23,8*
	S1	97,36	95,65	98,52	**97,77**	96,70	1,47	4.34
	S2	***98,21***	***100,0***	*97,14*	*95,45*	***97,67***	*2,85*	***0,00***

Tablo 5.1 Performans Kıyasları (devamı)

K1'de KFCV'nin CART için gerekliliğinden bahsedilmiş fakat uygulamada kullanılıp kullanılmadığı belirtilmemiştir. K2'de, K5'te, K6'da K7'de, K9'da ve K10'da KFCV kullanılmamıştır. *K3*'te, *K4*'te ve *K8*'de 10 katmanlı CV kullanılmıştır.

Tablo 5.1 de en iyi performanslar kalınlaştırılarak gösterilmektedir. Bunun yanı sıra 10 katmanlı çapraz doğrulama ile elde edilen sonuçlar italik olarak verilmiştir.

KAYNAKÇA

(1990). https://pages.cs.wisc.edu/~street/saves/xcyt1.gif adresinden alındı

(2024, 4 1). American Cancer Society: https://www.cancer.org/cancer/types/breast-cancer/about/what-is-breast-cancer.html#where-breast-cancer-starts adresinden alındı

(2024). 4 1, 2024 tarihinde Türk Tıbbi Onkoloji Derneği: https://www.kanser.org/saglik/toplum/kanser-turleri-alt-kategori/meme-kanseri adresinden alındı

(2024). 4 2, 2024 tarihinde Memorial Sloan Kettering Cancer Center: https://www.mskcc.org/cancer-care/types/breast/anatomy-breast adresinden alındı

Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013, 4 24). Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform, 4*(2). doi:10.4172/2157-7420.1000124

Ak, M. F. (2020, 4 26). A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. *In Healthcare, 8*(2), s. 111. doi:10.3390/healthcare8020111

Akar, M., & Sirakov, N. M. (2019, 10 1). Support vector machine skin lesion classification. *Applications of Mathematics, 64*(5), s. 581-598. doi:10.21136/AM.2019.0292-18

Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, 4). Breast cancer classification using machine learning. *In 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT)*, s. 1-4. 4 2, 2024 tarihinde alındı

Aruna, S., Rajagopalan, S. P., & Nandakishore, L. V. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. *Computer Science & Information Technology, 2*, s. 37-45. doi:10.5121/csit.2011.1205

Asif, K. (2024). *What are classification problems?* 3 24, 2024 tarihinde Educative: https://www.educative.io/answers/what-are-classification-problems adresinden alındı

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science, 83*, s. 1064-1069.

Badik, Ş. T., & Akar, M. (2024, 2). Machine learning classification models for the patients who have heart failure. *Sigma Journal of Engineering and Natural Sciences, 42*(1), s. 235−244. doi:10.14744/sigma.2023.00095

Bourke, P. (2003, 2). 4 3, 2024 tarihinde https://paulbourke.net/fractals/fracdim/ adresinden alındı

Chhillar, I., & Singh, A. (2024, 3 25). Performance evaluation of machine learning techniques for breast cancer detection using WDBC dataset. *In AIP Conference Proceedings, 2919*(1). doi:https://doi.org/10.1063/5.0184603

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2 b.). (N. Tache, Dü.) Sebastopol, United States of America: O'Reilly.

Gupta, P., & Garg, S. (2020). Breast cancer prediction using varying parameters of machine learning models. *Procedia Computer Science, 171*, s. 593-601. doi:10.1016/j.procs.2020.04.064

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020, 9 1). Breast cancer prediction: A comparative study using machine learning techniques. *SN Computer Science, 1*, s. 1-14. doi:https://doi.org/10.1007/s42979-020-00305-w

Kharel, S. (2020, 5 13). *Perceptron vs SVM: a quick comparison*. 5 1, 2024 tarihinde medium: https://medium.com/@subashkharel/perceptron-vs-svm-a-quick-comparison-6b5d6b5d64f adresinden alındı

Khourdifi, Y., & Bahaj, M. (2018, 12). Applying best machine learning algorithms for breast cancer prediction and classification. *2018 International conference on electronics, control, optimization and computer science (ICECOCS)*, (s. 1-5). 4 9, 2024 tarihinde alındı

Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021, 8 9-12). Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science, 191*, s. 487-492. doi:10.1016/j.procs.2021.07.062

National Cancer Institute. (2021, 10 11). 3 18, 2024 tarihinde https://www.cancer.gov/about-cancer/understanding/what-is-cancer adresinden alındı

National Cancer Institute. (2023, 6 26). 4 3, 2024 tarihinde National Cancer Institute: https://www.cancer.gov/types/breast/patient/breast-screening-pdq. adresinden alındı

Oyewola, D., Hakimi, D., Adeboye, K., & Shehu, M. D. (2016, 12 23). Using five machine learning for breast cancer biopsy predictions based on mammographic diagnosis. *International Journal of Engineering Technologies IJET, 2*(4), s. 142-145.

Sakib, S., Yasmin, N., Tanzeem, A. K., Shorna, F., Md. Hasib, K., & Alam, S. B. (2022). Breast cancer detection and classification: A comparative analysis using machine learning algorithms. *In Proceedings of Third International Conference on Communication, Computing and Electronics Systems: ICCCES 2021* (s. 703-717). Singapore: Springer Singapore. doi:https://doi.org/10.1007/978-981-16-8862-1_46

Samuel, A. L. (1959, 7). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), s. 210-229. doi:10.1147/rd.33.0210

Sevli, O. (2019, 8 31). Göğüs kanseri teşhisinde farklı makine öğrenmesi tekniklerinin performans karşılaştırması. *European Journal of Science and Technology*(16), s. 176-185. doi:10.31590/ejosat.553549

Sharma, S., Aggarwal, A., & Choudhury, T. (2018, 12). Breast cancer detection using machine learning algorithms. *In 2018 International conference on*

computational techniques, electronics and mechanical systems (CTEMS), s. 114-118. 4 14, 2024 tarihinde alındı

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2022, 2 4). doi: 10.3322/caac.21660

Supervised learning. (2024). Scikit Learn: https://scikit-learn.org/stable/index.html adresinden alındı

Tenekeci, N. (2024). 4 3, 2024 tarihinde NuriTenekeci: https://www.nuritenekeci.com/tr/icerik/1/prof-dr-nuri-tenekeci-tani-merkezi adresinden alındı

Westerdijk, L., & Bhulai, S. (2018, 7 10). Predicting malignant tumor cells in breasts. *Master Business Analytics*.

Wolberg, W. H., Mangasarian, O. L., & Street, N. (1990). *Machine Learning for Cancer Diagnosis and Prognosis*. 1 25, 2024 tarihinde wisc.edu: https://pages.cs.wisc.edu/~olvi/uwmp/cancer.html#diag adresinden alındı

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995, 10 31). *Breast Cancer Wisconsin (Diagnostic)*. doi:10.24432/C5DW2B

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. Demirel, Y. E., & Akar, M. (2024, March 13-16). The effects of machine learning classification algorithm parameters on breast cancer diagnosis prediction, *4th International Symposium of Scientific Research and Innovative Studies (ISSRIS'24)*, s. 573-574. Balıkesir, Türkiye. https://www.issris.org/