T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

KOLLEKTİF MAKİNE ÖĞRENMESİ METODLARI ile GÖĞÜS KANSERİ TEŞHİSİ

Fulya AKCAN

DANIŞMAN
Prof. Dr. Ahmet SERTBAŞ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

İSTANBUL-2021

Bu çalışma, 4.06.2021 tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Programında Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi

Prof. Dr. Ahmet SERTBAŞ (Danışman)
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Dr. Öğr. Üyesi Oğuzhan ÖZTAŞ
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Doç. Dr. Akhan AKBULUT
İstanbul Kültür Üniversitesi
Mühendislik Fakültesi

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi-Cerrahpaşa'nın abonesi olduğu intihal yazılım programı kullanılarak Lisansüstü Eğitim Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

ÖNSÖZ

Tezin yürütülmesi aşamasında bana yol gösteren ve yardımlarını esirgemeyen tez danışmanım Sayın Prof. Dr. Ahmet Sertbaş'a teşekkürlerimi sunarım.

Hayatım boyunca her zaman yanımda olan, maddi ve manevi desteklerini esirgemeyen aileme teşekkür ederim.

Haziran 2021 Fulya AKCAN

İÇİNDEKİLER

Sayfa No

ÖNSÖZ ..... **iv**
İÇİNDEKİLER ..... **v**
ŞEKİL LİSTESİ ..... **viii**
TABLO LİSTESİ ..... **x**
KISALTMA LİSTESİ ..... **xi**
ÖZET ..... **xii**
SUMMARY ..... **xiii**
1. GİRİŞ **1**
1.1. AMAÇ 2
1.2. TEZİN ORGANİZASYONU 2
2. GENEL KISIMLAR **3**
2.1. GÖĞÜS KANSERİ 3
2.1.1. Göğüs Kanseri Risk Faktörleri 3
2.1.2. Göğüs Kanseri Belirtileri 4
2.1.3. Göğüs Kanserinde Evrelendirme Sistemleri 5
2.1.3.1. TNM Evrelendirme Sistemi 5
2.1.3.2. Genel Evrelendirme Sistemi 8
2.1.4. Göğüs Kanseri Tanı Yöntemleri 9
2.1.5. Göğüs Kanserinde Tedavi Yöntemleri 11
2.1.5.1. Ameliyat 11
2.1.5.2. Radyasyon 12
2.1.5.3. İlaç Tedavisi 12
2.1.6. Tekrarlayan Göğüs Kanseri 13
2.2. MAKİNE ÖĞRENMESİ 14
2.3. TOPLULUK YÖNTEMLERİ 15
2.4. LİTERATÜR TARAMASI 16
3. MALZEME VE YÖNTEM **19**
3.1. ÖNERİLEN GÖĞÜS KANSERİ TESPİT SİSTEMİ 19
3.2. VERİ KÜMESİ BÖLÜMLENDİRME 20
3.3. VERİ ÖN İŞLEME 20

3.4. BOYUT İNDİRGEME....21
3.5. SINIFLANDIRMA YÖNTEMLERİ....22
3.5.1. Destek Vektör Makinesi....22
3.5.2. K-En Yakın Komşu....24
3.5.3. Saf Bayes....25
3.5.4. Karar Ağacı....27
3.5.4.1. Öznitelik Seçim Ölçüsü....28
3.6. TOPLULUK YÖNTEMLERİ....29
3.6.1. Rassal Orman....29
3.6.2. Oylama....30
3.6.2.1. Çoğunluğa Dayalı Oylama....31
3.6.2.2. Olasılığa Dayalı Oylama....31
3.6.3. Torbalama....32
3.6.4. Yükseltme....33
3.6.5. Yığma....35
3.6.6. Ekstra Ağaçlar....36
3.7. MATERYAL....37
3.7.1. Veri Kümesi....37
3.7.2. Colab Notebook....39
3.7.3. Sklearn....39
3.8. MODEL BAŞARIM ÖLÇÜTLERİ....39
3.8.1. Doğruluk....40
3.8.2. Kesinlik....41
3.8.3. Duyarlılık....41
3.8.4. F-Skor....41
3.8.5. Roc....41
3.8.6. Auc....42
4. BULGULAR....43
4.1. BİREYSEL SINIFLANDIRICILARIN MODEL BAŞARIM ORANLARI....43
4.1.1. Destek Vektör Makinesi....43
4.1.2. K-En Yakın Komşu....44
4.1.3. Saf Bayes....44
4.1.4. Karar Ağacı....45
4.2. TOPLULUK YÖNTEMLERİNİN MODEL BAŞARIM ORANLARI....46

4.2.1. Rassal Orman....46
4.2.2. Oylama....47
4.2.2.1. Çoğunluğa Dayalı Oylama....47
4.2.2.2. Olasılığa Dayalı Oylama....48
4.2.3. Torbalama....49
4.2.3.1. SVC Temel Öğrenicisi ile Torbalama Yöntemi....49
4.2.3.2. Gaussian Naive Bayes Temel Öğrenicisi ile Torbalama Yöntemi....50
4.2.3.3. Karar Ağacı Temel Öğrenici ile Torbalama Yöntemi....51
4.2.4. Yükseltme....52
4.2.4.1. Adaboost....52
4.2.4.2. Gradient Boosting....56
4.2.4.3. XGBoost....57
4.2.5. Yığma....58
4.2.6. Ekstra Ağaçlar....59
4.3. MODELLERİN PERFORMANS KARŞILAŞTIRMALARI....60
5. TARTIŞMA VE SONUÇ....62
KAYNAKLAR....65
ÖZGEÇMİŞ....69

ŞEKİL LİSTESİ

Sayfa No

Şekil 3.1: Önerilen sistemin çalışma mekanizması. 19
Şekil 3.2: SVM algoritmasının çalışma mekanizması. 23
Şekil 3.3: KNN algoritmasının çalışma mekanizması. 25
Şekil 3.4: Gaussian dağılımı. 26
Şekil 3.5: Karar Ağacı yapısı. 28
Şekil 3.6: İkili sınıflandırmaya göre entropi fonksiyonunun formu. 29
Şekil 3.7: Rassal Orman modeli. 30
Şekil 3.8: Oylama topluluk yöntemi iş akışı. 31
Şekil 3.9: Torbalama topluluk yöntemi iş akışı. 33
Şekil 3.10: Yükseltme topluluk yöntemi çalışma prensibi. 34
Şekil 3.11: Yığma topluluk yöntemi çalışma prensibi. 36
Şekil 3.12: FNA'nın sayısallaştırılmış görüntüleri: (a) İyi huylu, (b) Kötü huylu. 38
Şekil 3.13: Hata matrisi. 40
Şekil 3.14: ROC eğrisi. 42
Şekil 4.1: PCA ve Standardizasyon ön işleme işlemi sonrası SVC hata matrisi. 43
Şekil 4.2: PCA ve Standardizasyon ön işleme işlemi sonrası KNN hata matrisi. 44
Şekil 4.3: PCA ve Standardizasyon ön işleme işlemi sonrası Gaussian Naive Bayes hata matrisi. 45
Şekil 4.4: PCA ve Standardizasyon ön işleme işlemi sonrası Karar Ağacı hata matrisi. 45
Şekil 4.5: Rassal Orman hata matrisi. 46
Şekil 4.6: Rassal Orman ROC eğrisi. 46
Şekil 4.7: Hard Voting hata matrisi 47
Şekil 4.8: Hard Voting ROC eğrisi. 47
Şekil 4.9: Soft Voting hata matrisi. 48
Şekil 4.10: Soft Voting ROC eğrisi. 48
Şekil 4.11: Bagging(SVC) hata matrisi. 49
Şekil 4.12: Bagging(SVC) ROC eğrisi. 50
Şekil 4.13: Bagging(GaussianNB) hata matrisi. 50
Şekil 4.14: Bagging(GaussianNB) ROC eğrisi. 51
Şekil 4.15: Bagging(Karar Ağacı) hata matrisi. 51

Şekil 4.16: Bagging(Karar Ağacı) ROC eğrisi. 52
Şekil 4.17: Adaboost(SVC) hata matrisi. 53
Şekil 4.18: Adaboost(SVC) ROC eğrisi. 53
Şekil 4.19: Adaboost(GaussianNB) hata matrisi. 54
Şekil 4.20: Adaboost(GaussianNB) ROC eğrisi. 54
Şekil 4.21: Adaboost(Karar Ağacı) hata matrisi. 55
Şekil 4.22: Adaboost(Karar Ağacı) ROC eğrisi. 55
Şekil 4.23: Gradient Boosting hata matrisi. 56
Şekil 4.24: Gradient Boosting ROC eğrisi. 56
Şekil 4.25: XGBoost hata matrisi. 57
Şekil 4.26: XGBoost ROC eğrisi. 57
Şekil 4.27: Stacking hata matrisi. 58
Şekil 4.28: Stacking ROC eğrisi. 58
Şekil 4.29: Extra Trees hata matrisi. 59
Şekil 4.30: Extra Trees ROC eğrisi. 59

TABLO LİSTESİ

Sayfa No

Tablo 4.1: Bireysel sınıflandırıcıların Standardizasyon+PCA ön işleme işlemi sonrası model başarım oranları. 60

Tablo 4.2: Topluluk öğrenme yöntemlerinin model başarım ölçütlerine göre karşılaştırılması.....61

Tablo 5.1: SVC bireysel sınıflandırıcısının SVC tabanlı topluluk sınıflandırıcıları ile model başarım ölçütlerine göre karşılaştırılması. 62

Tablo 5.2: Gaussian Naive Bayes bireysel sınıflandırıcısının Gaussian Naive Bayes tabanlı topluluk sınıflandırıcıları ile karşılaştırılması.....63

Tablo 5.3: Karar Ağacı bireysel sınıflandırıcısının Karar Ağacı tabanlı topluluk sınıflandırıcıları ile karşılaştırılması. 64

KISALTMA LİSTESİ

Kısaltmalar	Açıklama
Adaboost	: Adaptive Boosting
AUC	: Area Under the Receiver Operating Characteristic Curve
KNN	: K-Nearest Neighbor
PCA	: Principal Component Analysis
ROC	: Receiver Operating Characteristic
SVM	: Support Vector Machine
WDBC	: Wisconsin Diagnostic Breast Cancer
XGBoost	: Extreme Gradient Boosting

ÖZET

KOLLEKTİF MAKİNE ÖĞRENMESİ METODLARI ile GÖĞÜS KANSERİ TEŞHİSİ

YÜKSEK LİSANS TEZİ

Fulya AKCAN

İstanbul Üniversitesi-Cerrahpaşa

Lisansüstü Eğitim Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman : Prof. Dr. Ahmet SERTBAŞ

Günümüzde, birçok insan göğüs kanserinden etkilenmektedir. Göğüs kanseri, erken evrelerde tespit edildiğinde tedavi edilebilmekte ve bu sayede göğüs kanseri hastalığına bağlı ölümler önlenebilmektedir. Son yıllarda, makine öğrenmesi yöntemleri ile göğüs kanseri teşhisinde, geleneksel yöntemlerle hastalığın tespitine göre daha yüksek performans elde edilmektedir. Öte yandan, Topluluk Öğrenme yöntemlerinin klasik makine öğrenme yöntemlerinin performansını artırdığı bilinmektedir. Bu çalışmada, topluluk öğrenmesine dayalı göğüs kanseri hastalığı tespit sistemi önerilmiştir. Tezde; Rassal Orman, Oylama, Torbalama, Yükseltme, Yığma ve Ekstra Ağaçlar Topluluk Öğrenme yöntemleri uygulanarak hastalığın teşhisinde kullanılan klasik sınıflandırma yöntemlerinin (Destek Vektör Makinesi, K-En Yakın Komşu, Saf Bayes ve Karar Ağacı algoritmaları) sınıflandırma performanslarını artırdığı gösterilmiştir. Ayrıca, kullanılan Topluluk Yöntemlerinin karşılaştırmalı performans analizleri yapılmıştır. Modellerin eğitim ve test aşaması için Wisconsin Diagnostic Breast Cancer veri kümesi kullanılmıştır. En yüksek doğruluk, Stacking topluluk yöntemi ile %95.80 oranında elde edilmiştir.

Mayıs 2021, 79 sayfa.

Anahtar kelimeler: Göğüs kanseri, makine öğrenmesi, topluluk öğrenmesi

SUMMARY

BREAST CANCER DIAGNOSIS USING ENSEMBLE MACHINE LEARNING METHODS

M.Sc. THESIS

Fulya AKCAN

Istanbul University-Cerrahpasa

Institute of Graduate Studies

Department of Computer Engineering

Supervisor : Prof. Dr. Ahmet SERTBAŞ

Today, many people are affected by breast cancer disease. Breast cancer can be treated when it is detected in the early stages, and as a result deaths due to breast cancer can be prevented. In recent years, higher performance has been achieved by using machine learning methods compared to traditional methods in the diagnosis of breast cancer disease. On the other hand, it is known that Ensemble Learning methods increase the performance of classical machine learning methods. In this study, a breast cancer disease detection system is proposed based on ensemble learning. In the thesis; it has been shown that classical classification methods' (Support Vector Machine, K-Nearest Neighbor, Naive Bayes and Decision Tree algorithms) classification performances which have been used in the diagnosis of the disease are increased by applying Random Forest, Voting, Bagging, Boosting, Stacking and Extra Trees ensemble learning methods. In addition, comparative performance analysis of the ensemble methods was made. Wisconsin Diagnostic Breast Cancer dataset was used for the training and testing phase of the models. Highest accuracy was achieved with a rate of 95.80 % by using Stacking ensemble method.

May 2021, 79 pages.

Keywords: Breast cancer, machine learning, ensemble learning

1. GİRİŞ

Göğüs kanseri hem kadın hem de erkeklerde görülebilen bir kanser türüdür. Ancak, kadınlarda rastlanan en yaygın hastalıklardan biridir. Göğüs kanseri büyük ölçüde 40 yaş ve üstü kadınlarda görülmektedir. Göğüs kanseri, en yaygın görülen kanser türleri arasında ikinci sırada yer almaktadır (Purnami vd., 2008). 2020 yılında, dünya çapında 2.3 milyon kadına göğüs kanseri tanısı konulmuş ve 685.000 göğüs kanseri kaynaklı ölüm gerçekleşmiştir (Sung vd., 2021).

Göğüs kanseri, erken evrelerde tespit edildiğinde tedavi edilebilmekte ve bu sayede göğüs kanseri hastalığına bağlı ölümler önlenebilmektedir. Yeni tıbbi teknolojilerin ortaya çıkışı ve büyük miktardaki hasta verisi, kanserin tespitinde yeni stratejilerin geliştirilmesine katkı sağlamıştır. Hekimlerin yaptığı değerlendirmelere yardımcı olmak için destekleyici araçlar kullanılabilmektedir. Göğüs kanseri teşhisinde kullanılan geleneksel yöntemlerin sınırlamalarından kaçınmak, hekimlere ikinci bir görüş sağlamak ve olası tanı hatalarını azaltmak için Makine Öğrenmesi yöntemleri kullanılmaktadır (Abdar vd., 2020). Makine öğrenmesi yöntemleri, hastalığın tanı sürecini otomatikleştirmeye yardımcı olmaktadır. Makine öğrenmesi yöntemleri ile hastalığın tespitinde, geleneksel yöntemlerle hastalığın tespitine göre daha yüksek doğruluk oranı elde edilmektedir.

Karar Destek Makineleri, K-En Yakın Komşuluk, Lojistik Regresyon, Karar Ağacı, Rassal Orman gibi birçok farklı makine öğrenmesi yöntemi, makine öğrenmesi yöntemleriyle göğüs kanseri teşhisi alanındaki önceki çalışmalarda kullanılmıştır. Göğüs kanseri teşhisinde tek bir makine öğrenmesi yönteminin kullanılması ile her zaman yüksek doğruluk elde edilememektedir. Her bir yöntemin avantajları ve dezavantajları vardır. Topluluk öğrenmesi, bireysel sınıflandırıcıların gücünden yararlanmak ve zayıflıklarını azaltmak için kullanılmaktadır. Topluluk sınıflandırma yöntemleri, birden fazla bireysel sınıflandırıcının belirli bir kombinasyon kuralı altında birleştirilmesi ile oluşmaktadır. Topluluk öğrenmesi yöntemleri; daha kısa sürede, daha doğru bir teşhis ve tedaviye ulaşmaya katkı sağlamaktadır.

1.1. AMAÇ

Bu çalışmanın amacı, göğüs kanseri teşhisinde kullanılan farklı makine öğrenmesi yöntemlerini analiz etmek ve bireysel sınıflandırıcıların doğruluk oranını arttıran topluluk yöntemlerini kullanarak göğüs kanserinin teşhisi alanında daha iyi sonuçlar üretmektir. Geleneksel yöntemlere göre daha doğru ve daha hızlı bir tahmin yapılarak, hastalığın erken teşhisi ve bunun sonucunda göğüs kanseri hastalarının ölümlerini azaltmaya yardımcı olabilecek karşılaştırmalı bir çalışma oluşturmak hedeflenmiştir.

1.2. TEZİN ORGANİZASYONU

Göğüs kanseri teşhisinde kullanılan bireysel sınıflandırıcıların doğruluk oranını artırmak amacıyla kullanılan topluluk yöntemlerini incelemek için yazılan bu tez çalışması, 5 bölümden oluşmaktadır:

Bölüm 1'de, makine öğrenmesi teknikleriyle göğüs kanseri teşhisi alanındaki temel bilgiler ve çalışmanın amacı yer almaktadır.

Bölüm 2'de, literatürdeki makine öğrenmesi yöntemleriyle göğüs kanseri teşhisi çalışmaları incelenmiştir.

Bölüm 3'te, çalışmada uygulanan sistemin çalışma mekanizması yer almaktadır. Bu çalışmada kullanılan bireysel sınıflandırma yöntemleri, topluluk yöntemleri, veri kümesi ve model başarım ölçütleri açıklanmıştır.

Bölüm 4'te, çalışmada uygulanan bireysel ve topluluk yöntemlerinin performans ölçütlerinden elde edilen sonuçlar yer almaktadır. Çalışmada kullanılan yöntemler, performans skorlarına göre birbirleriyle karşılaştırılmıştır.

Bölüm 5'te, çalışma özetlenmiş ve çalışmanın sonuçları tartışılmıştır.

2. GENEL KISIMLAR

2.1. GÖĞÜS KANSERİ

Günümüzde birçok insan göğüs kanserinden etkilenmektedir. Göğüs kanseri; akciğer kanseri, prostat kanseri, kolon kanseri ve pankreas kanseri ile birlikte en yaygın kanser türlerinden biridir (National Cancer Institute, 2017).

Göğüs kanseri yaygın olarak, lobül hücreleri olarak adlandırılan süt bezlerinde veya göğüs ucuna sütü boşaltan kanallarda gelişmektedir. Göğüs kanseri, lobül hücrelerinde anormallik olması ve hücrelerin hızlı bir şekilde bölünmesi durumunda ortaya çıkmaktadır. Kanser, nadiren göğsün yağlı ve lifli bağ dokularını içeren stromal dokularda da gelişebilmektedir. Tümör, iyi huylu(malign) ve kötü huylu(benign) olmak üzere iki grupta incelenmektedir. İyi huylu bir tümörün, hastaların ölümüne yol açma olasılığı düşüktür. İyi huylu tümör, vücudun yalnızca bir bölümünde büyüyebilmektedir. Kötü huylu bir tümörün, hastaların ölümüne yol açma olasılığı yüksektir. Kötü huylu tümör, hızlı bir şekilde başka organlara da yayılabilmektedir.

Kanserli bir hücre, sağlıklı hücrelerden daha büyük veya daha küçük boyutta olup normal görünümünü kaybetmiştir. Sağlıklı hücreler, eşit şekillere ve yapılara sahiptir. Ancak kanser hücreleri, düzensiz şekillere sahip hücrelerdir. Bir kanser hücresinin çekirdeği, genellikle normal bir boyut veya şekle sahip değildir. Sağlıklı hücre çekirdeğinin aksine kanserli hücre çekirdeğinin, hücre merkezinde konumlanma olasılığı düşüktür.

2.1.1. Göğüs Kanseri Risk Faktörleri

Göğüs kanserine hem genetik hem de çevresel risk faktörleri yol açmaktadır. Göğüs kanserine yol açan genetik risk faktörleri aşağıda verilmiştir:

1. Yaş: Göğüs kanserine yakalanma riski, yaşla birlikte artmaktadır.
2. İlk çocuk doğurma yaşı: Bir kadının, ilk çocuğunu 30 yaşında veya daha sonrasında doğurması durumunda göğüs kanseri riskinin arttığı gösterilmiştir (Takkar vd., 2017).

3. Hastalık geçmişi: Hastanın daha önce göğüs kanserine yakalanmış olması, hastada yeniden göğüs kanseri gelişimi bakımından önemli bir risk faktörü oluşturmaktadır.
4. Aile öyküsü: Bir kadının ailesinde göğüs kanseri öyküsü varsa, kadının göğüs kanseri olma riski yüksektir (Pharoah vd., 1997).

Göğüs kanserine yol açan çevresel risk faktörleri aşağıda verilmiştir:

1. Alkol tüketimi: Alkol tüketimi, göğüs kanseri riskini artırmaktadır.
2. Fiziksel aktivite: Düzenli olarak yapılan fiziksel aktiviteler, göğüs kanseri riskini azaltmaktadır.
3. Obezite: Kadınlarda menopoz dönemi sonrasında, yüksek bir vücut kitle indeksi değeri göğüs kanseri riskini artırmaktadır (Takkar vd., 2017).
4. Radyasyon: Radyasyona maruz kalmak, göğüs kanseri riskini artırmaktadır.

2.1.2. Göğüs Kanseri Belirtileri

Göğüs kanseri hastalığının erken evrelerinde, bir tümör elle hissedilemeyecek kadar küçük olduğu için çoğu zaman herhangi bir belirti göstermemektedir. Ancak herhangi bir belirtiye rastlanmamasına rağmen mamografide bir anormallik görülebilmektedir.

Göğüs kanserinde görülen yaygın belirtiler şunlardır:

1. Göğüste veya koltuk altında kalınlaşma hissi veren, geçmeyen bir yumru görülebilir. Yumru, genellikle göğüs kanserinin ilk belirtisidir.
2. Göğüs boyutunda veya şeklinde bir değişiklik meydana gelebilir.
3. Aniden oluşan ve sadece bir göğüste oluşan göğüs ucu akıntısı görülebilir.
4. Göğüs ucunun içe doğru dönmesi veya göğüs ucu bölgesinde yara oluşması vb. fiziksel değişiklikler görülebilir.
5. Koltuk altında veya köprücük kemiğinin yakınında şişme oluşabilir. Bu, göğüs kanserinin o bölgedeki lenf düğümlerine yayıldığı anlamına gelebilmektedir.
6. Göğüste geçmeyen ağrı veya hassasiyet hissedilebilir.
7. Göğüste çukurlu bir alan oluşabilir. Göğüste çukurlaşma, görülemeyen veya hissedilemeyen bir tümör nedeniyle olabilmektedir.
8. Göğüste veya göğüs ucunda; döküntü, soyulma, kızarıklık veya sıcaklık değişimi olabilir.

9. Göğüste veya göğüs ucunda; kaşıntı veya yanma hissi oluşabilir.

2.1.3. Göğüs Kanserinde Evrelendirme Sistemleri

Göğüs kanseri; tümörün veya tümörlerin ne kadar büyük olduğuna, göğüs kanserinin ne kadar yayıldığına, lenf düğümlerine yayılıp yayılmadığına, tümörlü bölge yakınındaki dokulara veya organlara yayılıp yayılmadığına bağlı olarak evrelere ayrılabilmektedir.

2.1.3.1. TNM Evrelendirme Sistemi

TNM evrelendirme sistemi, kanserin yayılma derecesini sınıflandırmak için dünya çapında tanınan bir standarttır. TNM evrelendirme sisteminde evre numarasını hesaplamak için gerekli olan Tümör(T), Lenf Düğümü (N) ve Metastaz (M) klinik özellikleri aşağıda yer almaktadır (Singletary vd., 2002):

- Tümör klinik özelliği, tümörün boyutu ile ilgilidir.
- Lenf Düğümü klinik özelliği; tümörün, lenf düğümlerine yayılıp yayılmadığı ile ilgilidir.
- Metastaz klinik özelliği; tümörün, vücudun göğüs dışındaki diğer bölgelerine yayılıp yayılmadığı ile ilgilidir.

TNM evrelendirme sisteminde; tümörün boyutunu ve konumunu açıklamak için "T" harfinin yanına, başka bir harf veya bir sayı (0-4 arasında) eklenmektedir. Tümör boyutu, cm cinsinden ölçülmektedir.

Tümör evrelendirme sistemi aşağıda yer almaktadır:

- TX: Birincil tümör değerlendirilemez.
- T0: Göğüste, tümöre dair bir bulguya rastlanmamıştır.
- Tis: Tümör, göğüs dokusunun kanalları içinde olup ve göğsü çevreleyen dokuya yayılmamıştır. 2 tip Tis vardır:
    - Tis (DCIS): DCIS, istilacı olmayan bir kanserdir, ancak çıkarılmazsa daha sonra istilacı bir göğüs kanserine dönüşebilir. DCIS, kanser hücrelerinin göğüs kanallarında bulunduğu ve henüz yayılmaya başlamadıkları anlamına gelmektedir.

    - Tis (Paget's): Memenin Paget hastalığı, yalnızca göğüs ucunun hücrelerinde ortaya çıkan, genellikle istilacı olmayan bir kanser biçimidir.
- T1: Tümörün, 2 cm veya daha küçük olduğu anlamına gelmektedir. T1 evresi, tümörün boyutuna bağlı olarak 4 alt evreye ayrılmaktadır:
    - T1mi: Tümörün, 0,1 cm veya daha küçük olduğu anlamına gelmektedir.
    - T1a: Tümörün, 0,1 cm'den büyük olduğu ancak 0,5 cm'den büyük olmadığı anlamına gelmektedir.
    - T1b: Tümörün, 0,5 cm'den büyük olduğu ancak 1 cm'den büyük olmadığı anlamına gelmektedir.
    - T1c: Tümörün, 1 cm'den büyük olduğu ancak 2 cm'den büyük olmadığı anlamına gelmektedir.
- T2: Tümörün, 2 cm'den büyük olduğu ancak 5 cm'den büyük olmadığı anlamına gelmektedir.
- T3: Tümörün, 5 cm'den daha büyük olduğu anlamına gelmektedir.
- T4: T4 evresi, 4 alt evreye ayrılmaktadır:
    - T4a: Tümörün, göğüs duvarına yayıldığı anlamına gelmektedir.
    - T4b: Tümörün, cilde yayıldığı ve göğsün şişmiş olabileceği anlamına gelmektedir.
    - T4c: Tümörün hem cilde hem de göğüs duvarına yayıldığı anlamına gelmektedir.
    - T4d: T4d, iltihaplı göğüs kanseri anlamına gelmektedir. Tümörün üzerini çevreleyen deri; kızarık, şiş ve ağrılı olabilir.

Lenf düğümleri, bulundukları bölgeye göre aşağıdaki gibi sınıflandırılmaktadır:

- Aksiller lenf düğümleri olarak adlandırılan, koltuk altında bulunan lenf düğümleri
- Köprücük kemiğinin üstünde ve altında bulunan lenf düğümleri
- Göğüs kemiğinin arkasında bulunan ve internal mammarian lenf düğümleri adlandırılan lenf düğümleri
- Uzak lenf düğümleri olarak adlandırılan, vücudun diğer bölgelerinde bulunan lenf düğümleri

Lenf düğümlerini, TNM evrelendirme sisteminde "N" harfi temsil etmektedir. Lenf düğümleri evrelendirme sistemi aşağıda yer almaktadır:

- NX: Lenf düğümlerinin değerlendirilemeyeceği anlamına gelmektedir.
- N0, aşağıdaki koşullardan birini sağlamaktadır:
    - Lenf düğümlerinde kanser bulunmamaktadır.
    - Lenf düğümlerinde, sadece 0,2 mm'den küçük kanser alanları bulunmaktadır.
- N1: 1-3 adet aksiller lenf düğümünde ve / veya internal mammarian lenf düğümlerinde tutulum görülmektedir. Lenf düğümündeki kanser; 0,2 mm'den büyükse ancak 2 mm'den daha büyük değilse, bu evre "N1mi" olarak adlandırılmaktadır.
- N2, 2 gruba ayrılmıştır:
    - N2a: 4-9 adet aksiller lenf düğümünde kanser hücreleri olduğu ve bu kanser hücrelerinden en az birinin 2 mm'den büyük olduğu anlamına gelmektedir.
    - N2b: Internal mammarian lenf düğümlerinde tutulum görüldüğü ancak aksiller lenf düğümlerinde kansere dair bir bulgu bulunmadığı anlamına gelmektedir.
- N3, üç grupta incelenmektedir:
    - N3a: Lenf düğümleri, aşağıdaki şartlardan birini sağlamaktadır:
        - ❖ 10 veya daha fazla aksiller lenf düğümünde kanser hücreleri olduğu ve bu kanser hücrelerinden en az birinin 2 mm'den büyük olduğu anlamına gelmektedir.
        - ❖ Köprücük kemiğinin altındaki düğümlerde kanser hücreleri olduğu anlamına gelmektedir.
    - N3b: Aksiller ve internal mammarian lenf düğümlerinde kanser hücreleri olduğu anlamına gelmektedir.
    - N3c: Köprücük kemiğinin üzerindeki lenf düğümlerinde kanser hücreleri olduğu anlamına gelmektedir.

TNM evrelendirme sisteminde; uzak metastaz adı verilen, kanserin vücudun diğer bölgelerine yayılıp yayılmama durumunu tanımlamak için "M" harfi kullanılmaktadır.

- MX: Kanserin, vücudun diğer bölgelerine yayılma durumu değerlendirilemez.
- M0: Uzak metastazlara dair hiçbir bulguya rastlanmadığı anlamına gelmektedir.

- M0(i+): Uzak metastazlara dair klinik veya radyografik kanıt bulunamadığı anlamına gelmektedir. Bununla birlikte kanda, kemik iliğinde veya diğer bölgesel olmayan lenf düğümlerinde tümör hücrelerine rastlanmıştır.

- M1: Kanserin, vücudun başka bir bölümüne yayıldığına dair bir bulguya rastlandığı anlamına gelmektedir.

2.1.3.2. Genel Evrelendirme Sistemi

Tümörün genel evresini tanımlarken; TNM sisteminde kullanılan Tümör, Lenf Düğümü ve Metastaz klinik özelliklerinin kombinasyonundan yararlanılmaktadır. Göğüs kanseri; tümörün büyüklüğüne, tümörün yakındaki doku veya organlara yayılıp yayılmamasına bağlı olarak evrelere ayrılmaktadır.

Göğüs kanseri, beş ana evrede incelenmektedir:

- Evre 0: İstilacı olmayan göğüs kanserlerini tanımlamak için kullanılmaktadır (Tis, N0, M0).
- Evre 1A: Birincil tümör, 2 cm veya daha küçük büyüklüktedir. Lenf düğümlerinde tutulum görülmemiştir (T1, N0, M0).
- Evre 1B: Kanser, lenf düğümlerine yayılmıştır. Lenf düğümündeki tümör, 0.2 mm- 2 mm büyüklüğündedir. Evre 1B, aşağıdaki koşullardan birini sağlamaktadır:
  - ❖ Göğüste tümör olduğuna dair bir bulguya rastlanmamıştır (T0, N1mi, M0).
  - ❖ Göğüsteki tümör, 2 cm'den büyük değildir (T1, N1mi, M0).
- Evre 2A, aşağıdaki koşullardan herhangi birini sağlamaktadır:
  - Göğüste tümör olduğuna dair herhangi bir bulguya rastlanmamıştır. 1-3 adet aksiller lenf düğümünde tutulum görülmektedir. Kanser, vücudun uzak bölgelerine yayılmamıştır (T0, N1, M0).
  - Tümör, 2 cm'den daha büyük değildir. Kanser, 1-3 adet aksiller lenf düğümüne yayılmıştır. Vücudun uzak bölgelerine yayılmamıştır (T1, N1, M0).
  - Tümör, 2-5 cm büyüklüğündedir. Kanser, aksiller lenf düğümlerine ve vücudun uzak bölgelerine yayılmamıştır (T2, N0, M0).
- Evre 2B, aşağıdaki koşullardan herhangi birini sağlamaktadır:
  - Tümör, 2-5 cm büyüklüğündedir. Kanser, 1-3 adet aksiller lenf düğümüne yayılmıştır. Vücudun uzak bölgelerine yayılmamıştır (T2, N1, M0).

- Tümör, 5 cm'den büyüktür. Kanser, aksiller lenf düğümlerine ve vücudun uzak bölgelerine yayılmamıştır (T3, N0, M0).

- Evre 3A, aşağıdaki koşullardan herhangi birini sağlamaktadır:
  - Herhangi bir boyuttaki tümör, 4-9 adet aksiller lenf düğümüne veya internal mammarian lenf düğümlerine yayılmıştır. Vücudun uzak bölgelerine yayılmamıştır (T0, N2, M0 veya T1, N2, M0 veya T2, N2, M0 veya T3, N2, M0).
  - Tümör, 5 cm'den büyüktür. Kanser, 1-3 adet aksiller lenf düğümüne yayılmıştır. Vücudun uzak bölgelerine yayılmamıştır (T3, N1, M0).
- Evre 3B aşağıdaki koşullardan herhangi birini sağlamaktadır:
  - Tümör, göğüs duvarını veya cildi istila etmiştir. Lenf düğümlerinde kanser bulunmamakta veya sadece 0,2 mm'den küçük kanser alanları bulunmaktadır. Vücudun uzak bölgelerine yayılmamıştır (T4, N0, M0).
  - Tümör, göğüs duvarını veya cildi istila etmiştir. 1-3 adet aksiller lenf düğümünde ve / veya internal mammarian lenf düğümlerinde tutulum görülmektedir. Vücudun uzak bölgelerine yayılmamıştır (T4, N1, M0).
  - N2a: Tümör, göğüs duvarını veya cildi istila etmiştir. 4-9 adet aksiller veya internal mammarian lenf düğümünde kanser hücreleri bulunmaktadır. Vücudun uzak bölgelerine yayılmamıştır (T4, N2, M0).
- Evre 3C: Göğüste, tümöre dair bir bulguya rastlanmamış olabilir veya tümör varsa herhangi bir boyutta olabilir. Kanser, 10 adet veya daha fazla aksiller, internal mammarian veya köprücük kemiği yakınındaki lenf düğümlerine yayılmıştır. Vücudun uzak bölgelerine yayılmamıştır (herhangi bir T, N3, M0).
- Evre IV (Metastatik): Tümör, herhangi bir boyutta olabilir. Herhangi bir sayıda aksiller, internal mammarian veya köprücük kemiği yakınındaki lenf düğümlerine yayılmış olabilir. Vücudun başka bir bölümüne yayıldığına dair bir bulguya rastlanmıştır. (herhangi bir T, herhangi bir N, M1).

2.1.4. Göğüs Kanseri Tanı Yöntemleri

Erken teşhis, ölüme yol açabilen göğüs kanserinin gelişimini önlemenin en önemli yollarından biridir. Göğüs kanseri taraması; bireyde göğüs kanseri oluşmasını önlememekle

birlikte, hastalığın erken teşhis edilmesine yardımcı olmakta ve bu sayede göğüs kanserinden kaynaklanan ölüm oranlarının azalmasına katkıda bulunmaktadır.

Klinik göğüs kanseri muayenesi; göğüs kanseri semptomlarını veya diğer anormallikleri tespit etmek için bir sağlık uzmanının, hastanın göğsünü fiziksel olarak inceleyerek yaptığı bir tarama yöntemidir.

Kendi kendine meme muayenesi; kadınların göğüs kanserini tespit etmek için kendi kendilerine, bir sağlık uzmanı yardımı olmadan uyguladıkları bir tarama yöntemidir. Bu tarama yöntemi; kişinin olası yumrular, bozulmalar, şişlikler için her iki göğsü de bakarak incelemesini ve her bir göğsü, göğsün yaklaşık bir inç aşağısından köprücük kemiğine kadar elle muayene etmeyi içermektedir (Senie vd., 1981).

Mamografi, kanser ve diğer göğüs hastalıklarının erken teşhisi için göğsü incelemede kullanılan bir röntgen görüntüleme yöntemidir. Kanser hücresinin, genellikle dışarıdan görülemeyecek kadar küçük olması nedeniyle göğüs kanserini ilk evrede tespit etmek neredeyse imkansızdır (Islam vd., 2020). Kanseri, erken evrede mamografi testi ile teşhis etmek mümkün olmaktadır ve mamografi testini yaptırmak yalnızca birkaç dakika sürmektedir. Göğüs kanseri, 40 ve üzeri yaşlardaki kadınlar için erken ölümlerin önde gelen nedenlerinden biridir. Mamografik tarama, göğüs kanserine bağlı ölüm riskini önemli ölçüde azaltmaktadır (Berg, 2010).

Ultrason, göğüs kanserini ve diğer göğüs anormalliklerini taramak için yaygın olarak kullanılan bir görüntüleme tekniğidir. Göğüslerin içerdeki durumunu gözlemlemek için, yüksek frekanslı ses dalgaları kullanarak göğsün ayrıntılı görüntülerini üretmektedir.

Manyetik rezonans görüntüleme (MRI); göğüs kanseri teşhisinde, göğsün ayrıntılı görüntülerini üretmek için manyetizma, radyo dalgaları ve bilgisayarın kombinasyonunu kullanmaktadır. Göğsün, dört bir yanındaki açılardan görüntüler üretilmektedir. Manyetik rezonans görüntüleme yöntemiyle, yumuşak dokular net bir biçimde gösterilmektedir. MRI yöntemi ile mamografide gözden kaçabilen bazı küçük meme lezyonları görülebilmektedir.

Biyopsi; göğüste bir yumru, bozukluk veya mamografide anormal bir bulgu görüldüğünde uygulanan bir tekniktir. Göğsün, tümör olduğu şüphelenilen bölgesinden hücreler veya doku

çıkarılarak mikroskop altında incelenmektedir. Aşağıda, göğüs kanseri teşhisinde uygulanan birkaç farklı göğüs biyopsisi yöntemine yer verilmiştir:

- İnce iğne aspirasyon biyopsisi (FNA); göğüs kanserinin varlığından şüphelenilen bir alandan, küçük miktarda doku almak için çok ince bir iğnenin kullanıldığı bir biyopsi türüdür. FNA biyopsisi için kullanılan iğne, kan testleri için kullanılan iğnelerden daha ince bir boyuttadır.
- Çekirdek iğne biyopsisi; genellikle ultrason, mamogram veya MRI görüntüleri rehberliğinde yapılmaktadır (Łukasiewicz vd., 2017). Bu yöntem uygulanırken, ince iğne aspirasyon biyopsisine göre daha büyük bir iğne kullanılmaktadır.
- Cerrahi biyopsi; lezyonun çoğunu veya tamamını test etmek için ameliyatla çıkarma işlemi yapmaktadır. Cerrahi biyopsi, çekirdek iğne biyopsisine göre daha yüksek kanama ve enfeksiyon riski taşımaktadır (Bernardi vd., 2012).

2.1.5. Göğüs Kanserinde Tedavi Yöntemleri

Göğüs kanserinde tedavi seçenekleri hastaya göre değişkenlik göstermektedir. Tedavi şekli; tümörün evresi, hastanın genel sağlık ve menopoz durumu gibi çeşitli faktörlere bağlıdır.

Göğüs kanserinin teşhis ve tedavisinde; kanser tedavisinin farklı alanlarında uzmanlaşmış doktorlar, radyologlar ve patologların dahil olduğu multidisipliner bir ekip ortak bir çalışma gerçekleştirmektedir. Göğüs kanseri için yaygın olarak kullanılan tedavi türleri; ameliyat, radyasyon ve ilaçla tedavidir. Kanserin evresine bağlı olarak, hastaya bu tedavilerden biri veya birkaçını içeren bir tedavi planı oluşturulur.

2.1.5.1. Ameliyat

Ameliyat, hastalığı tedavi etmek için tümörün ve çevresindeki bazı sağlıklı dokuların çıkarıldığı bir yöntemdir. Göğüs kanseri tedavisi için uygulanan lumpektomi ve mastektomi ameliyat türleri aşağıda açıklanmıştır:

- Lumpektomi: Bu ameliyat türünde, tümör ve tümörün etrafındaki az miktarda sağlıklı doku çıkarılmaktadır. Göğsün büyük bir kısmı, yerinde kalmaktadır. Lumpektomi ameliyatında çıkarılması gereken göğüs dokusu, tümörün boyutu ve konumuna bağlı olarak belirlenmektedir (University of Rochester Medical Center).

- Mastektomi: Bu ameliyat türünde, kanserli göğsün tümü ameliyatla alınmaktadır. Mastektomi ameliyatını, hastanın sağlıklı göğsüne olabildiğince çok benzeyen bir göğüs şekli oluşturmayı sağlayan rekonstrüktif cerrahi takip edebilmektedir.

2.1.5.2. Radyasyon

Radyasyon, kanser hücrelerini yok etmek için yüksek enerjili ışınların kullanıldığı bir tedavi yöntemidir. Radyasyon tedavisinde kullanılan yöntemler şunlardır:

- Dış ışın radyasyon tedavisi: Bu radyasyon tedavisi türü, göğüs kanseri tedavisinde en yaygın kullanılan radyasyon tedavisi türüdür. Radyasyon, büyük bir makine yoluyla vücuda dışarıdan verilmektedir.
- Intraoperatif radyasyon tedavisi: Ameliyathanede; ameliyat sırasında tek seans X ışınının, sadece tümörlü bölgeye verilmesiyle gerçekleştirilen bir radyasyon tedavisi yöntemidir.
- Brakiterapi: Ameliyat sırasında veya ameliyattan kısa bir süre sonra, radyoaktif kaynakların tümör yatağına yerleştirilerek verildiği bir radyasyon tedavisi türüdür (National Breast Cancer Foundation).

2.1.5.3. İlaç Tedavisi

Göğüs kanseri tedavisinde, kanser hücrelerini yok etmek için ilaçların kullanıldığı tedavi türleri şunlardır:

- Kemoterapi
- Hormonal terapi
- Hedefe yönelik ilaç terapisi

Hastaya bu tedavilerden biri veya birkaçının kombinasyonunu içeren bir tedavi planı oluşturulmaktadır. Cerrahi ve/veya radyasyon tedavi türünü içeren bir tedavi planının parçası olarak da uygulanabilmektedir.

Kemoterapi; kanser hücrelerini yok etmek veya kanser hücrelerinin büyümesini yavaşlatmak için ilaçların kullanıldığı bir tedavi türüdür. Neoadjuvan kemoterapi; büyük bir tümörü küçültmek ve bunun neticesinde yapılacak ameliyatı kolaylaştırmak için ameliyattan önce uygulanan bir kemoterapi türüdür. Adjuvan kemoterapi; çıkarılmamış kanser hücrelerini yok etmek veya nüks riskini azaltmak için ameliyattan sonra uygulanan bir kemoterapi türüdür

(Cancer Research UK, 2020). Kemoterapide kullanılacak ilaç seçimi ve farklı ilaçların kombinasyonu; hastadaki tümörün türüne, büyüklüğüne ve evresine bağlı olarak belirlenmektedir. Kemoterapinin yan etkileri; hastaya, kullanılan ilaçlara, kullanılan ilaçların dozuna bağlı olarak değişmektedir. Yan etkiler arasında; yorgunluk, enfeksiyon riski, bulantı, iştah kaybı, saç dökülmesi, ishal ve kabızlık yer almaktadır.

Hormonal terapi; hormonların bloke edilmesini sağlayarak kanserin nüksetmesini önlemeye yardımcı olmaktadır. Neoadjuvan hormonal tedavi; bir tümörün boyutunu küçültmeye veya yayılan kanser hücrelerini öldürmeye yardımcı olmak için ameliyattan önce verilmektedir (Sikov vd., 2020). Adjuvan hormonal tedavi; kanserin nüksetme riskini azaltmak için ameliyattan sonra verilmektedir. Uygulanacak hormon tedavisinin türü; kanserin evresine, hastanın yaşına ve menopoz durumuna bağlı olarak belirlenmektedir.

Hedefe yönelik ilaç tedavisi; kanser hücrelerinin büyümesine katkıda bulunan proteinler ve genler gibi kanser hücrelerinin belirli kısımlarını hedefleyen bir tedavi türüdür (Smith, 2021). Bu tedavi türü; kanserli hücrelerin büyümesini ve yayılmasını engellerken, sağlıklı hücrelerin korunabilmesine yardımcı olmaktadır.

2.1.6. Tekrarlayan Göğüs Kanseri

Tekrarlayan göğüs kanseri, tedaviden sonra geri gelen göğüs kanserini ifade etmektedir. Göğüs kanserinin tekrarlanması riski, hastaya uygulanan tedaviyi izleyen ilk iki yıl en yüksek orandadır (John Hopkins Medicine). Tekrarlayan göğüs kanseri, nüksettikleri bölgelere göre 3 farklı tipte incelenmektedir:

- Yerel Nüks: Kanserin, göğüste tümörün ilk rastlandığı bölgede tekrar ortaya çıkarak tekrarlama biçimine yerel nüks denilmektedir.
- Bölgesel Nüks: Kanserin, göğüste tümörün ilk rastlandığı bölgenin yakınında bulunan koltuk altındaki lenf düğümlerinde veya göğüs duvarında ortaya çıkarak tekrarlama biçimine bölgesel nüks denilmektedir.
- Metastatik Nüks: Kanserin, vücudun diğer bölgelerinde ortaya çıkarak tekrarlama biçimine metastatik nüks denilmektedir. En yaygın görülen göğüs kanseri metastaz bölgeleri; kemikler, akciğerler, beyin ve karaciğerdir.

2.2. MAKİNE ÖĞRENMESİ

Makine öğrenmesi; sistemlerin açıkça programlanmadan, verilerden otomatik olarak öğrenebilmesini ve kendi kendini iyileştirilebilmesini sağlayan bir yapay zeka alt dalıdır. Makine öğrenmesi algoritmaları, yeni verilerin beslenmesiyle daha doğru tahminler yapmaktadırlar. Makine öğrenmesi; bilgi teknolojileri, istatistik, psikoloji, nörobiyoloji ve diğer birçok bilim dalında kullanılmaktadır. Makine öğrenmesi algoritmaları, makinenin öğretilme şekline göre 4 farklı kategoride gruplandırılmaktadır:

Gözetimli öğrenme; makinenin etiketlenmiş veriler kullanılarak eğitildiği bir öğrenme biçimidir. Yeni girdi verilerinin çıktı değerleri hakkında tahminde bulunmak için eğitim veri kümesinin kullanılmaktadır. Makineyi eğitmek için ilk önce etiketli bir eğitim veri kümesi kullanılmaktadır. Daha sonra yeni girdi verilerini benzer gruplara ayırmak için; eğitilmiş algoritma, etiketsiz test veri kümesine beslenmektedir. Gözetimli öğrenme, iki kategoriye ayrılmaktadır:

- Sınıflandırma: Girdi değişkenlerine dayalı ayrık bir çıktı değeri tahmin etmektedir. Sınıflandırma algoritmaları, çıktı değişkeni “kadın”, “erkek” gibi bir etiket veya kategori olduğunda kullanılmaktadır. Lojistik regresyon, K-En Yakın Komşu, Destek Vektör Makineleri, Saf Bayes, Karar Ağacı ve Rassal Orman, sınıflandırma algoritması türlerindendir.
- Regresyon: Girdi değişkenlerine dayalı sürekli bir çıktı değeri tahmin etmektedir. Regresyon algoritmaları gelir, puan, boy, kilo gibi gerçek değerleri tahmin etmek için kullanılmaktadır. Basit Doğrusal Regresyon, Çok Değişkenli Doğrusal Regresyon, Polinomsal Regresyon ve Destek Vektör Regresyonu, regresyon algoritması türlerindendir.

Yarı-gözetimli öğrenme, gözetimli ve gözetimsiz öğrenmenin bir kombinasyonudur. Eğitim için hem etiketlenmiş hem de etiketlenmemiş verilerin kullanıldığı bir öğrenme biçimidir. Yarı-gözetimli öğrenme kullanarak hem gözetimli ve gözetimsiz öğrenmenin avantajlarından yararlanılmakta, hem de büyük miktarda veri etiketlemenin yüksek maliyetiyle karşılaşılmamaktadır.

Gözetimsiz öğrenme, eğitim için etiketlenmemiş veya sınıflandırılmamış verilerin kullanıldığı bir öğrenme biçimidir. Gözetimsiz öğrenmede makinenin görevi; etiketlenmemiş verilerdeki gizli yapıyı kendi kendine bulmak için veri kümelerinden çıkarımlar yapmaktır. Kümeleme algoritmaları, verilerdeki gizli kalıpları bulmak için kullanılan en yaygın gözetimsiz öğrenme algoritmalarıdır (El Bouchefry & de Souza, 2020). Kümeleme; veri noktalarını, aralarındaki benzerlik ve farklılıklara göre kategorize edip gruplara ayırmaktadır. K-means, Fuzzy C-Means ve Hiyerarşik Kümeleme, en yaygın kullanılan kümeleme algoritmalarıdır.

Pekiştirmeli öğrenmede; makine, eylemler üreterek çevre ile etkileşime girmekte ve hangi eylemlerin en büyük ödülü verdiğini keşfetmektedir. Pekiştirmeli öğrenme biçiminde; ajan (öğrenen), çevre ve eylemler olmak üzere 3 ana bileşen bulunmaktadır. Ajanın hangi eylemin en iyi olduğunu öğrenmesi için gerekli olan ödül geri bildirimi, takviye sinyali olarak adlandırılmaktadır.

2.3. TOPLULUK YÖNTEMLERİ

Topluluk öğrenmesi; birden fazla temel öğreniciyi topluluk üyeleri olarak eğitmek ve bu topluluk üyelerinin tahminlerini hedef veri setleri üzerinde, topluluk üyelerine göre daha yüksek doğruluk oranının elde edildiği tek bir çıktıda birleştirmeyi sağlamaktadır (Yang, 2016). Birden çok sınıflandırıcının tahminlerini bir araya getirerek sınıflandırma doğruluğunu iyileştirmeyi amaçlamaktadır. Topluluk öğrenmesi; temel öğrenicilerin güçlü ve zayıf yönlerini yöneterek, mümkün olan en iyi kararın alınmasını sağlamaktadır. Bir modelin performansını iyileştirmek, kötü bir seçim yapma olasılığını azaltmak veya her ikisini birden yapmak için kullanılmaktadır (Manikandan vd., 2019).

Çoğu topluluk yöntemi, homojen temel öğrenenler üretmek için tek bir temel öğrenme algoritması kullanmaktadır. Bununla birlikte, heterojen öğrenenler üretmek için birden fazla öğrenme algoritması kullanan topluluk yöntemleri de bulunmaktadır. Topluluk öğrenmesi; Torbalama, Yükseltme, Yığma, Ekstra Ağaçlar ve Oylama yöntemlerini içermektedir.

2.4. LİTERATÜR TARAMASI

Yeni teknolojilerin gelişmesi ile birlikte göğüs kanserinin tahmininde makine öğrenmesi yöntemleri uygulanmaya başlanmıştır. Makine öğrenmesi yöntemleri kullanılarak göğüs kanseri teşhisi alanında yapılan çalışmalardan bazıları bu bölümde incelenmiştir.

Abdar ve Makarenkov (2019), CWV (Confidence-Weighted Voting) ve Boosting topluluk öğrenme yöntemlerine dayalı yeni bir göğüs kanseri tespit yaklaşımı önermişlerdir. İlk olarak SVM algoritması; C, ε ve γ parametrelerinin çeşitli değerlerini kullanarak test edilmiştir. İlk deneyin bir sonucu olarak, düzenleme parametrelerinin ayarlanmasının, göğüs kanseri tespiti için uygulanan geleneksel SVM algoritmasının performansını büyük ölçüde artırabildiği gözlemlenmiştir. CWV-BANNSVM adı verilen model, ilk deney sırasında seçilen optimum parametreleri kullanarak Boosting Yapay Sinir Ağları (BANN) ve iki SVM'yi birleştirmiştir. Deneyler, Wisconsin Breast Cancer Dataset üzerinde yürütülmüştür. Uygulanan yöntemlerin performansı, özgüllük, duyarlılık, kesinlik, FPR, FNR, F1 Skoru, AUC, Gini ve doğruluk gibi model başarım ölçütleri kullanılarak değerlendirilmiştir. Önerilen CWV-BANNSVM modeli, göğüs kanseri tespiti için uygulanan geleneksel makine öğrenimi algoritmalarının performansını artırarak %100 doğruluk yüzdesine ulaşmıştır.

Moon ve diğ. (2020), farklı görüntü içeriği temsilleri ile kombine edilmiş bir görüntü füzyon yöntemi kullanılarak tümör teşhisi için bir CAD (Computer Aided Diagnosis) sistemi önerilmiştir. Çalışmada önerilen CNN (Convolutional Neural Network) tabanlı yöntem; VGGNet, ResNet ve DenseNet içermektedir. 953 iyi huylu ve 734 kötü huylu olmak üzere toplam 1687 tümör ultrason görüntüsünün yer aldığı, Seoul National University Hospital (SNUH) tarafından oluşturulan özel veri kümesi kullanılmıştır. SNUH veri kümesi ile önerilen yöntemin doğruluğu, duyarlılığı, özgüllüğü, kesinliği, F1 puanı ve AUC skoru sırasıyla %91.10, % 85.14, % 95.77, %94.03, %89.36 ve 0.9697 olarak elde edilmiştir. BUSI veri kümesi ile önerilen yöntemin doğruluğu, duyarlılığı, özgüllüğü, kesinliği, F1 puanı ve AUC skoru sırasıyla; %94.62, %92.31, %95.60, %90, %91.14 ve 0.9711 olarak ölçülmüştür.

Liu ve diğ. (2020), kenar tabanlı özellikler ve morfolojik özellik bilgilerini birleştiren yeni bir ultrasonik görüntü özelliği çıkarma algoritması önermişlerdir. SVM sınıflandırıcısının sonuçlarına göre, uç tabanlı özelliklerin daha yüksek sınıflandırma doğruluğuna sahip olduğu görülmüştür (%82.69). Veri kümesi için kullanılan görüntüler, ultrason teşhis cihazı ile 192

vakadan elde edilmiştir. Veri kümesi, 71 kötü huylu tümör ve 121 iyi huylu tümör görselinden oluşmaktadır.

Gu ve diğ. (2020), göğüs kanseri nüksü tahmini için CBR (Case-Based Reasoning) ve topluluk öğrenmesini birleştiren bir yöntem önermişlerdir. Önerilen sistemde, XGBoost sınıflandırıcısının tahmin sürecindeki her bir değişkenin öneminin bir ölçüsü olarak SHAP (Shapley Additive Explanations) değeri kullanılmıştır. İş birliği yapılan hastaneden; 84 pozitif ve 133 negatif vaka olmak üzere toplam 217 standart göğüs kanseri nüks tahmini vakası hakkında veri toplanmıştır. XGBoost sınıflandırıcısı ile; % 90.28 duyarlılık, %91.62 doğruluk ve %89.39 F1-skor elde edilmiştir.

Agarap (2018), göğüs kanseri teşhisinde GRU-SVM (Gated Recurrent Unit – SVM), Lojistik Regresyon, MLP (Multilayer Perceptron), En Yakın Komşu, Softmax Regresyon ve Destek Vektör Makinesi yöntemlerini karşılaştırmıştır. Deneyler, Wisconsin Diagnostic Breast Cancer veri kümesi üzerinde gerçekleştirilmiştir. %70 eğitim ve %30 test olmak üzere veri kümesi bölünmüştür. En yüksek doğruluk, MLP yöntemi ile elde edilmiştir (%99.04).

Omondiagbe ve diğ. (2019), göğüs kanseri teşhisi için Doğrusal Diskriminant Analizi (LDA) ile özelliklerin yüksek boyutluluğunu azaltıp ardından Destek Vektör Makinesi'ne indirgenmiş özellik veri kümesinin uygulandığı hibrit bir yaklaşım önermişlerdir. Eğitim ve test aşaması için Wisconsin Diagnostic Breast Cancer veri kümesini kullanılmıştır. Önerilen yaklaşım ile %98.82 doğruluk, %98.41 duyarlılık, %99.07 özgüllük ve 0,9994 AUC skoru elde edilmiştir.

Ahmad ve diğ. (2013), iki yıl boyunca izlenen hastalarda göğüs kanseri nüksü için öngörücü modeller geliştirmek amacıyla Karar Ağacı (C4.5), Destek Vektör Makinesi ve Yapay Sinir Ağları yöntemlerinin performansını karşılaştırmıştır. Kullanılan veri kümesi, Iranian Center for Breast Cancer'dan elde edilmiştir. Çalışma sonucunda Karar Ağacı, Yapay Sinir Ağları ve Destek Vektör Makinesi yöntemlerinin doğrulukları sırasıyla 0.936, 0.947 ve 0.957 olarak ölçülmüştür. SVM sınıflandırma modeli, göğüs kanseri nüksünü en az hata oranı ve en yüksek doğruluk ile öngörmüştür.

Nematzadeh ve diğ. (2015), göğüs kanserini sınıflandırmada Karar Ağacı, Saf Bayes, NN (Neural Network), SVM-Linear, SVM-MLP ve SVM-RBF (SVM- Radial Basis Function) yöntemleri ile karşılaştırmalı bir çalışma yürütmüştür. Deneyler için Wisconsin Prognostic

Breast Cancer (WPBC) ve Wisconsin Breast Cancer (WBC) veri kümesi kulllanılmıştır. Çalışma sonucunda; WBC veri kümesinde en yüksek doğruluk NN (10 kat) ile elde edilmiştir (%98.09). WPBC veri kümesinde en yüksek doğruluk SVM-RBF (10 kat) yöntemi ile elde edilmiştir (%98.32).

Ojha ve Goel (2017), tekrarlayan göğüs kanseri vakalarını tahmin etmek için dördü kümeleme algoritmalarından (K-Means, EM (Expectation Maximization), PAM (Partition Around Medoids) ve Fuzzy C-Means) ve dördü sınıflandırma algoritmalarından (SVM, Karar Ağacı C5.0, KNN ve Saf Bayes) olmak üzere sekiz popüler veri madenciliği yöntemlerini karşılaştırmışlardır. Deneyler, Wisconsin Prognostic Breast Cancer veri kümesi üzerinde yürütülmüştür. En yüksek doğruluk (%81), SVM ve Karar Ağacı ile elde edilirken; en düşük doğruluk (%37) Fuzzy C-Means yöntemi ile elde edilmiştir.

Ghosh ve diğ. (2014), göğüs kanseri teşhisi için MLP-BPN (Multilayer Perceptron using Back Propagation Neural Network) ve SVM yöntemlerini kullanmışlardır. Deneyler, Wisconsin Breast Cancer veri kümesi üzerinde yürütülmüştür. Çalışmanın sonunda, SVM yöntemi ile daha yüksek doğruluk elde edilmiştir (9 Dimension- %96.85).

Osareh ve Shadgar (2010), göğüs kanseri teşhisi ve kanser nüksünün risk değerlendirmesi konularını SVM, KNN ve olasılıklı sinir ağı (PNN) kullanarak araştırmışlardır. Bu sınıflandırıcılar, sinyal-gürültü oranı (SNR) özellik sıralama yöntemi, sıralı ileri seçim tabanlı (SFS) özellik seçimi ve PCA özellik dönüşümü ile birleştirilmiştir. Çalışmada, FNAB (Fine Needle Aspirate of Breast Lesions) veri kümesi ve gen mikro dizilerinden oluşan bir veri kümesi kullanılmıştır. En yüksek doğruluk, SVM-RBF yöntemi ile elde edilmiştir (%98.80).

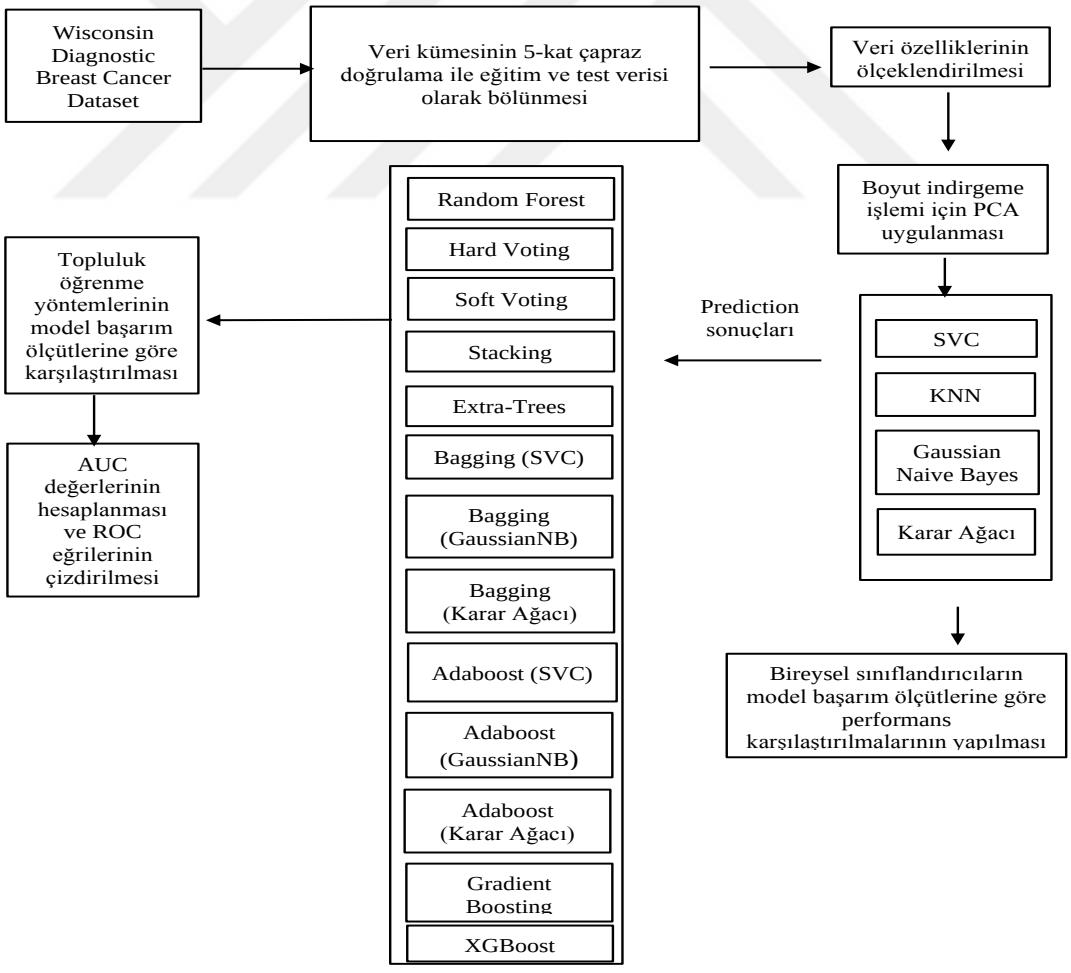
Azmi ve Cob (2010), göğüs kanserini tespit etmek için ileri beslemeli geri yayılım algoritmasına dayalı otomatik bir tanı sistemi sunmuşlardır. Deneyler, Wisconsin Breast Cancer veri kümesi üzerinde yürütülmüştür. Deneysel sonuçlara göre en yüksek doğruluk, gizli 7 katmanı olan sinir ağı ile elde edilmiştir (%96.63).

Gayathri ve Sumathi (2016), göğüs kanserini tespit etmek için RVM (Relevance Vector Machine) yöntemini uygulamışlardır. Boyut indirgeme işlemi için Doğrusal Diskriminant Analizi kullanılmıştır. Çalışmada, Wisconsin Original Breast Cancer veri kümesi kullanılmıştır. Çalışma sonucunda; %96 doğruluk, %98 duyarlılık ve %94 özgüllük elde edilmiştir.

3. MALZEME VE YÖNTEM

3.1. ÖNERİLEN GÖĞÜS KANSERİ TESPİT SİSTEMİ

Bu çalışmada; Destek Vektör Makinesi, K-En Yakın Komşu, Saf Bayes ve Karar Ağacı algoritmaları bireysel sınıflandırıcı olarak kullanılmıştır. Çalışmada uygulanan bireysel sınıflandırıcıların doğruluk oranlarını arttırmak için Rassal Orman, Oylama, Torbalama, Yükseltme, Yığma ve Ekstra Ağaçlar topluluk öğrenmesi yöntemleri uygulanmış ve model başarım ölçütlerine göre birbirleriyle karşılaştırılmıştır. Bu tez çalışmasında kullanılan sistemin çalışma mekanizması, Şekil 3.1'de gösterilmiştir.



Şekil 3.1: Önerilen sistemin çalışma mekanizması.

3.2. VERİ KÜMESİ BÖLÜMLENDİRME

Makine öğrenmesi yöntemleri uygulanırken algoritmaların performansını değerlendirebilmek için, veri kümesi genellikle eğitim ve test veri kümesi olmak üzere ikiye ayrılmaktadır. Eğitim veri kümesi, orijinal veri kümesinin modeli eğitmek için kullandığımız bir parçasıdır. Eğitim seti, modeli eğitmek için kullanılan veri kümesidir. Test veri kümesi, modelin eğitim aşaması tamamlandıktan sonra değerlendirilmesini sağlamak için kullanılan veri kümesidir. Test veri kümesi, model tamamen eğitildikten sonra kullanılmalıdır. Eğitim aşaması bitene kadar, modelin test setine maruz bırakılmaması gerekmektedir. Bu sayede, değerlendirme sonucunda elde edilen doğruluk oranının güvenilir olduğu söylenebilir. Bu çalışmada, veri kümelerini eğitim ve test kümesi olarak iki kategoriye ayırmak için k-kat çapraz doğrulama yöntemi kullanılmıştır. K-kat çapraz doğrulama yönteminde, örnek verilerin bölünebileceği grup sayısını gösteren k parametresinin seçilmesi gerekmektedir. Bu çalışmada, 5-kat çapraz doğrulama kullanılmıştır (k=5).

K-kat çapraz doğrulama yönteminin çalışma mekanizması şu şekildedir:

1. Veri kümesi, k parçaya bölünür. Her parça için şu adımlar uygulanır:
   1.1. Seçilen parça, test kümesi olarak alınır. Geriye kalan k-1 parça, eğitim kümesi olarak alınır.
   1.2. Model, eğitim kümesinde eğitilir.
   1.3. Model, test kümesinde doğrulanır.
2. Süreç tamamlandığında, her bir kat için elde edilen model başarım ölçütlerinin ortalaması alınır. Bu ortalama değer, modelin performansı olarak kabul edilir.

Bu çalışmada, aşırı öğrenme problemini önlemek için k-kat çapraz doğrulama yöntemi kullanılmıştır. Veri kümesinin eğitim ve test kümelerine bölünmesi işlemi, sklearn kütüphanesinin model_selection paketinin KFold sınıfı kullanılarak gerçekleştirilmiştir.

3.3. VERİ ÖN İŞLEME

Veri kümesindeki özelliklerin değer aralığı değişkenlik göstermektedir. Özellik Ölçeklendirme, veri kümesini sabit bir aralıkta standartlaştırmak için kullanılmaktadır. Makine öğreniminde özellik ölçeklendirme işlemi; model oluşturmadan önce, verilerin ön

işlemesi sırasında uygulanan önemli bir adımdır. Eğitim setindeki örnekler üzerinde ilgili istatistikleri hesaplayarak, her özellik için bağımsız olarak gerçekleştirilmektedir. Bu çalışmada ölçeklendirme işlemi için; sklearn kütüphanesinin modülü olan sklearn.preprocessing.StandardScaler kullanılmıştır.

Özellik ölçeklendirme işlemi gerçekleştirilirken; öncellikle her özellik için dağılım ortalaması ve standart sapma belirlenmektedir. Daha sonra, her bir özellikten ortalama değer çıkarılmakta ve elde edilen bu değer, standart sapma değerine bölünmektedir. X standartlaştırılacak özellik, μ özelliğin ortalama değeri ve σ özelliğin standart sapması olup özellik ölçeklendirme aşağıdaki şekilde hesaplanmaktadır:

$$z = \frac{x-\mu}{\sigma} \tag{3.1}$$

3.4. BOYUT İNDİRGEME

Boyut indirgeme işlemini gerçekleştirme amacıyla birçok teknik geliştirilmiştir, ancak Temel Bileşen Analizi en eski ve en yaygın kullanılan tekniklerden biridir. Temel Bileşen Analizi; büyük bir değişken kümesini, büyük kümedeki bilgilerin çoğunu içermekte olan küçük bir kümeye indirgeyen bir boyut indirgeme yöntemidir. Temel Bileşen Analizi, büyük veri kümelerinin boyutlarını azaltmak ve bu işlemi gerçekleştirirken mümkün olduğunca çok değişkenliği koruyarak bilgi kaybını en aza indirmek için kullanılan bir tekniktir. Bir dizi ilişkili değişkeni, ortogonal bir dönüşüm kullanarak ana bileşenler adı verilen ilişkisiz doğrusal değişkene dönüştüren bir prosedürdür. Bu temel bileşenler, orijinal değişkenlerin sayısına eşit veya orijinal değişkenlerin sayısından daha az olabilir. Temel Bileşen Analizi ile boyut indirgeme gerçekleştirirken, veri kümesinin özellik boyutunu azaltmada kümülatif varyans kullanılmaktadır. Kümülatif varyans kuralı; temel bileşenlerin sayısını, Eigen değer boyutlarına veya her bir temel bileşenin açıkladığı varyans oranına göre seçmektedir (Hasan & Tahir, 2010).

Temel Bileşen Analizi algoritmasını uygulamadan önce, her bir özelliğin ortalama değerini veri kümesinden çıkararak ve her bir boyut aynı aralıkta olacak şekilde ölçeklendirerek normalize etmek gerekmektedir. PCA hem eğitim hem de test verilerine uygulanmalıdır.

3.5. SINIFLANDIRMA YÖNTEMLERİ

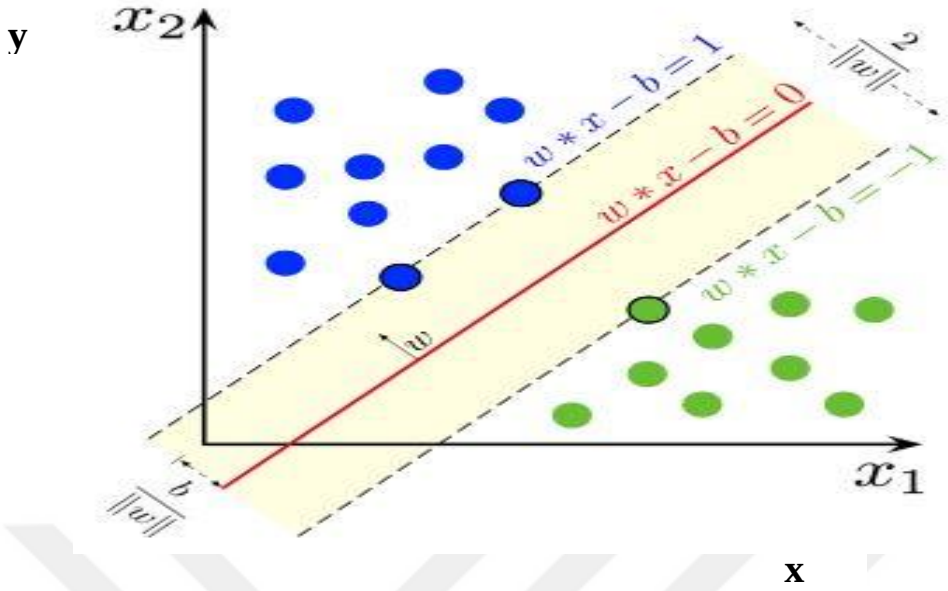
Bu çalışmada; Destek Vektör Makinesi, K-En Yakın Komşu, Saf Bayes ve Karar Ağacı algoritmaları bireysel sınıflandırıcı olarak kullanılmıştır. Doğruluk, kesinlik, duyarlılık ve f-skor model başarım ölçülerine göre bireysel sınıflandırıcıların performansları birbirleriyle karşılaştırılmıştır. Ardından, bireysel sınıflandırıcıların doğruluk oranını arttırmak için topluluk öğrenmesi yöntemleri kullanılmıştır.

3.5.1. Destek Vektör Makinesi

Destek Vektör Makinesi, kanser teşhisi alanında yaygın olarak uygulanan gözetimli makine öğrenmesi sınıflandırma tekniklerinden biridir. Boser ve diğ. (1992), şu an Destek Vektör Makineleri olarak bilinen algoritmaya evrilen gözetimli bir sınıflandırma algoritması önermişlerdir. Hem sınıflandırma hem de regresyon işlemi için kullanılabilecek bir makine öğrenmesi yöntemi olup, esas olarak sınıflandırma için kullanılmaktadır. Destek Vektör Makinesi yöntemi; aşırı öğrenme, küçük veri kümesi ve doğrusal olmayan verilerin yol açtığı sorunları üstesinden gelebilmektedir (Li & Zhang, 2009).

Destek vektör makinesi algoritması hem doğrusal hem de doğrusal olmayan verileri sınıflandırabilmektedir. SVM algoritması; veri noktalarını kategorize eden n boyutlu bir uzayda bir hiper düzlem bulmaya yardımcı olmaktadır. Burada n, özelliklerin sayısıdır. Hiper düzlem ile en yakın veri noktaları arasındaki mesafe, margin olarak adlandırılmaktadır. Sınıflandırmayı gerçekleştirmek için, iki sınıfı maksimum margin ile ayıran en uygun hiper düzlemin bulunması gerekmektedir.

Bu çalışmada, Destek vektör makinesi yöntemini uygulamak için sklearn kütüphanesinin SVC sınıfı kullanılmıştır. Şekil 3.2'de, Destek Vektör Makinesi'nin basitleştirilmiş bir örneklemesi yer almaktadır.



Şekil 3.2: SVM algoritmasının çalışma mekanizması.

Girdi verisi $x_i \in R^m$, eğitim seti $\{(x_i, y_i)\}_{i=1}^N$ ve karşılık gelen sınıf etiketleri $y_i \in \{-1, +1\}$ olarak verilmiştir. İki sınıfı ayırmak için hiper düzlem kullanan sınıflandırıcı modeli aşağıda yer almaktadır:

$$Y(x) = sgn\ (w^T{}_\varphi(x) + b) \tag{3.2}$$

Burada; x özellik vektörü, w ağırlık vektörü, b sapma değeri, ∅ doğrusal olmayan fonksiyon ve y sınıf etiketi olarak ifade edilmektedir. En uygun hiper düzlemi bulurken marjı (M) maksimize edebilmek için, || w || 'yi en aza indirmemiz gerekmektedir:

$$M = \frac{2}{|w|} \tag{3.3}$$

$$\frac{||w^2||}{2} = \frac{1}{2} w^T w \tag{3.4}$$

Aşağıda gösterilen durumda, çıkış sınıfı $y_i = +1$ değerini almaktadır:

$$w^T{}_\varphi(x_i) + b \geq +1 \tag{3.5}$$

Aşağıda gösterilen durumda, çıkış sınıfı $y_i = -1$ değerini almaktadır:

$$w^T{}_\varphi(x_i) + b \leq -1 \tag{3.6}$$

(3.5) ve (3.6) eşitsizliği, (3.7) eşitsizliğinde birleştirilmiştir:

$$y(w^T{}_{\varphi}(x_i) + b) \geq +1\,,\ i = 1,2,\dots N \tag{3.7}$$

3.5.2. K-En Yakın Komşu

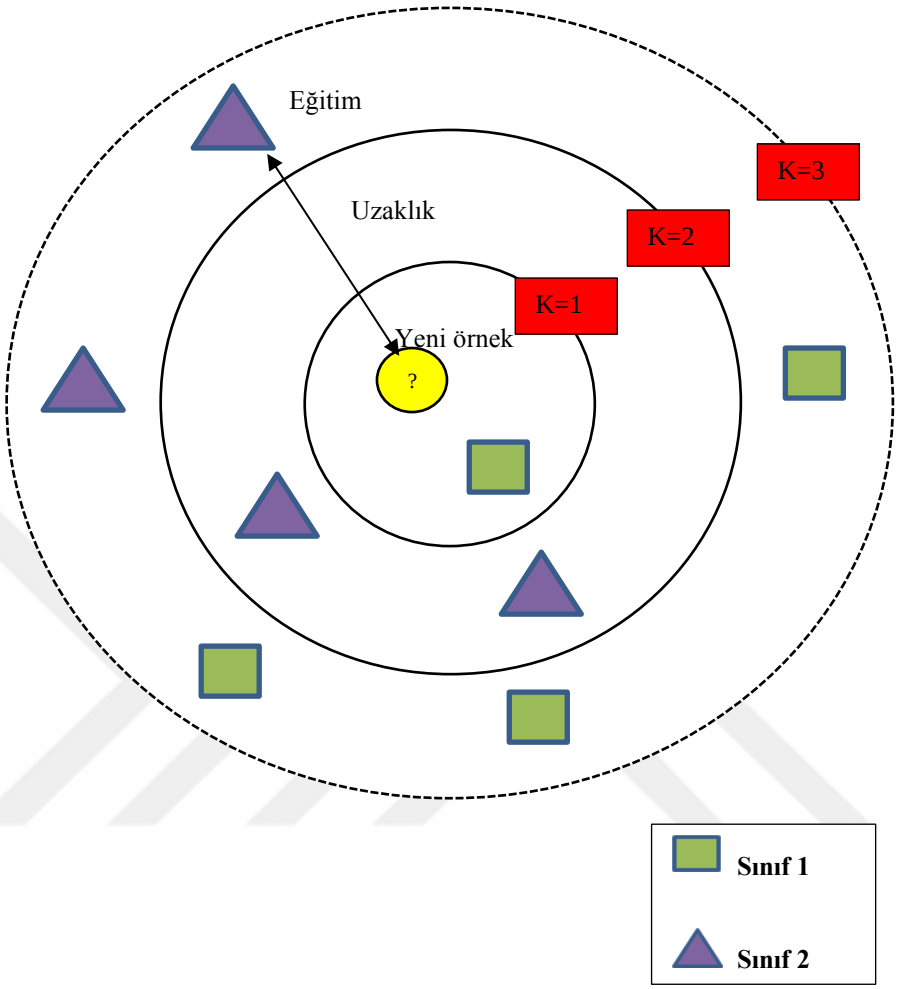
K-En Yakın Komşu hem sınıflandırma hem de regresyon problemlerini çözmenin bir yolu olarak kullanılabilen gözetimli bir algoritmadır. Test veri noktasının sınıflandırılması, veri kümesinin parametreleri yerine en yakın eğitim veri noktalarına dayandığı için, KNN parametrik olmayan bir yöntemdir (Islam vd., 2017).

KNN yöntemi ile oluşturulan model, bir uzaklık fonksiyonu ve sınıfın seçim fonksiyonu ile ilişkili eğitim örneğinin sonuçlarıdır. Yeni bir örnek sınıflandırılmadan önce, benzerlik ölçüsü kullanılarak diğer örneklerle karşılaştırılmaktadır. KNN algoritmasında; bir nokta ile örnekler arasındaki uzaklıklar bulunmakta, uzaklık kullanılarak komşular ağırlıklandırılmakta, noktaya en yakın gerekli sayı örnekleri (K) seçilmekte ve komşular arasında en yaygın olan sınıf, sınıflandırılacak örneğe atanmaktadır. KNN algoritmasındaki "K", oy alındığı düşünülen en yakın komşuların sayısıdır. "K" için farklı değerlerin seçilmesi, aynı örnek için farklı sınıflandırma sonuçları elde edilmesine yol açabilmektedir.

KNN algoritmasında; bir düzlemdeki iki nokta arasındaki uzaklığın hesaplanması, Öklid uzaklığı olarak adlandırılan matematiksel formüle dayanmaktadır. Bir düzlemdeki $A(x_0, y_0)$ ve $B(x_1, y_1)$ noktalarının aralarındaki Öklid uzaklığı, aşağıdaki şekilde hesaplanmaktadır:

$$\sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2} \tag{3.8}$$

K-En Yakın Komşuluk algoritmasının, yeni bir nesneyi sınıflandırmak için kullandığı çalışma mekanizmasının bir örneklemesi, Şekil 3.3'te gösterilmiştir.



Şekil 3.3: KNN algoritmasının çalışma mekanizması.

3.5.3. Saf Bayes

Saf (Naive) Bayes, sınıflandırma görevleri için kullanılan olasılığa dayalı bir makine öğrenmesi algoritmasıdır. Bayes teoremine dayalı bir sınıflandırma tekniğidir. Saf Bayes sınıflandırma tekniği kullanılarak, belirli bir sınıfa ait bir örneğin olasılığı tahmin edilmektedir. Bayes teoremine göre; bir özelliğin değerinin bir sınıf üzerindeki etkisinin, diğer özellik değerlerinden bağımsız olduğu varsayılmaktadır. Bayes teoreminin formülü aşağıda verilmiştir:

$$P\ (A|B) = \frac{P\ (B|A)P(A)}{P(B)} \tag{3.9}$$

Bayes teoremi formülünde;

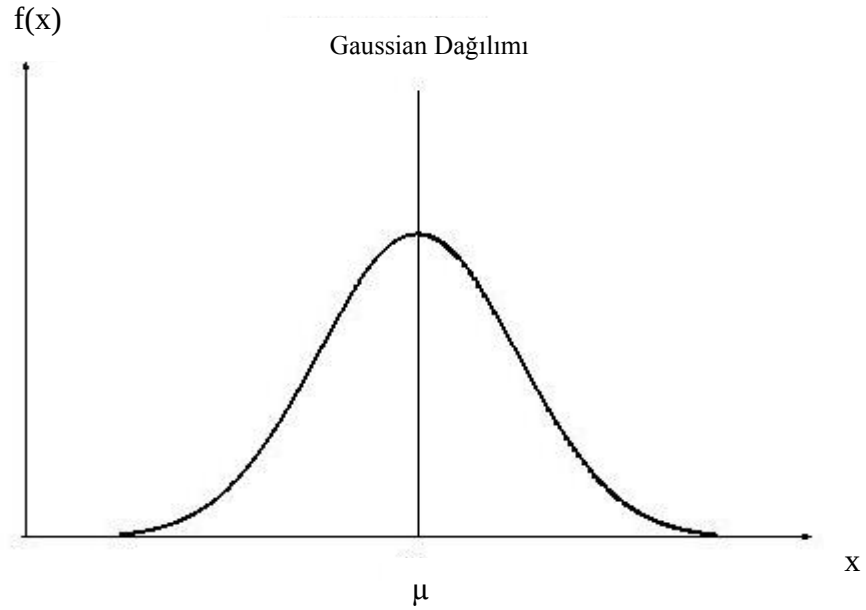
- P (A|B): B olayının meydana gelmesi durumunda A olayının gerçekleşme olasılığını,
- P (B|A): A olayının meydana gelmesi durumunda B olayının gerçekleşme olasılığını,
- P (A): A olayının meydana gelme olasılığını,
- P (B): B olayının meydana gelme olasılığını belirtmektedir.

Saf Bayes sınıflandırıcısının avantajlarından biri, eğitim verilerinin büyüklüğüne bağlı olmamasıdır. Bu sayede, sınıflandırma için gerekli parametreleri tahmin etmek için az sayıda eğitim verisi kullanabilmektedir.

Multinominal, Bernoulli ve Gaussian gibi farklı Naive Bayes sınıflandırıcı türleri bulunmaktadır. Bu çalışmada Gaussian Naive Bayes sınıflandırıcısı kullanılmıştır. Sürekli verileri işlerken, her sınıfa karşılık gelen sürekli değerlerin bir Gauss dağılımına göre dağıtılması esasına dayanmaktadır. Gauss olasılık fonksiyonu aşağıda yer alan denklem ile tanımlanmaktadır:

$$P\,(x_i|y) = \frac{1}{\sqrt{2\pi\sigma^2 y}}\, exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right),\ i = 1,2,..n \qquad (3.10)$$

Burada; μ, ortalamayı ve σ^2, varyansı ifade etmektedir. Gaussian dağılımı, Şekil 3.4'te verilmiştir.



Şekil 3.4: Gaussian dağılımı.

3.5.4. Karar Ağacı

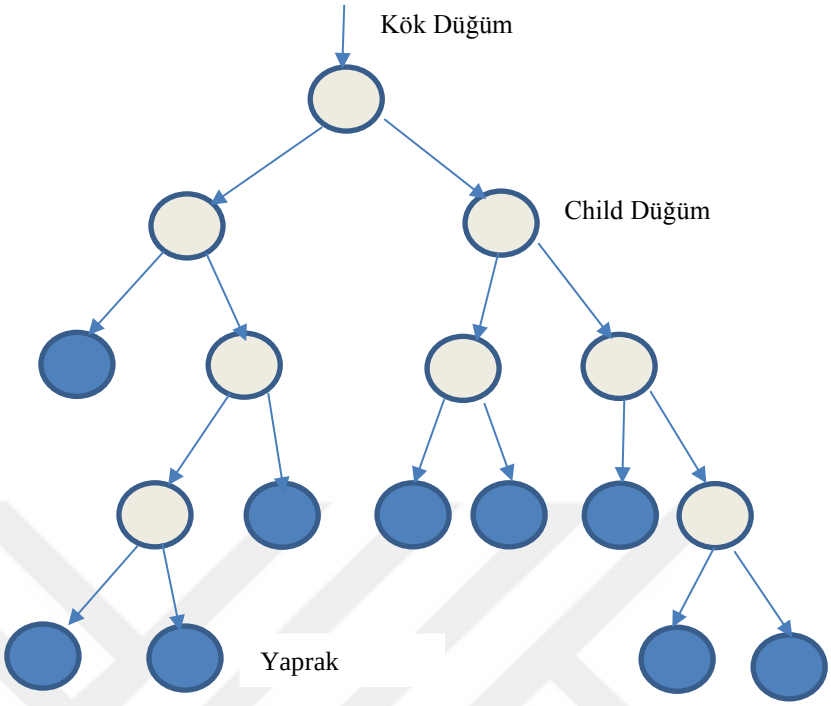
Karar Ağacı, gözetimli bir öğrenme algoritmasıdır. Regresyon ve sınıflandırma ile ilgili problemleri çözmek için kullanılmaktadır. Karar ağaçları hem kategorik hem de nümerik verileri işleyebilmektedir.

Karar ağacı; kök düğüm, bir özelliği test eden iç karar düğümleri, özellik değerlerine karşılık gelen dallar ve bir sınıflandırma atayan yaprak düğümlerden oluşmaktadır. Bir ağaçtaki en iyi tahminci olan en üstteki karar düğümüne, kök düğüm denilmektedir. Bir karar ağacı oluşturmak için ilk adım bölmektir. Örnekler, kök düğümden başlayarak sınıflandırılmaktadır. Özellik değerlerine göre ağaç, bir yaprak düğüme doğru sıralanmaktadır. Karar ağacı algoritmasında; dallar bir karar kuralını belirtmekte ve her bir yaprak düğüm, sonucu belirtmektedir. Karar ağacı algoritması ile; veri kümeleri daha küçük alt gruplara ayrılırken, aynı zamanda bir karar ağacı oluşturulmaktadır. Veri kümelerinin bölünmesi işleminden sonra, aşırı öğrenme sorununun üstesinden gelmek için budama işlemi yapılabilmektedir. Budama; bazı dalları yaprak düğümlerine çevirerek ve orijinal dalın altındaki yaprak düğümlerini kaldırarak ağacın boyutunu küçültme işlemidir (Avhad vd., 2020).

Basit bir karar ağacı algoritmasının çalışma mekanizması şu şekildedir:

1. Ağaca, tüm veri kümesini içeren kök düğüm ile başlanılır.
2. Öznitelik Seçimi Ölçüsü kullanılarak veri kümesindeki en iyi öznitelik seçilir.
3. Seçilen en iyi öznitelik, bir karar düğümü yapılır ve veri kümesi daha küçük alt kümelere bölünür.
4. Düğümler, daha fazla sınıflandırılamayana ve son düğüm, yaprak düğümü olarak adlandırabilene kadar; bu işlem her çocuk için tekrarlanılarak ağaç oluşturmaya devam edilir.

Şekil 3.5’te, karar ağacının yapısı gösterilmiştir.



Şekil 3.5: Karar Ağacı yapısı.

3.5.4.1. Öznitelik Seçim Ölçüsü

Öznitelik seçim ölçüsü, verileri en iyi şekilde bölümlere ayırmak için kullanılan bölme ölçütünü seçerken uygulanmaktadır. En popüler seçim ölçüleri; bilgi kazancı, kazanım oranı, gini endeksidir. Bu çalışmada, verileri en iyi şekilde bölmek için bilgi kazancı kullanılmıştır.

Bilgi kazancı, belirli bir özelliğin eğitim örneklerini ne kadar iyi ayırdığını ölçen istatistiksel bir özelliktir. Bir veri kümesinin bir özniteliğe göre bölümlendirilmesinden sonra entropideki değişikliklerin ölçülmesi ile elde edilmektedir. Bilgi kazancı aşağıdaki (3.11) ve (3.12) formülleri kullanılarak hesaplanabilmektedir:

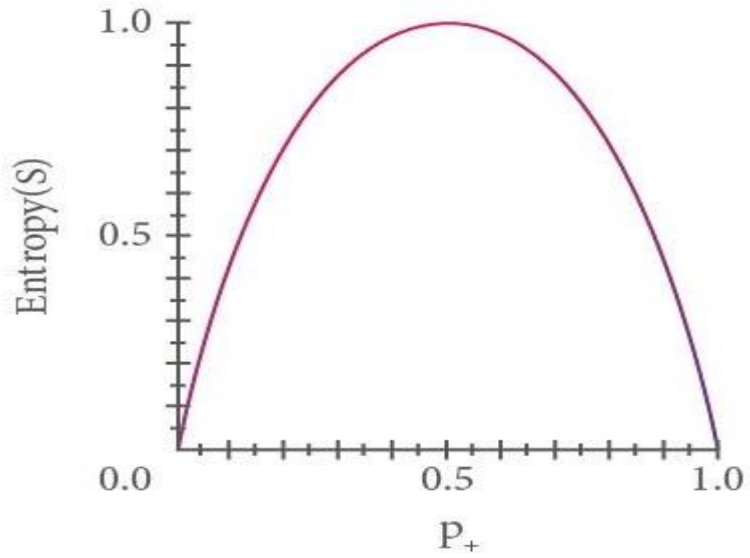
$$Bilgi\ Kazancı = Entropi(parent) - [Ortalama\ Entropi(children)] \quad (3.11)$$

$$Entropi\ (S) = -p_+ log_2 p_+ - p_- log_2 p_- \quad (3.12)$$

Verilen (3.12) formülünde;

- S, eğitim örneklerinden birini,
- p_+, ilgili düğümdeki başarı olasılığını,
- p_-, ilgili düğümdeki başarısızlık olasılığını ifade etmektedir.

İkili sınıflandırmaya göre entropi fonksiyonunun formu, Şekil 3.6'da gösterilmiştir.



Şekil 3.6: İkili sınıflandırmaya göre entropi fonksiyonunun formu.

3.6. TOPLULUK YÖNTEMLERİ

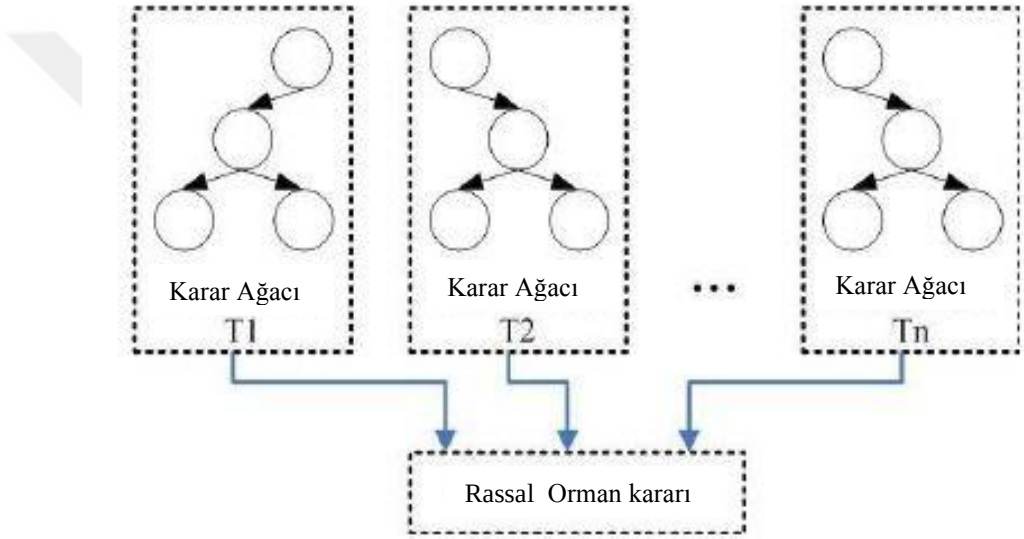
Bu çalışmada bireysel sınıflandırıcıların performansını iyileştirmek için; Rassal Orman (Random Forest), Oylama (Voting), Torbalama (Bagging), Yükseltme (Boosting), Yığma (Stacking) ve Ekstra Ağaçlar (Extra-Trees) topluluk yöntemleri kullanılmıştır. Doğruluk, kesinlik, duyarlılık ve f-skor, roc eğrisi ve auc-skor model başarım ölçütlerine göre topluluk yöntemlerinin performansları birbirleriyle karşılaştırılmıştır.

3.6.1. Rassal Orman

Rassal orman, daha iyi bir tahmin performansı elde etmek için birkaç karar ağacının birden çok modelini kullanan bir toplu sınıflandırıcıdır. Bunun arkasındaki temel düşünce, bir grup zayıf öğrenicinin bir araya gelerek güçlü bir öğrenci oluşturabilmesidir. Rassal orman hem sınıflandırma hem de regresyon görevlerinde kullanılabilmektedir. Birçok karar ağacından oluşmakta ve sınıfın çıktısının modu olan sınıfı, tek tek ağaçlar aracılığıyla çıkartmaktadır. Örnek setler üzerinde karar ağaçları oluşturulduktan sonra birçok ağaç üretilmekte, böylece bir orman yaratılmaktadır. Rassal orman modelinde; sınıflandırma için girdi vektörü, her

ağaca beslenmekte ve her ağaç bir sınıf için oy kullanmaktadır (Mushtaq & Mellouk, 2017). Son olarak rastgele orman modeli, en yüksek oy sayısına sahip sınıfı seçmektedir. Her ağacın girdisi, orijinal veri kümesinden alınan örneklenmiş verilerdir. Ayrıca, her düğümde ağacı büyütmek için rastgele bir özellik alt kümesi seçilmektedir. Rassal Orman yöntemi, büyük verileri doğru bir şekilde sınıflandırabilme yeteneğine sahiptir.

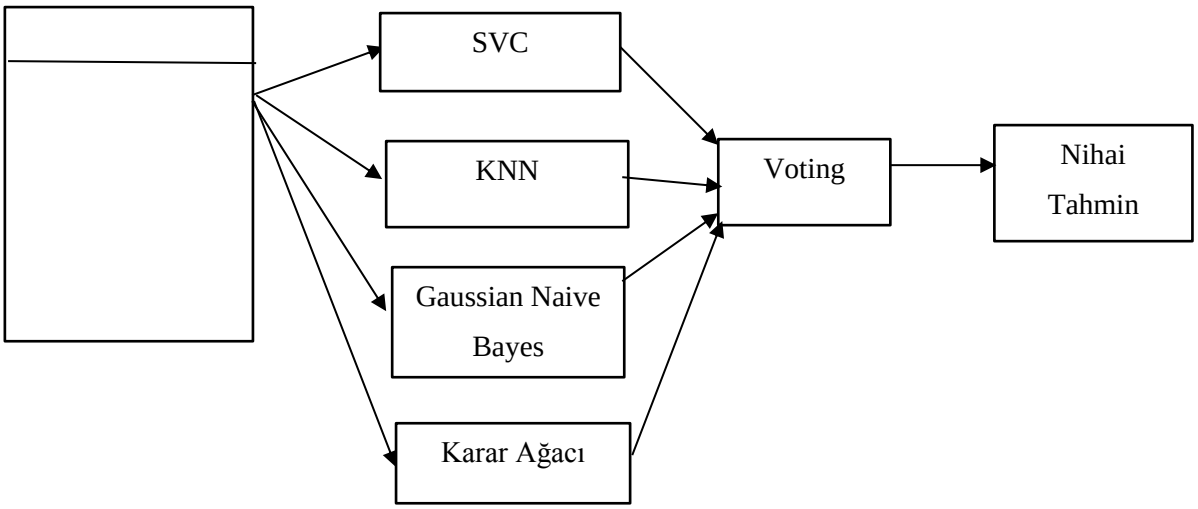
Rassal orman modelinin çalışma mekanizması, Şekil 3.7'de verilmiştir.



Şekil 3.7: Rassal Orman modeli.

3.6.2. Oylama

Oylama topluluk yöntemi, birden fazla bireysel sınıflandırıcı ile elde edilen tahminleri birleştirmek için oylamayı kullanmaktadır. Bununla birlikte, Rassal Orman yönteminden farklı olarak bireysel sınıflandırıcı için kullanılan temel modeller homojen olmayabilir. Başka bir deyişle, farklı temel öğreniciler eğitilebilmektedir. Örneğin bu çalışmada; SVC, K-En Yakın Komşuluk, Gaussian Naive Bayes, Karar Ağacı ve ardından sonuçları birleştirmek için Oylama topluluk yöntemleri kullanılmıştır. Bu çalışmada; çoğunluğa dayalı (hard) ve olasılığa dayalı (soft) oylama kullanılmıştır. Oylama topluluk yönteminin iş akışı, Şekil 3.8'de verilmiştir.



Şekil 3.8: Oylama topluluk yöntemi iş akışı.

3.6.2.1. Çoğunluğa Dayalı Oylama

Çoğunluğa dayalı oylama (Hard Voting), topluluk sınıflandırma yöntemleri arasında yaygın olarak kullanılmaktadır. Çoğunluğa dayalı oylamada, çoğunluk kuralları kullanılmaktadır. Her bireysel sınıflandırıcı, bir sınıf için oy kullanmaktadır. Son sınıf tahmini, çoğunluk oyuyla belirlenmektedir. Nihai çıktı tahmini, oyların yarısından fazlasını alan olmaktadır. Modellerin yaptığı tahminlerin hiçbiri oyların yarısından fazlasını alamazsa, uygulanan çoğunluğa dayalı topluluk yöntemi ile istikrarlı bir tahmin yapılamayacağı söylenebilmektedir.

Sınıf etiketi Y, her bir Cj sınıflandırıcısının çoğunluk oyu ile aşağıda verilen şekilde belirlenmektedir:

$$Y = mode\{C_1(x), C_2(x), \dots, C_m(x)\} \qquad [j = 1,2,3,\dots,m] \tag{3.13}$$

Bu çalışmada çoğunluğa dayalı oylamada; SVC, K-En Yakın Komşuluk, Gaussian Naive Bayes ve Karar Ağacı bireysel sınıflandırıcıları temel öğreniciler olarak kullanılmıştır.

3.6.2.2. Olasılığa Dayalı Oylama

Olasılığa dayalı oylamada (Soft Voting); her modelin aynı öneme sahip olduğu çoğunluğa dayalı oylama yönteminin aksine, bir veya birden fazla modelin önemi artırılabilmektedir. Olasılığa dayalı oylama yönteminde; her bir sınıflandırıcı, belirli bir veri noktasının belirli bir hedef sınıfa ait olduğuna dair bir olasılık değeri vermektedir. Tahminler, sınıflandırıcının önemine göre ağırlıklandırılmaktadır. En yüksek ağırlıklandırılmış olasılık toplamına sahip hedef etiket, oylamayı kazanmaktadır.

Olasılığa dayalı oylama yönteminde; nihai çıktı, sınıflandırıcılar için tahmin edilen olasılıklara bağlı olarak tahmin edilmektedir. Olasılığa dayalı oylama yöntemi, tahmin edilen olasılıkların toplamının argmax / en büyük tahmin edilen değerine dayalı bir etiket tahmin etmektedir. Sınıf etiketi Y, aşağıda verilen şekilde belirlenmektedir:

$$Y = argmax_i \sum_{j=1}^{m} W_j P_{ij} \text{ [j = 1, 2,3, …, m; i = 1, 2, 3, …,n]} \quad (3.14)$$

Yukarıda verilen formüle göre; i ve j sınıflandırıcı olup Pij, bu sınıflandırıcıların tahmini olasılık değeridir. Wj, j. sınıflandırıcının ağırlığıdır.

Bu çalışmada olasılığa dayalı oylamada; SVC, K-En Yakın Komşuluk, Gaussian Naive Bayes ve Karar Ağacı bireysel sınıflandırıcıları temel öğreniciler olarak kullanılmıştır.

3.6.3. Torbalama

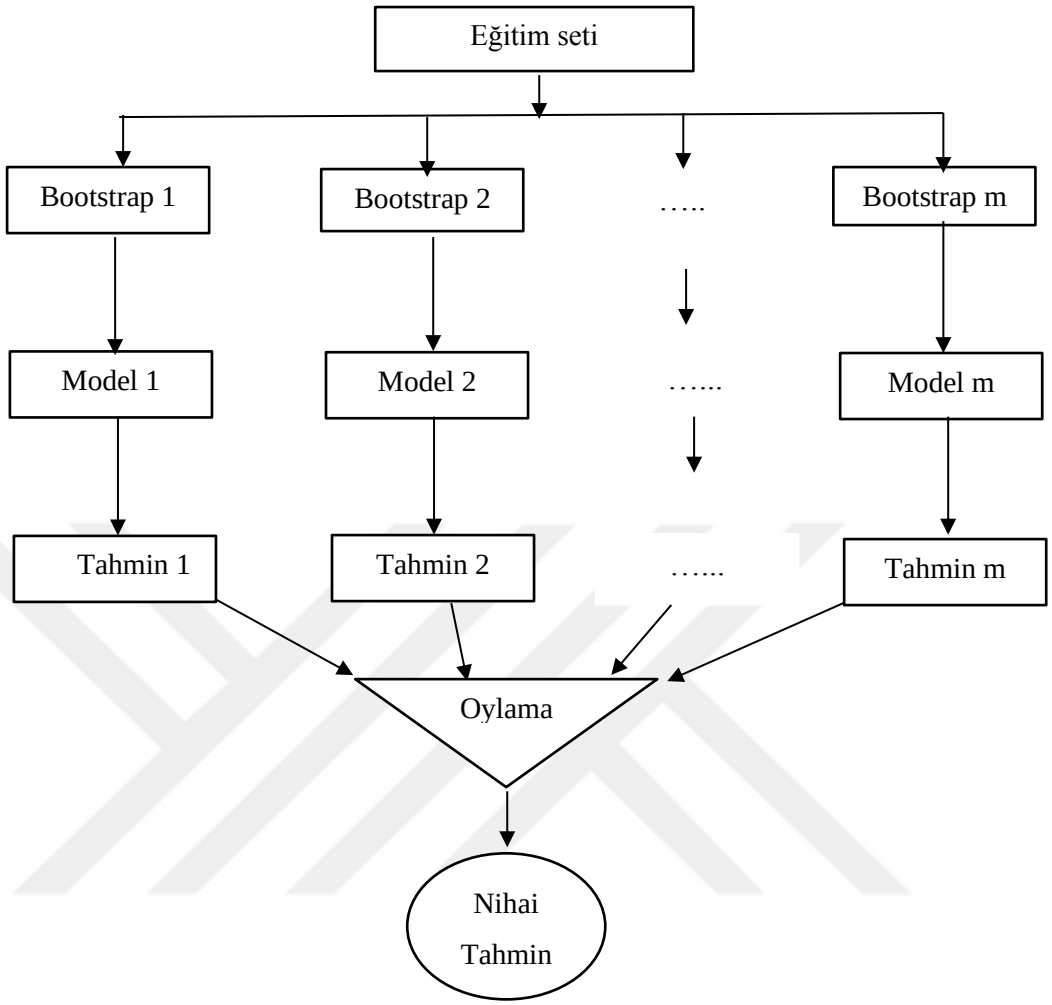
Zayıf temel öğreniciler, bireysel öğrenicilerden daha güçlü bir öğrenici oluşturmak için birleştirilmektedir. Torbalama, Bootstrap prosedürünün yüksek varyanslı bir makine öğrenimi algoritmasına uygulanması işlemidir.

Torbalama topluluk yönteminde, örnek setleri değiştirilerek rastgele N kez çekilmekte ve alt kümeler oluşturulmaktadır. Her bir bireysel sınıflandırıcı, çekilen alt kümelerden biriyle eğitilmektedir. Torbalama işleminde değiştirilerek çekiliş yapmak, toplulukta temsil edilme olasılığının eşit olmasını sağladığı için önemlidir. Bu yöntemde, birkaç farklı model ayrı ayrı eğitilmekte ve her bir model test örnekleri için oy kullanmaktadır. Modeller, birbirinden bağımsız çalışmaktadır. Nihai tahmin, modellerin çoğunluk oyu ile belirlenmektedir. Torbalama yöntemi, modellerin aşırı öğrenme sorununu azaltmaktadır.

$$c_i = argmax_i\{\frac{1}{N} \sum_{T=1}^{N} h_T(x, c_i)\}, \text{ [T = 1, …, N]} \quad (3.15)$$

Yukarıda verilen formüle göre; x test örneğini, T bootstrap eğitim setini temsil etmektedir. $h_T(x)$, T üzerindeki bireysel tahmin modelidir.

Bu tez çalışmasında; SVC, Gaussian Naive Bayes ve Karar Ağacı bireysel sınıflandırıcıları ile Bagging(SVC), Bagging(GaussianNB), Bagging(Karar Ağacı) olmak üzere ayrı ayrı torbalama sınıflandırıcıları oluşturulmuştur. Torbalama topluluk yöntemi iş akışı, Şekil 3.9'da verilmiştir.



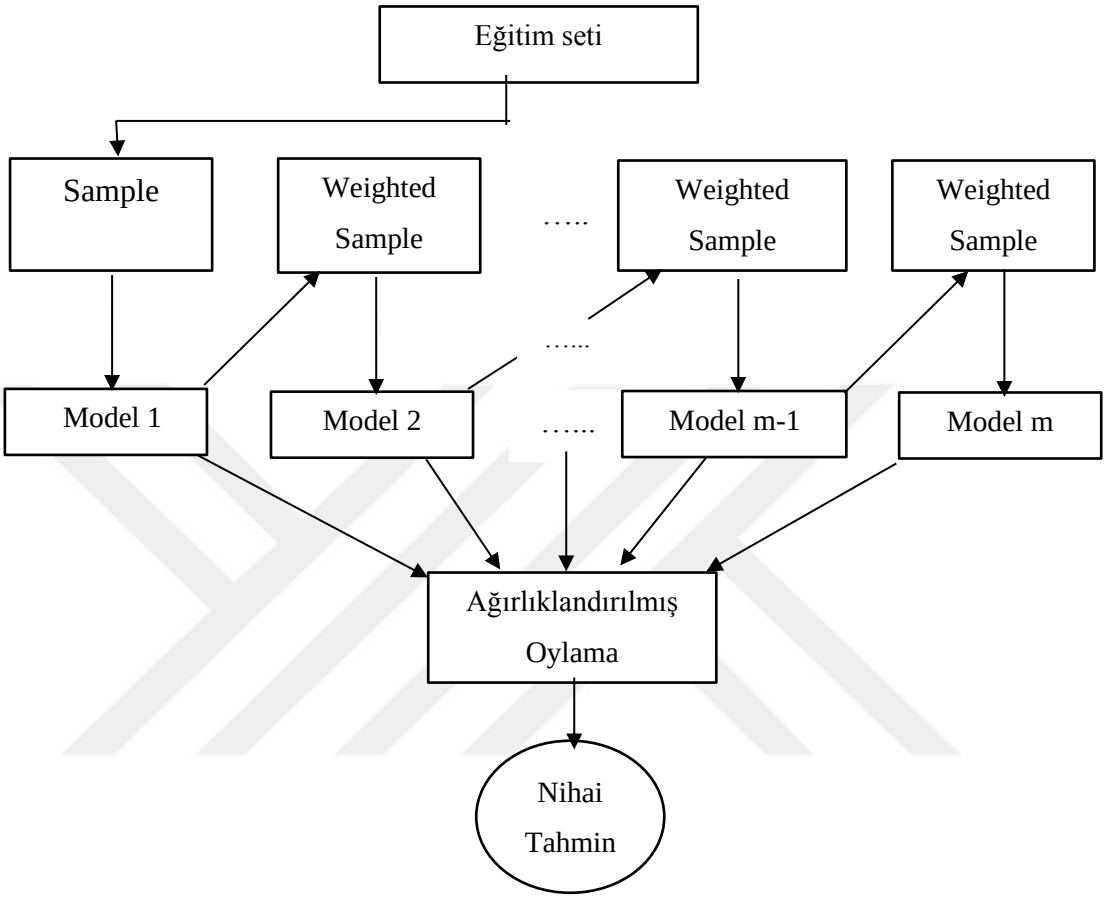
Şekil 3.9: Torbalama topluluk yöntemi iş akışı.

3.6.4. Yükseltme

Yükseltme topluluk yönteminde; zayıf temel öğreniciler, daha güçlü bir öğrenici modeli oluşturmak için birleştirilmektedir. Yükseltme yönteminde, sınıflandırıcılar önceki sınıflandırıcıların hatalarına odaklanmaktadırlar. Daha iyi tahminler yapabilmek için önceki tahmin hatalarından öğrenme işlemi gerçekleştiren bir topluluk öğrenmesi yöntemidir.

Yükseltme topluluk yöntemi, sınıflandırıcıları bir sıraya göre düzenleyerek eğitmekte ve bu sınıflandırıcıları tahmin için birleştirerek çalışmaktadır. Bir temel öğrenicinin çıktısı, diğer temel öğrenicinin girdisi durumundadır. Yükseltme; örneklerin ağırlıklarının, son tahminin hatasına göre ayarlandığı bir topluluk oluşturmaktadır. Torbalama işleminden farklı olarak sıralı bir yöntem olması nedeniyle, bu yöntem uygulanırken paralel işlemler

kullanılamamaktadır. Yükseltme topluluk yönteminin çalışma prensibi, Şekil 3.10'da verilmiştir.



Şekil 3.10: Yükseltme topluluk yöntemi çalışma prensibi.

Aşağıda verilen yükseltme yönteminin çalışma formülünde; y_m, m. hedef değerini, $h_m(x)$ bireysel sınıflandırıcıyı temsil etmektedir.

$$H_m(x) = \sum_{m=1}^{M} y_m \; h_m(x) \qquad [m=1,\ldots,M] \tag{3.16}$$

Birçok farklı yükseltme topluluk yöntemi bulunmakla birlikte en yaygın kullanılan yükseltme yöntemleri; Adaptive Boosting (Adaboost), Gradient Boosting (GBM), Stochastic Gradient Boosting (SGBM) ve XGBoost (XGBM) yöntemleridir. Bu tez çalışmasında; Adaboost, Gradient Boosting ve XGBoost yükseltme topluluk yöntemleri uygulanmıştır.

Adaboost topluluk yöntemi, ilk olarak Yoav Freund ve Robert Schapire (1995) tarafından önerilmiştir. Adaboost; temelde, karar ağacının ikili sınıflandırma problemleri üzerindeki

performansını artırmak için kullanılmaktadır. Aynı eğitim seti için farklı öğrenme algoritmaları eğitmekte ve daha sonra bu zayıf öğrenme algoritmalarını birleştirerek daha güçlü bir nihai öğrenme algoritması oluşturmakta olan yinelemeli bir algoritmadır. Her yeni tahmin edici için ağırlık örneğini ayarlamaktadır. Adaboost topluluk yönteminde, yanlış sınıflandırılmış veri noktalarının ağırlığı arttırılarak hatalardan öğrenme gerçekleştirilmektedir.

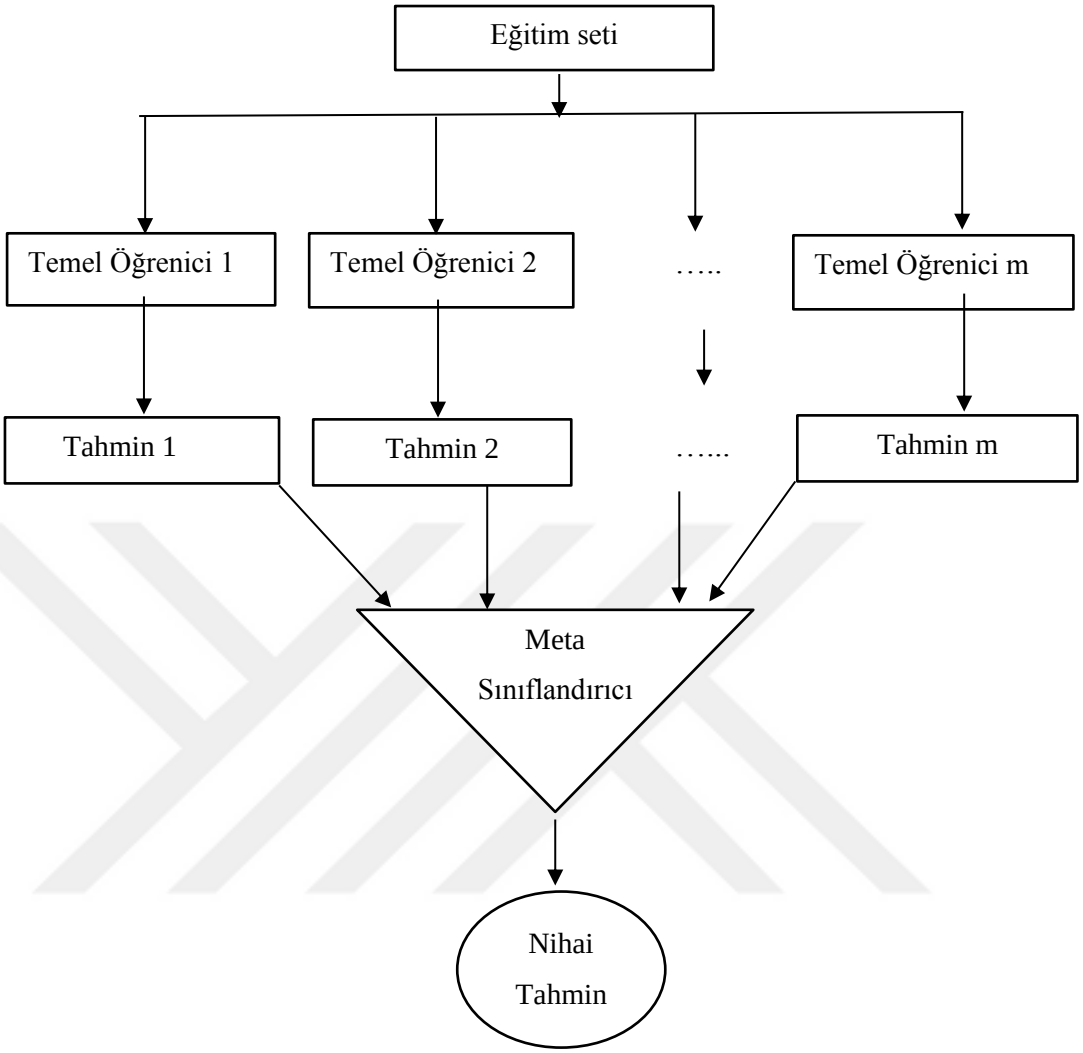
Gradient Boosting, Friedman (2001) tarafından önerilmiştir. Yükseltme kavramını, Adaboost yönteminin regresyon problemleri ve çoklu kayıp fonksiyonları ile genelleştirmesine izin veren önemli istatistiksel kavramlarla ilişkilendirmiştir. Gradient boosting yöntemi ile kayıp minimuma indirilmektedir. Gradyan descent kullanılarak ve tahminler, bir öğrenme oranına göre güncellenerek kayıp fonksiyonunun minimum olduğu değerler bulunabilmektedir.

XGBoost topluluk yöntemi; belirli kayıp işlevlerini optimize eden, Gradient Boosted Trees algoritmasının en popüler uygulamalarından biridir. Artırılmış eğimli karar ağaçlarından yararlanarak daha yüksek hız ve performans sağlamaktadır.

Bu tez çalışmasında; SVC, Gaussian Naive Bayes ve Karar Ağacı bireysel sınıflandırıcıları ile Adaboost(SVC), Adaboost(GaussianNB), Adaboost(Karar Ağacı) olmak üzere ayrı ayrı torbalama sınıflandırıcıları oluşturulmuştur. Gradient Boosting ve XGBoost yüksetme topluluk yöntemleri, temel öğrenici olarak karar ağacı algoritmasını kullanmıştır.

3.6.5. Yığma

Yığma; bireysel sınıflandırma modellerinden gelen tahminleri bir meta sınıflandırıcı aracılığıyla birleştiren bir topluluk öğrenmesi tekniğidir. Yığma, bireysel sınıfların performansını iyileştirmek için tasarlanmıştır. Bir yığma modelinin yapısı; seviye 0 modelleri olarak adlandırılan iki veya daha fazla temel öğrenici ve seviye-1 modeli olarak adlandırılan bir meta modelden oluşmaktadır. Meta model, temel modellerin tahminlerini verimli bir şekilde birleştirmek amacıyla eğitilmektedir. Yığma topluluk yönteminin çalışma prensibi, Şekil 3.11'de verilmiştir.



Şekil 3.11: Yığma topluluk yöntemi çalışma prensibi.

Yığma topluluk yöntemi, genellikle farklı öğrenme algoritmalarından oluşmaktadır. Bu nedenle; yığma toplulukları, genellikle heterojendir. Hem denetimli öğrenme görevlerinde hem de denetimsiz öğrenme görevlerinde kullanılmaktadır. Bir temel öğrenici; eğer yığma topluluğu kadar veya yığma topluluğundan daha iyi performans elde gösteriyorsa, topluluk modellerine kıyasla daha düşük karmaşıklığa sahip olduğu için yığma topluluğu yerine kullanılabilmektedir.

3.6.6. Ekstra Ağaçlar

Ekstra Ağaçlar topluluk yöntemi, karar ağaçlarına dayanan bir topluluk öğrenme yöntemidir. Bir ormanda toplanmış birden fazla ilişkisiz karar ağacının tahminlerini birleştirerek nihai tahmini elde eden bir topluluk öğrenme yöntemidir. Ekstra Ağaçlar, verilerden aşırı

öğrenmeyi en aza indirmek için belirli kararları ve veri alt kümelerini rastgele hale getirmesi bakımından Rassal Orman yöntemine benzemektedir. Topluluğun üyeleri olarak kullanılan karar ağaçlarını oluşturmada, Rassal Orman yöntemine kıyasla daha basit bir algoritma kullanmasına rağmen, Rastgele orman algoritmasından daha iyi performans elde edilebilmektedir.

Her bir karar ağacının eğitim veri kümesinin bir bootstrap örneğinden geliştirildiği rassal orman yönteminin aksine, Ekstra Ağaçlar yönteminde her bir karar ağacı orijinal eğitim örneğinden oluşturulmaktadır. Ekstra Ağaçlar algoritması, rastgele bir bölme noktası seçerek düğümleri dallara ayırmaktadır (Geurts vd., 2006). Nihai tahmin, regresyon durumunda karar ağaçlarının tahminlerinin ortalamasını alarak veya sınıflandırma durumunda karar ağaçlarının çoğunluk oyu alınarak elde edilmektedir. Bu yöntem genellikle, sapmanın artmasına neden olurken modelin varyansının azalmasını sağlamaktadır.

3.7. MATERYAL

Bu çalışmada, Wisconsin Diagnostic Breast Cancer veri kümesi kullanılmıştır. Uygulama, Colab Notebook ortamında gerçeklenmiştir. Sklearn kütüphanesinden yararlanılmıştır. Materyal başlığı altında; kullanılan veri kümesi, Colab Notebook ve Sklearn kütüphanesi incelenmiştir.

3.7.1. Veri Kümesi

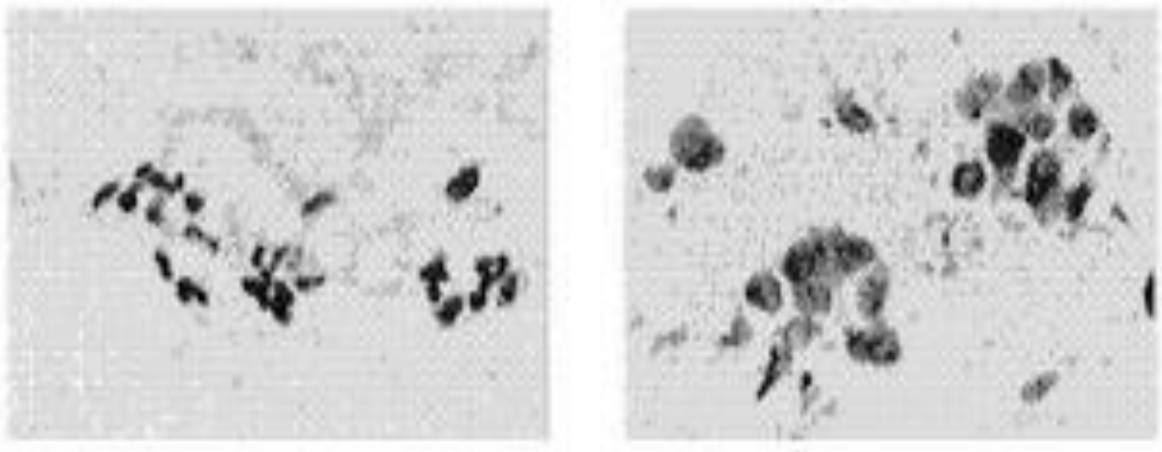
Bu tez çalışmasında, verilerin eğitim ve test aşaması için Wisconsin Diagnostic Breast Cancer (WDBC) veri kümesi kullanılmıştır. Wisconsin Diagnostic Breast Cancer veri kümesi, Wisconsin Üniversitesi Hastanesi'nde doktor olan Dr. William H. Wolberg tarafından oluşturulmuştur. Veri kümesi, 212 kötü huylu ve 357 iyi huylu olmak üzere 569 örnekten oluşmaktadır. Wisconsin Diagnostic Breast Cancer veri kümesi, bir göğüs kitlesinin ince iğne aspirasyon biyopsisinin (FNA) dijitalleştirilmiş bir görüntüsünden hesaplanan özelliklerden oluşmaktadır (Wolberg vd., 1992). Id, tanı sınıfı ve 30 gerçek-değerli özellikten oluşan toplam 32 özellik bilgisi kullanılmıştır.

WDBC veri kümesi özellikleri aşağıda verilmiştir:

- Id
- Tanı sınıfı (M = malignant, B = benign)
- Kanserli hücrelerin her birinden aşağıdaki gibi on özellik hesaplanmıştır:
    1. Yarıçap
    2. Doku
    3. Çevre
    4. Alan
    5. Pürüzsüzlük
    6. Kompaktlık
    7. İçbükeylik
    8. İçbükey Noktalar
    9. Simetri
    10. Fraktal Boyut

10 özelliğin her biri için; ortalama, standart hata ve en kötü (ilgili özelliğin en büyük üç değerinin ortalaması) olmak üzere 3 değer hesaplanmıştır. Böylece toplam 30 adet özellik elde edilmiştir.

Şekil 3.12, Wolberg ve diğ. (1995) tarafından göğsün ince iğne biyopsilerinden alınan iyi huylu ve kötü huylu tümöre ait iki görüntüyü göstermektedir.



Şekil 3.12: FNA'nın sayısallaştırılmış görüntüleri: (a) İyi huylu, (b) Kötü huylu.

3.7.2. Colab Notebook

Colab Notebook, Google Drive ile entegre olmuş Jupyter not defterleridir. Jupyter Notebook Uygulaması, bir sunucu-istemci uygulamasıdır. Jupyter adı; desteklediği Julia, Python ve R programlama dillerinden gelmektedir. Colab Notebook; kod, veri görselleştirmeleri, matematiksel denklemler ve anlatım metni tek bir belgede birleştirmek için kullanılabilen ve bu belgeyi paylaşabilmeye olanak sağlayan ücretsiz, açık kaynaklı bir web uygulamasıdır. Kodu görüntülemeye, çalıştırabilmeye ve sonuçları doğrudan web tarayıcıda görüntüleyebilmeye olanak sağlamaktadır.

Colab Notebook; statik değildir, doğrudan tarayıcıda sağlanan geri bildirimlerle gerçek zamanlı olarak düzenlenebilir ve yeniden çalıştırılabilir. Yürütülen kodun sonuçları, kod bloklarından hemen sonra görünmektedir. Kod blokları; herhangi bir sırada, istenilen sıklıkta yeniden çalıştırılabilmektedir. Matplotlib modülü aracılığıyla grafikler oluşturulabilmektedir. NumPy gibi Python sayısal ve bilimsel kütüphanelerini desteklemektedir.

3.7.3. Sklearn

Sklearn kütüphanesi, Scikit-learn adıyla da bilinmektedir. Sklearn, en yaygın kullanılan makine öğrenmesi kütüphanelerinden biridir. Scikit-learn projesi; 2007 yılında David Cournapeau tarafından, bir Google Summer of Code projesi olarak scikits.learn adıyla geliştirilmiştir.

Sklearn, Python programlama dili için ücretsiz, açık kaynak kodlu bir makine öğrenmesi kütüphanesidir (Pedregosa vd., 2011). Sınıflandırma, regresyon, kümeleme ve boyut indirgeme gibi makine öğrenmesi için birçok araç içermektedir. Birçok gözetimli ve gözetimsiz öğrenme algoritması sağlamaktadır. Sklearn kütüphanesi; destek vektör makinesi, k-en yakın komşu, saf bayes, karar ağacı ve rassal orman gibi çeşitli algoritmalara sahiptir. NumPy, SciPy ve Matplotlib gibi sayısal ve bilimsel kütüphaneleri desteklemektedir.

3.8. MODEL BAŞARIM ÖLÇÜTLERİ

Bu çalışmada; doğruluk(accuracy), kesinlik(precision), duyarlılık(recall), f-skor, ROC grafiğinin altındaki alan (AUC) ve ROC model başarım ölçütleri kullanılmıştır.

Hata matrisi, sınıflandırma problemlerini çözerken kullanılan bir performans ölçüsüdür. Matris düzeni, bir algoritmanın performansını görselleştirmeye yardımcı olmaktadır. Matrisin her satırı gerçek bir sınıftaki örnekleri temsil ederken, matristeki her sütun tahmin edilen bir sınıftaki örnekleri temsil etmektedir. İkili sınıflandırma için bir hata matrisi örneği Şekil 3.13'te gösterilmektedir.

		Tahmin Edilen Değer	
		Negative	Positive
Gerçek Değer	Negative	TN	FP
	Positive	FN	TP

Şekil 3.13: Hata matrisi.

Hata matrisinde yer alan TP, FP, TN ve FN aşağıdaki şekilde ifade edilmektedir:

- Gerçek pozitif (TP): Pozitif tahmin edilen ve pozitif gerçek değer elde edilen örneklerin sayısı.
- Yanlış pozitif (FP): Pozitif tahmin edilen ancak negatif gerçek değer elde edilen örneklerin sayısı.
- Gerçek negatif (TN): Negatif tahmin edilen ve negatif gerçek değer elde edilen örneklerin sayısı.
- Yanlış negatif (FN): Negatif tahmin edilen ancak pozitif gerçek değer elde edilen örneklerin sayısı.

3.8.1. Doğruluk

Doğruluk, sınıflandırıcının vakaları kategorilerine göre ne kadar iyi tahmin edebildiğinin bir ölçüsüdür. Modelin tüm sınıflarda nasıl bir performans gösterdiğini açıklayan bir metriktir. Tüm sınıflar eşit öneme sahip olduğunda kullanışlıdır. Doğru tahmin sayısının, veri kümesindeki toplam örnek sayısına bölünmesiyle elde edilen orandır. Doğruluk, aşağıdaki denklem kullanılarak hesaplanabilmektedir:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.17}$$

3.8.2. Kesinlik

Kesinlik, modelin bir örneği pozitif olarak sınıflandırmadaki doğruluğunu ölçmektedir. Gerçek pozitiflerin sayısının, gerçek pozitiflerin sayısı ile yanlış pozitiflerin sayısının toplamına bölünmesiyle elde edilmektedir. Kesinlik, aşağıdaki denklem kullanılarak hesaplanabilmektedir:

$$Precision = \frac{TP}{TP + FP} \tag{3.18}$$

3.8.3. Duyarlılık

Duyarlılık; gerçek pozitiflerin sayısının, gerçek pozitiflerin sayısı ile yanlış negatiflerin sayısının toplamına bölünmesiyle elde edilmektedir. Modelin tahmin ettiği yanlış pozitiflerin sayısına bakmak yerine yanlış negatiflerin sayısına bakmaktadır. Duyarlılık, aşağıdaki denklem kullanılarak hesaplanabilmektedir:

$$Recall = \frac{TP}{TP + FN} \tag{3.19}$$

3.8.4. F-Skor

Optimum bir kesinlik ve duyarlılık karışımını bulmak istediğimiz durumlarda, F1 skoru ölçütü kullanılarak iki ölçüm birleştirilebilmektedir. F1 skoru, kesinlik ve duyarlılığın harmonik ortalamasıdır. F-skor, aşağıdaki denklem kullanılarak hesaplanabilmektedir:

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.20}$$

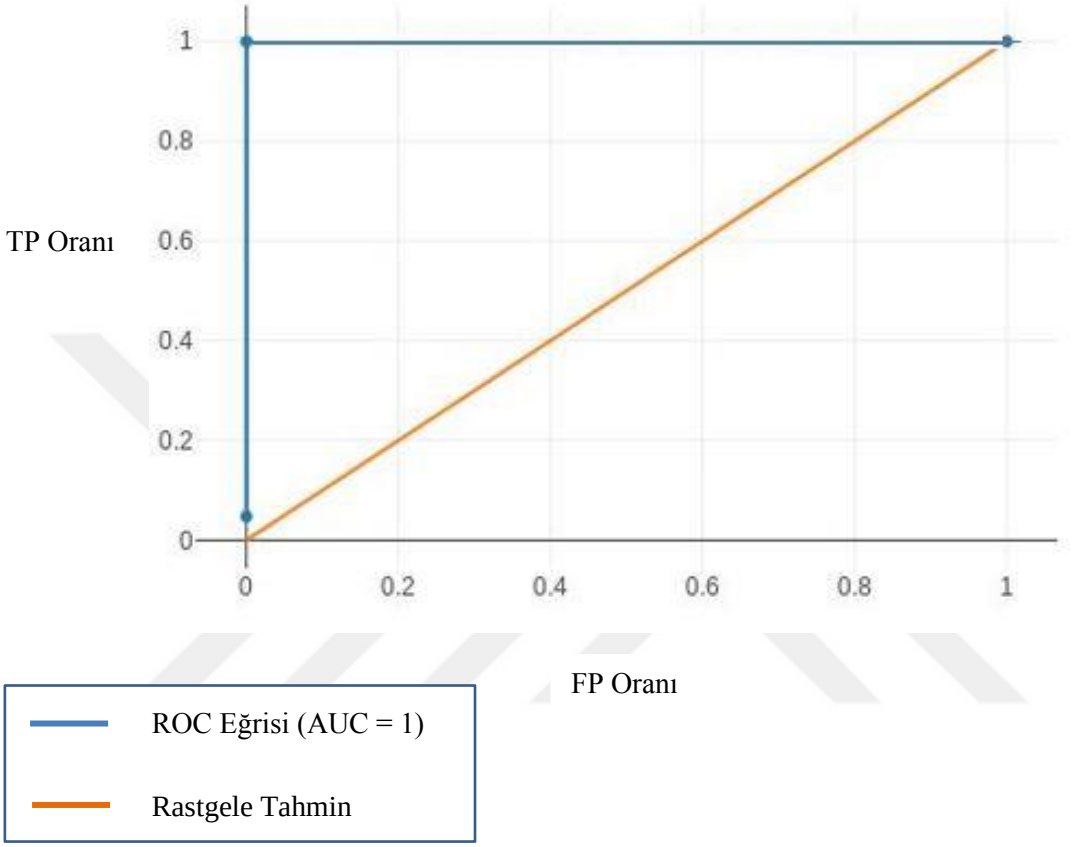
3.8.5. Roc

Bir sınıflandırma modelinin performansını göstermek için kullanılan görselleştirme tekniklerinden biri, İşlem Karakteristik Eğrisi (ROC) yöntemidir. ROC grafikleri, TP oranının Y ekseninde ve FP oranının X ekseninde çizildiği iki boyutlu grafiklerdir. TP ve FP oranları şu şekilde hesaplanmaktadır:

$$TP\ oranı = \frac{TP}{TP+FN} \tag{3.21}$$

$$FP\ oranı = \frac{FP}{FP+TN} \tag{3.22}$$

ROC eğrisi grafiği Şekil 3.14'te verilmiştir. Şekil 3.14'te ROC eğrisi mavi çizgi ile gösterilmiştir. İyi bir tahmin için ROC eğrisi, grafiğin sol üst köşesine yakındır. Kötü bir tahmin için ROC eğrisi, turuncu çizgiye (y = x) yakındır.



Şekil 3.14: ROC eğrisi.

3.8.6. Auc

AUC, ROC eğrisi altındaki alan anlamına gelmektedir. AUC, (0,0) 'dan (1,1)' e kadar tüm ROC eğrisinin altındaki iki boyutlu alanın tamamını ölçmektedir. AUC değerleri, [0, 1] aralığında yer almaktadır. Bununla birlikte elde edilen AUC değeri, 0.5 değerinin altındaysa ilgili sınıflandırıcı ile kötü bir sonuç elde edildiği söylenebilmektedir. İkili sınıflandırma problemleri için en yaygın kullanılan değerlendirme kriterlerinden biridir.

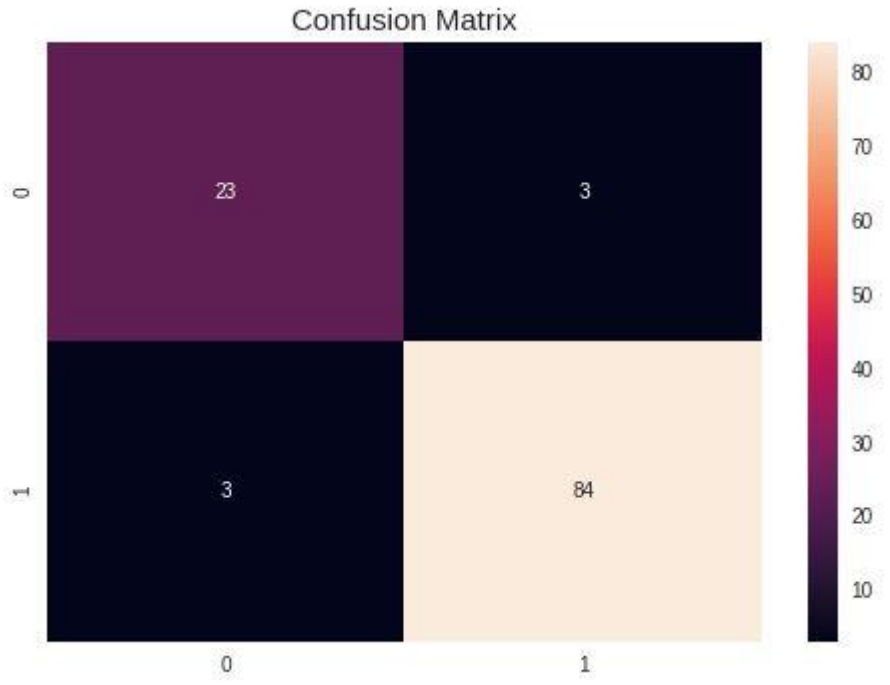
1. BULGULAR

Bu bölümde Destek Vektör Makinesi, K-En Yakın Komşu, Saf Bayes ve Karar Ağaçları bireysel sınıflandırıcılarının model başarım ölçütlerine göre elde edilen kesinlik, duyarlılık, f-skor, doğruluk oranları yer almaktadır. Bireysel sınıflandırıcılarının performansını artırmak için kullanılan Rassal Orman, Oylama, Torbalama, Yükseltme, Yığma ve Ekstra Ağaçlar topluluk yöntemlerinin kesinlik, duyarlılık, f-skor, doğruluk, ROC ve ROC grafiğinin altındaki alan (AUC) değerleri gösterilmiştir.

4.1. BİREYSEL SINIFLANDIRICILARIN MODEL BAŞARIM ORANLARI

4.1.1. Destek Vektör Makinesi

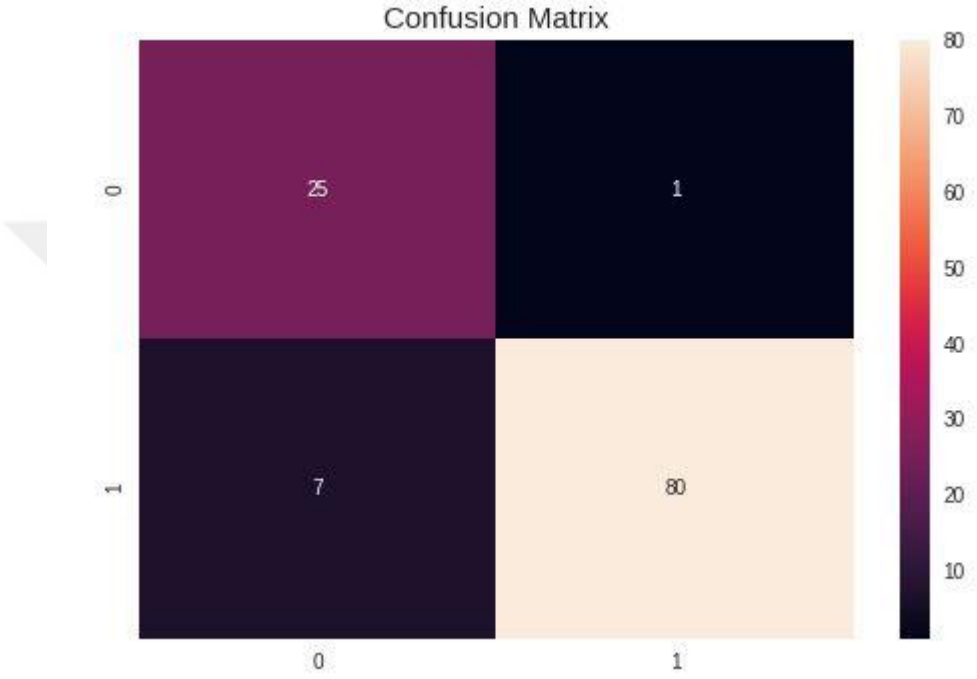
Bu çalışmada, SVC sınıflandırıcısı ile PCA ve Standardizasyon işlemi sonrası; 0.94 kesinlik, 0.95 duyarlılık, 0.95 f-skor ve %93.71 doğruluk elde edilmiştir. SVC sınıflandırıcısı ile PCA ve Standardizasyon işlemi uygulanarak elde edilen hata matrisi, Şekil 4.1'de verilmiştir.



Şekil 4.1: PCA ve Standardizasyon ön işleme işlemi sonrası SVC hata matrisi.

4.1.2. K-En Yakın Komşu

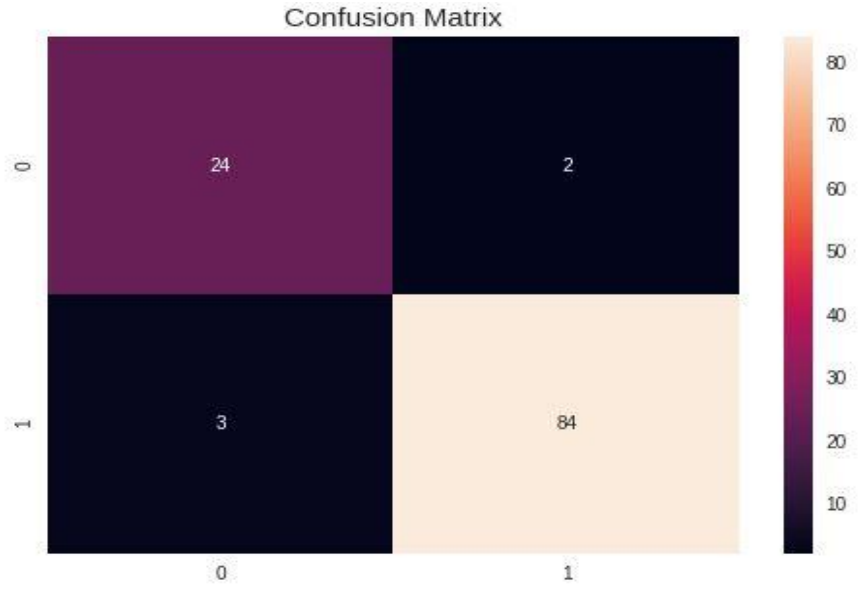
Bu çalışmada, KNN sınıflandırıcısı ile PCA ve Standardizasyon işlemi sonrası; 0.94 kesinlik, 0.93 duyarlılık, 0.94 f-skor ve %92.31 doğruluk elde edilmiştir. KNN sınıflandırıcısı ile; PCA ve Standardizasyon işlemi uygulanarak elde edilen hata matrisi aşağıda yer almaktadır (Şekil 4.2).



Şekil 4.2: PCA ve Standardizasyon ön işleme işlemi sonrası KNN hata matrisi.

4.1.3. Saf Bayes

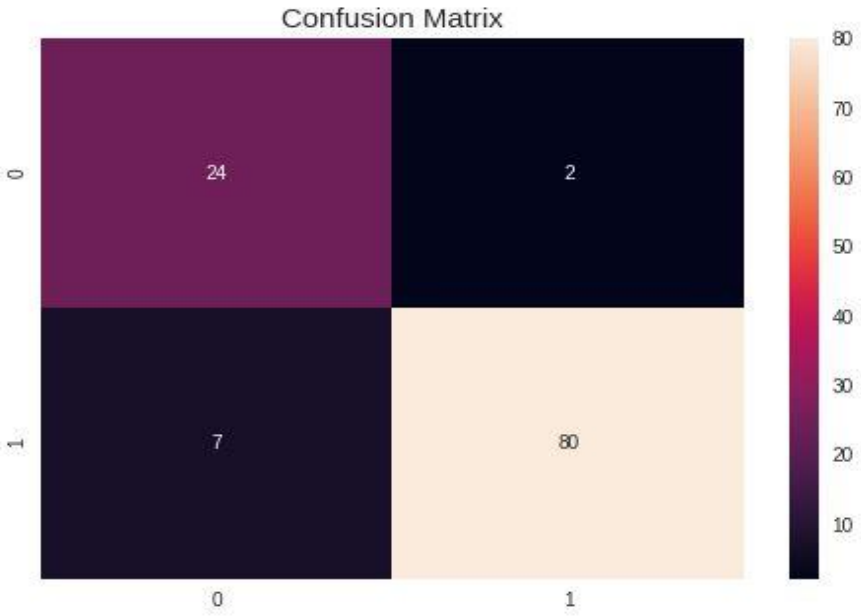
Bu çalışmada, Gaussian Naive Bayes sınıflandırıcısı ile PCA ve Standardizasyon işlemi sonrası; 0.91 kesinlik, 0.95 duyarlılık, 0.92 f-skor ve %91.61 doğruluk elde edilmiştir. Gaussian Naive Bayes sınıflandırıcısı ile PCA ve Standardizasyon işlemi uygulanarak elde edilen hata matrisi, Şekil 4.3'te verilmiştir.



Şekil 4.3: PCA ve Standardizasyon ön işleme işlemi sonrası Gaussian Naive Bayes hata matrisi.

4.1.4. Karar Ağacı

Bu çalışmada, Karar Ağacı sınıflandırıcısı ile PCA ve Standardizasyon işlemi sonrası; 0.90 kesinlik, 0.88 duyarlılık, 0.89 f-skor ve %86.01 doğruluk elde edilmiştir. Karar Ağacı sınıflandırıcısı ile PCA ve Standardizasyon işlemi uygulanarak elde edilen hata matrisi, Şekil 4.4'te verilmiştir.

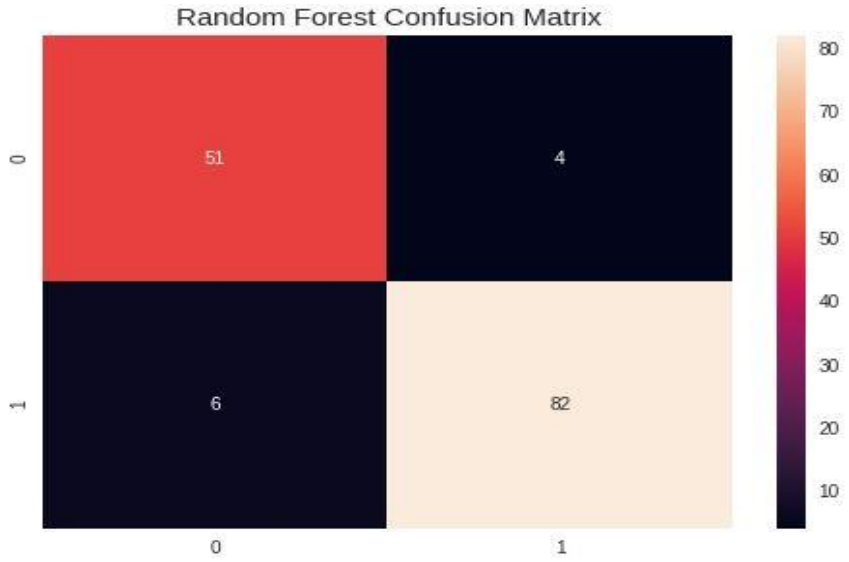


Şekil 4.4: PCA ve Standardizasyon ön işleme işlemi sonrası Karar Ağacı hata matrisi.

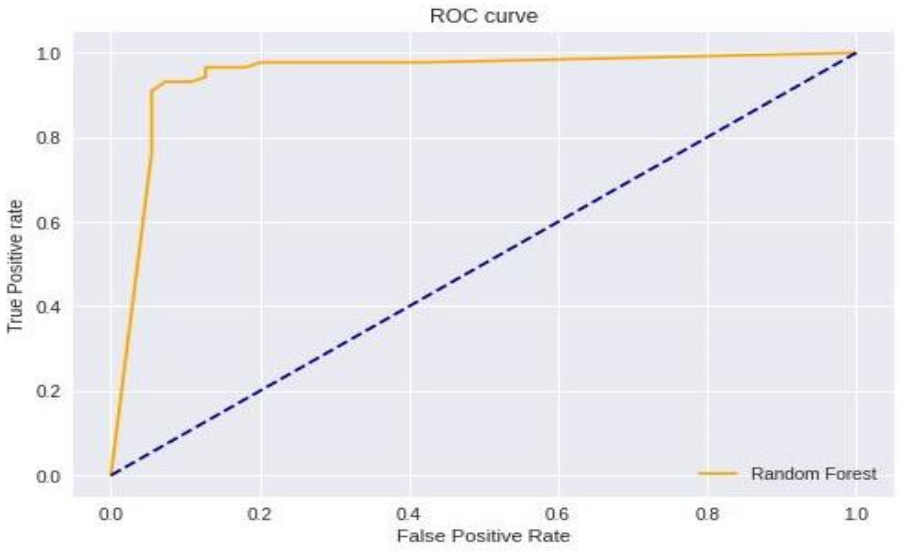
4.2. TOPLULUK YÖNTEMLERİNİN MODEL BAŞARIM ORANLARI

4.2.1. Rassal Orman

Bu çalışmada, Rassal Orman sınıflandırıcısı ile; 0.95 kesinlik, 0.93 duyarlılık, 0.95 f-skor, %93.01 doğruluk, %92.95 AUC skor elde edilmiştir. Şekil 4.5'te, Rassal Orman topluluk yönteminin hata matrisi verilmiştir. Şekil 4.6'da, Rassal Orman topluluk yönteminin ROC eğrisi gösterilmiştir.



Şekil 4.5: Rassal Orman hata matrisi.



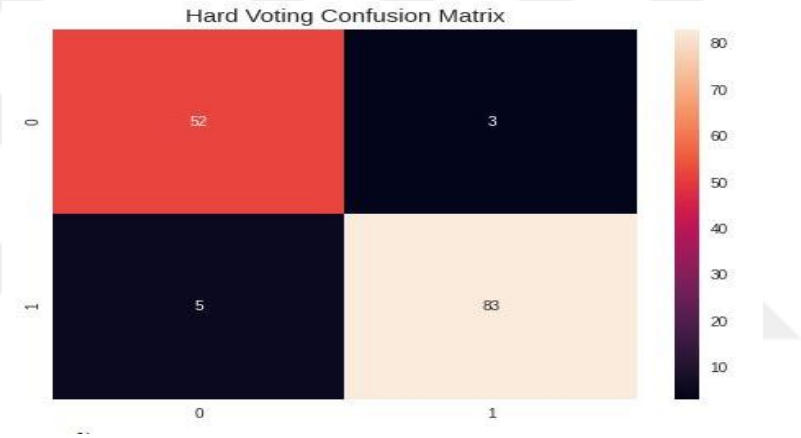
Şekil 4.6: Rassal Orman ROC eğrisi.

4.2.2. Oylama

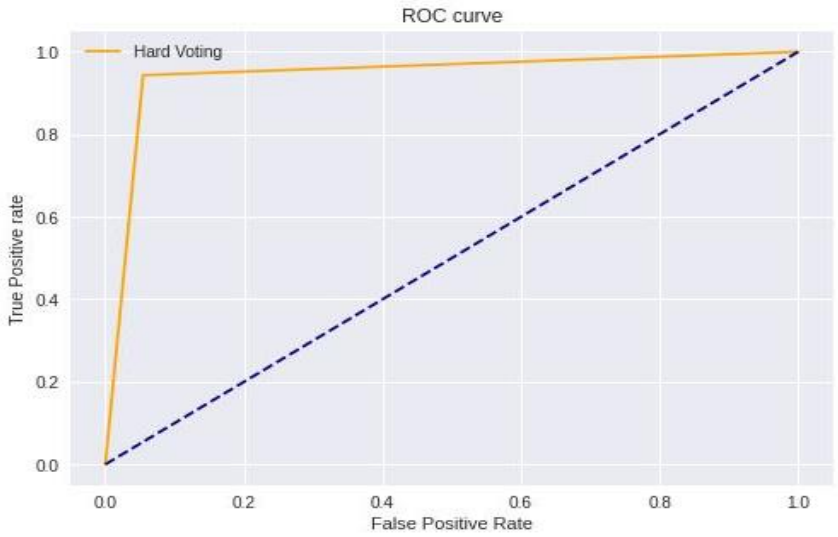
Bu çalışmada; SVC, K-En Yakın Komşuluk, Gaussian Naive Bayes, Karar Ağacı ve ardından sonuçları birleştirmek için Oylama topluluk yöntemleri kullanılmıştır. Çoğunluğa dayalı (hard) ve olasılığa dayalı (soft) oylama yöntemleri kullanılmıştır.

4.2.2.1. Çoğunluğa Dayalı Oylama

Çoğunluğa dayalı oylama için Hard Voting sınıflandırıcısı ile; 0.97 kesinlik, 0.94 duyarlılık, 0.96 f-skor, %94.41 doğruluk, %94.43 AUC skor elde edilmiştir. Şekil 4.7'de, Hard Voting topluluk yönteminin hata matrisi verilmiştir. Hard Voting sınıflandırıcısının ROC eğrisi, Şekil 4.8'de gösterilmiştir.



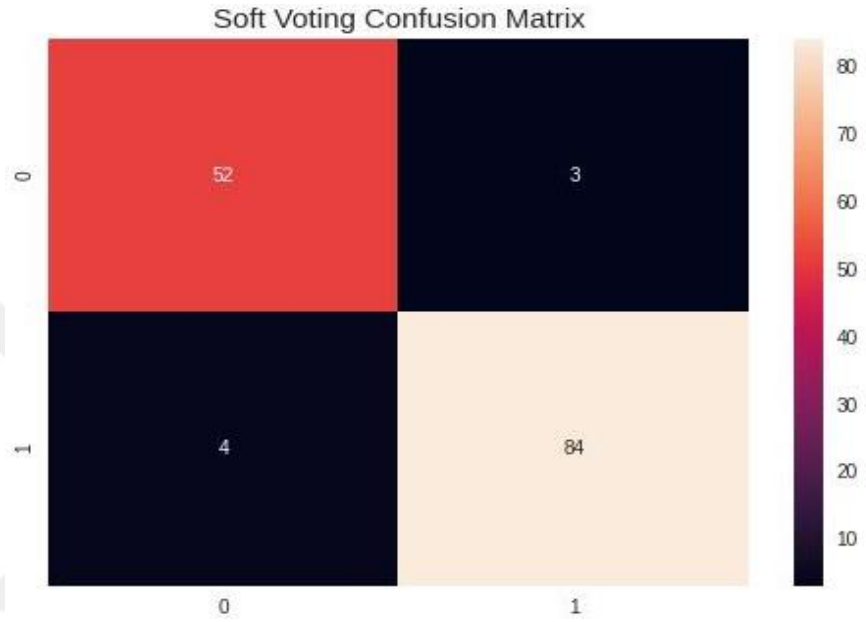
Şekil 4.7: Hard Voting hata matrisi.



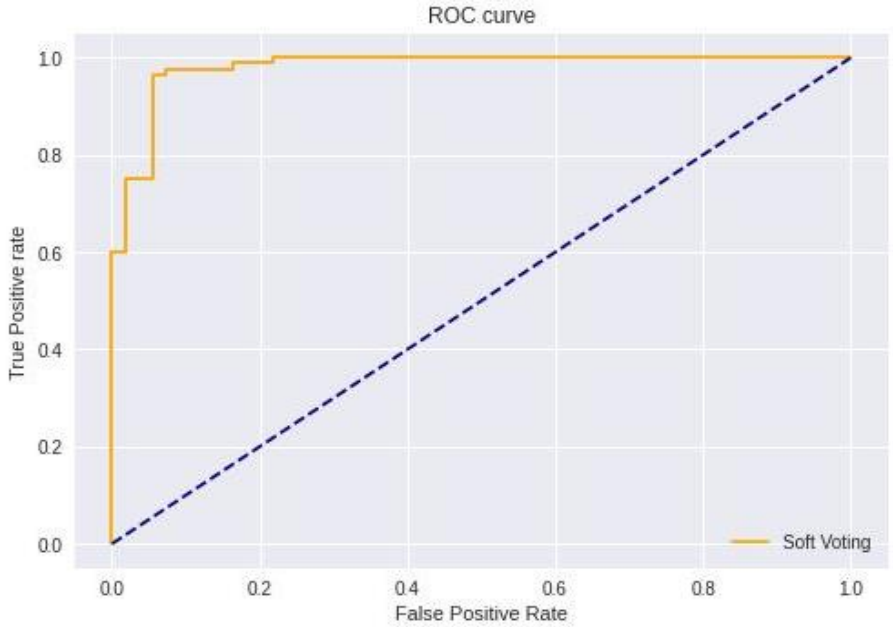
Şekil 4.8: Hard Voting ROC eğrisi.

4.2.2.2. Olasılığa Dayalı Oylama

Olasılığa dayalı oylama için Soft Voting sınıflandırıcısı ile; 0.97 kesinlik, 0.95 duyarlılık, 0.96 f-skor, %95.10 doğruluk, %95 AUC skor elde edilmiştir. Soft Voting topluluk yönteminin hata matrisi, Şekil 4.9'da verilmiştir. Soft Voting topluluk yönteminin ROC eğrisi, Şekil 4.10'da verilmiştir.



Şekil 4.9: Soft Voting hata matrisi.



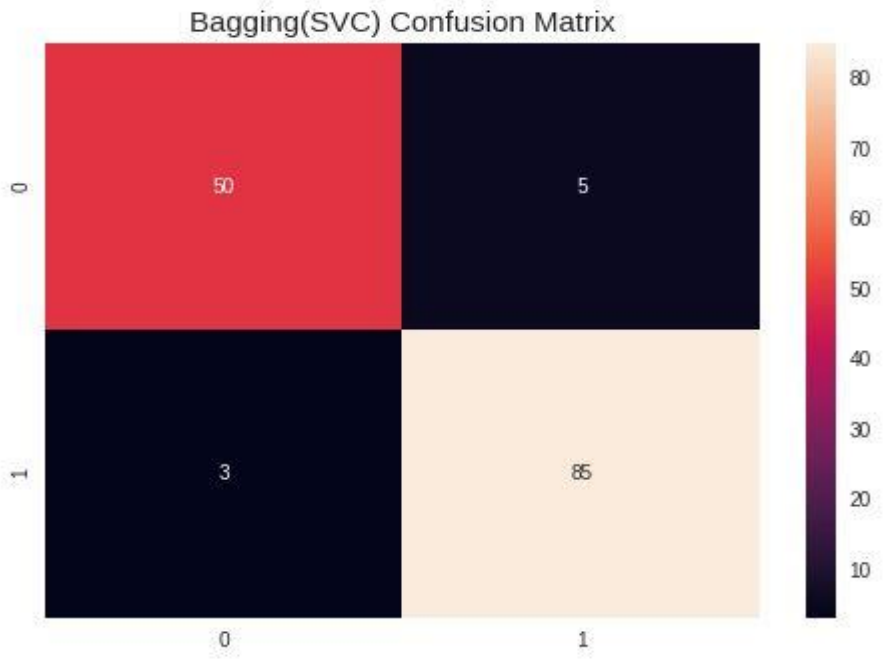
Şekil 4.10: Soft Voting ROC eğrisi.

4.2.3. Torbalama

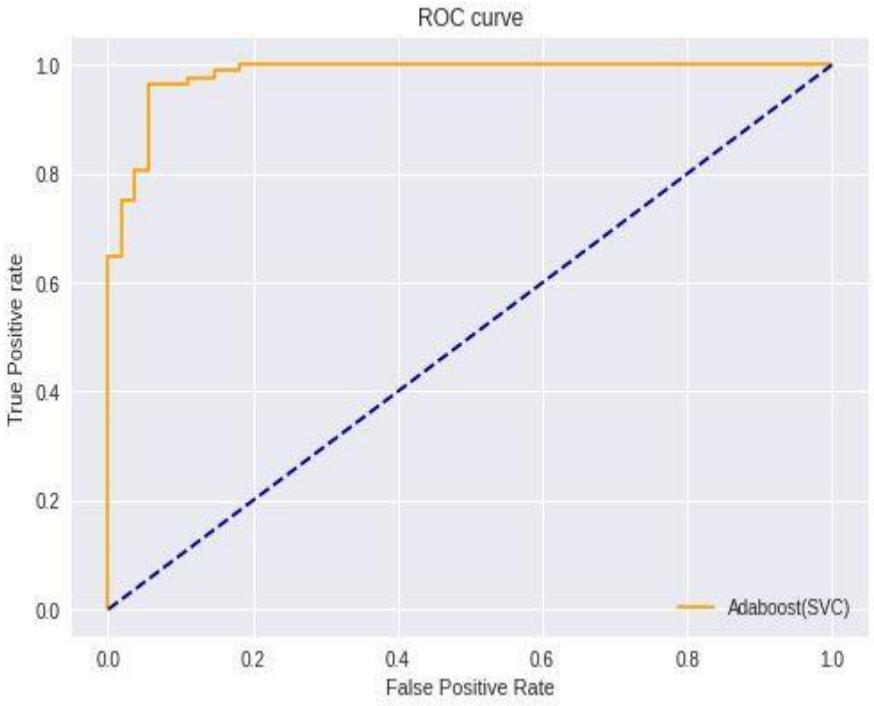
Bu çalışmada; SVC, Gaussian Naive Bayes ve Karar Ağacı bireysel sınıflandırıcıları ile Bagging(SVC), Bagging(GaussianNB), Bagging(Karar Ağacı) olmak üzere ayrı ayrı torbalama sınıflandırıcıları oluşturulmuştur.

4.2.3.1. SVC Temel Öğrenicisi ile Torbalama Yöntemi

Bagging(SVC) sınıflandırıcısı ile; 0.94 kesinlik, 0.97 duyarlılık, 0.95 f-skor, %94.41 doğruluk, %93.75 AUC skor elde edilmiştir. Şekil 4.11'de, Bagging(SVC) topluluk yönteminin hata matrisi verilmiştir. Şekil 4.12'de, Bagging(SVC) topluluk yönteminin ROC eğrisi gösterilmiştir.



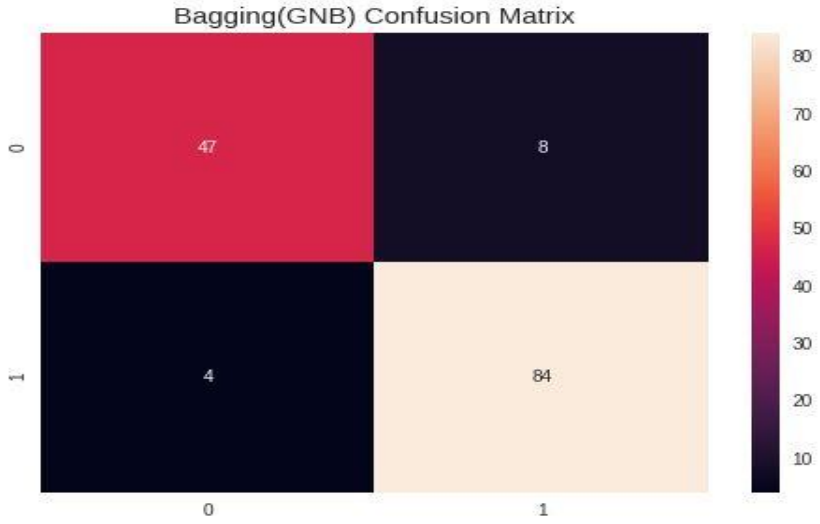
Şekil 4.11: Bagging(SVC) hata matrisi.



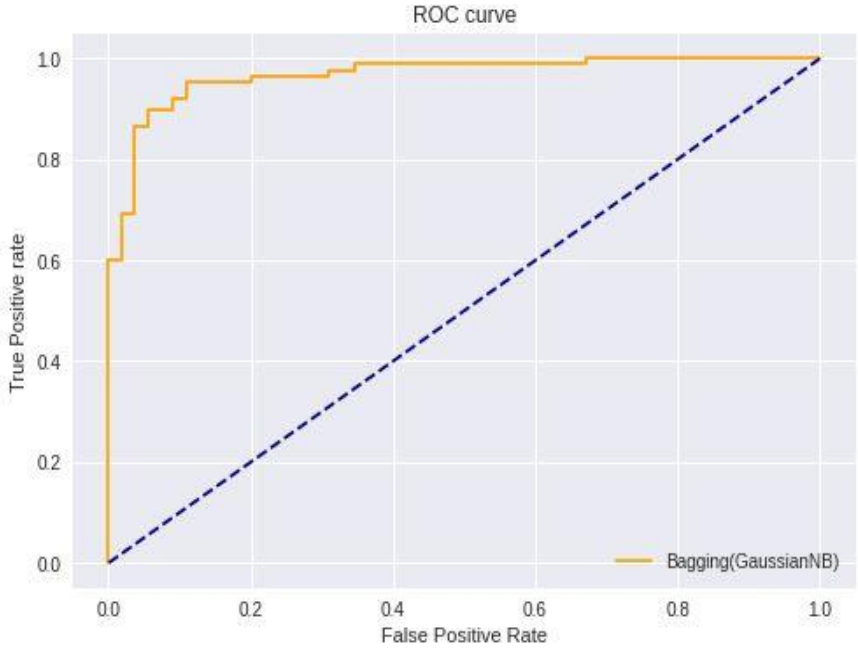
Şekil 4.12: Bagging(SVC) ROC eğrisi.

4.2.3.2. Gaussian Naive Bayes Temel Öğrenicisi ile Torbalama Yöntemi

Bagging(GaussianNB) sınıflandırıcısı ile; 0.91 kesinlik, 0.95 duyarlılık, 0.92 f-skor, %91.61 doğruluk, %90.45 AUC skor elde edilmiştir. Şekil 4.13'te, Bagging(GaussianNB) topluluk yönteminin hata matrisi verilmiştir. Şekil 4.14'te, Bagging(GaussianNB) topluluk yönteminin ROC eğrisi gösterilmiştir.



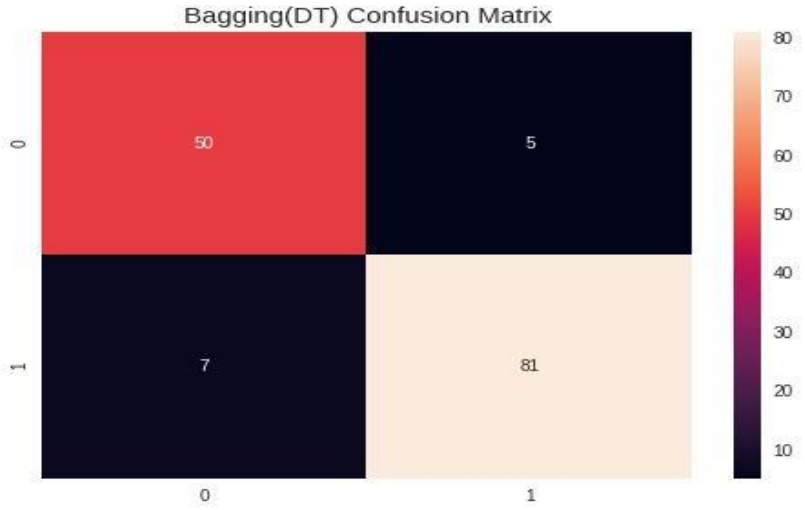
Şekil 4.13: Bagging(GaussianNB) hata matrisi.



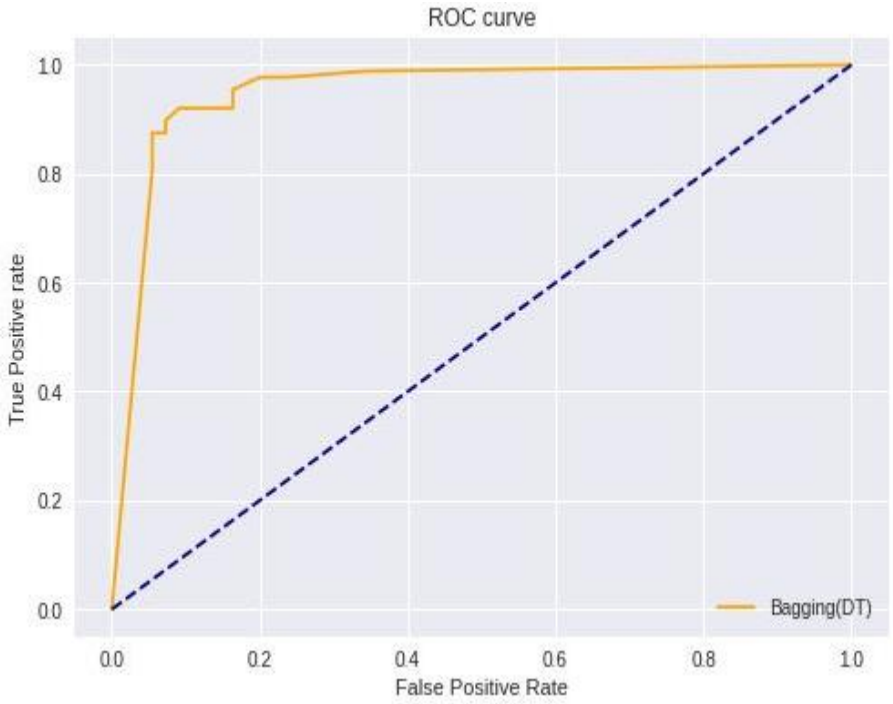
Şekil 4.14: Bagging(GaussianNB) ROC eğrisi.

4.2.3.3. Karar Ağacı Temel Öğrenici ile Torbalama Yöntemi

Bagging(Karar Ağacı) sınıflandırıcısı ile; 0.94 kesinlik, 0.92 duyarlılık, 0.94 f-skor, %91.61 doğruluk, %91.48 AUC skor elde edilmiştir. Şekil 4.15'te, Bagging(Karar Ağacı) topluluk yönteminin hata matrisi verilmiştir. Şekil 4.16'da, Bagging(Karar Ağacı) topluluk yönteminin ROC eğrisi gösterilmiştir.



Şekil 4.15: Bagging(Karar Ağacı) hata matrisi.



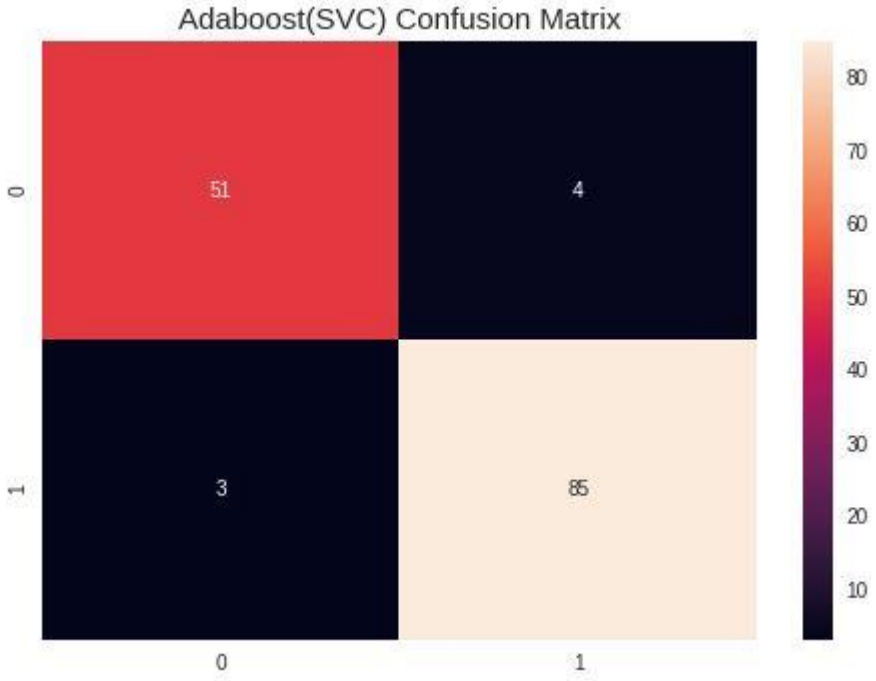
Şekil 4.16: Bagging(Karar Ağacı) ROC eğrisi.

4.2.4. Yükseltme

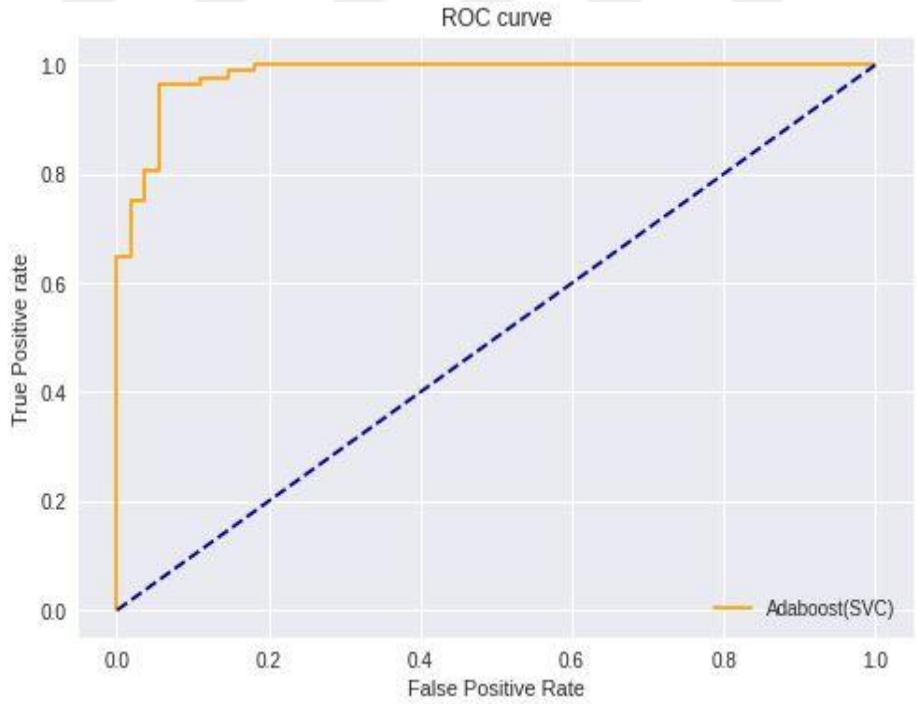
Bu çalışmada; Adaboost, Gradient Boosting ve XGBoost yükseltme topluluk yöntemleri uygulanmıştır. SVC, Gaussian Naive Bayes ve Karar Ağacı bireysel sınıflandırıcıları ile Adaboost(SVC), Adaboost(GaussianNB), Adaboost(Karar Ağacı) olmak üzere ayrı ayrı torbalama sınıflandırıcıları oluşturulmuştur. Gradient Boosting ve XGBoost yüksetme topluluk yöntemleri, temel öğrenici olarak karar ağacı algoritmasını kullanmıştır.

4.2.4.1. Adaboost

Adaboost(SVC) sınıflandırıcısı ile; 0.96 kesinlik, 0.97 duyarlılık, 0.96 f-skor, %95.10 doğruluk, %94.66 AUC skor elde edilmiştir. Şekil 4.17'de, Adaboost(SVC) topluluk yönteminin hata matrisi verilmiştir. Şekil 4.18'de, Adaboost(SVC) topluluk yönteminin ROC eğrisi gösterilmiştir.



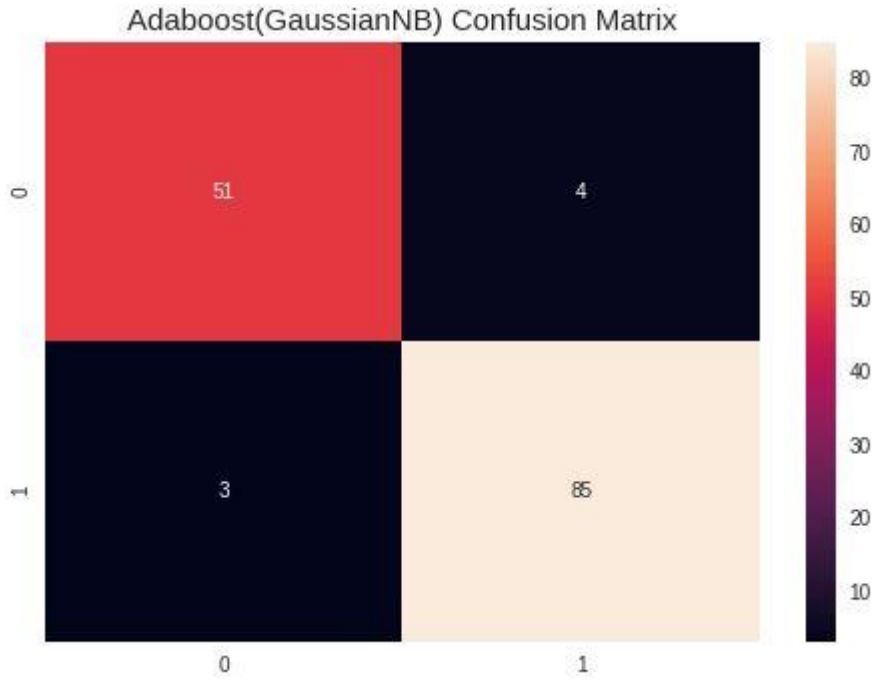
Şekil 4.17: Adaboost(SVC) hata matrisi.



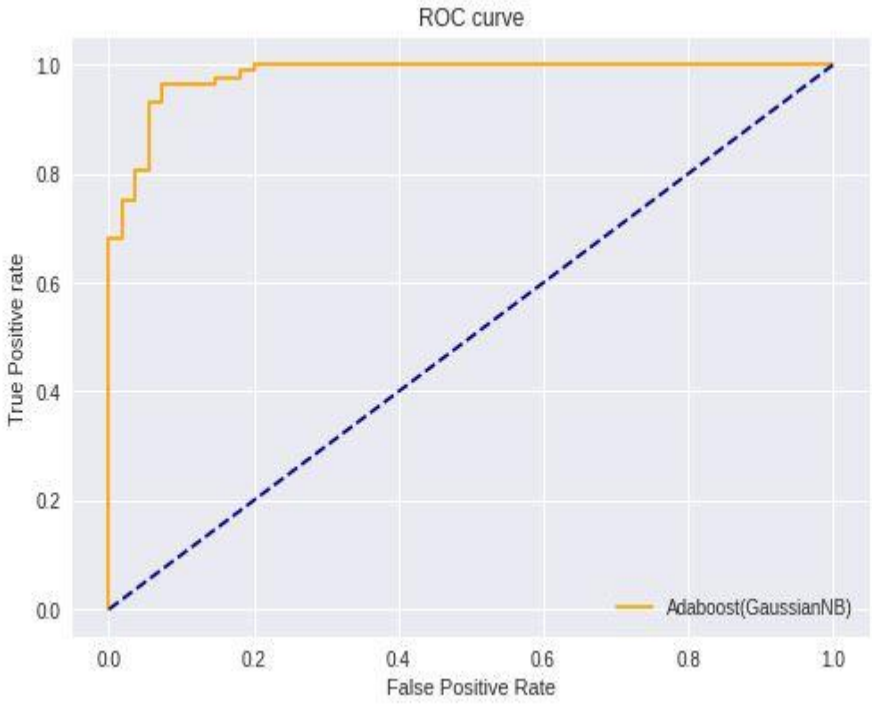
Şekil 4.18: Adaboost(SVC) ROC eğrisi.

Adaboost(GaussianNB) sınıflandırıcısı ile; 0.96 kesinlik, 0.97 duyarlılık, 0.96 f-skor, %95.10 doğruluk, %94.66 AUC skor elde edilmiştir.

Şekil 4.19'da, Adaboost(GaussianNB) topluluk yönteminin hata matrisi verilmiştir. Şekil 4.20'de, Adaboost(GaussianNB) topluluk yönteminin ROC eğrisi gösterilmiştir.

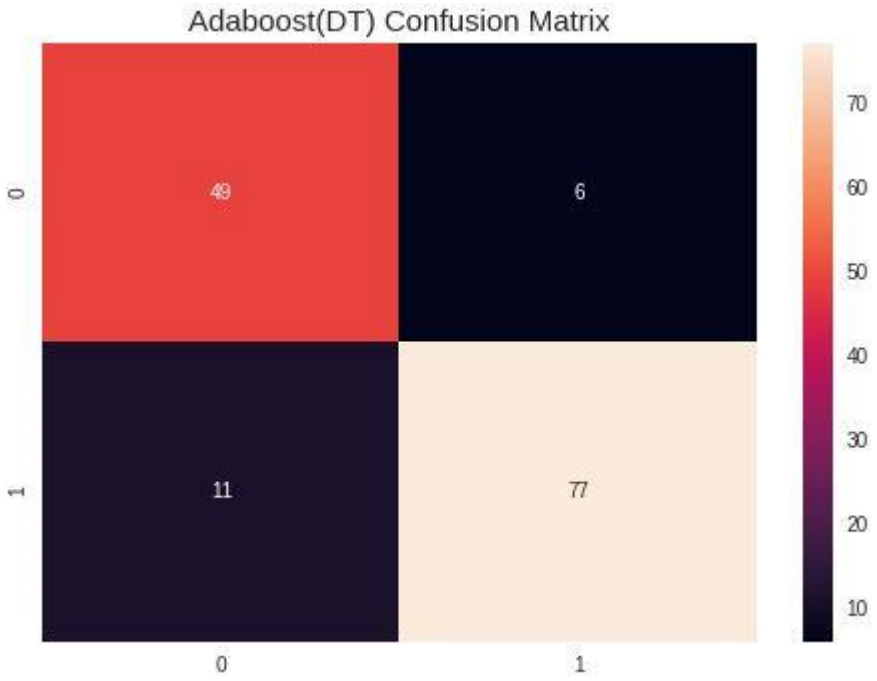


Şekil 4.19: Adaboost(GaussianNB) hata matrisi.

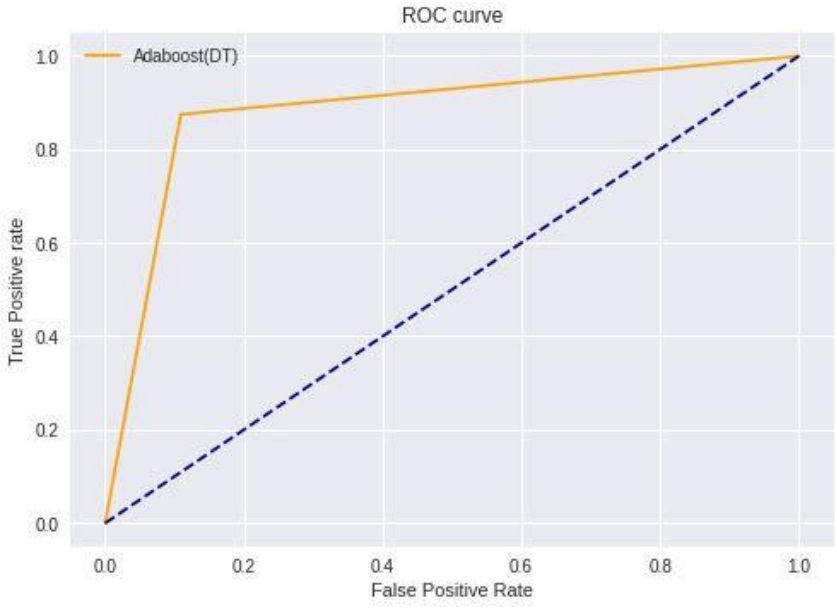


Şekil 4.20: Adaboost(GaussianNB) ROC eğrisi.

Adaboost(Karar Ağacı) sınıflandırıcısı ile; 0.93 kesinlik, 0.88 duyarlılık, 0.92 f-skor, %88.11 doğruluk, %88.30 AUC skor elde edilmiştir. Şekil 4.21'de, Adaboost(Karar Ağacı) topluluk yönteminin hata matrisi verilmiştir. Şekil 4.22'de, Adaboost(Karar Ağacı) topluluk yönteminin ROC eğrisi gösterilmiştir.



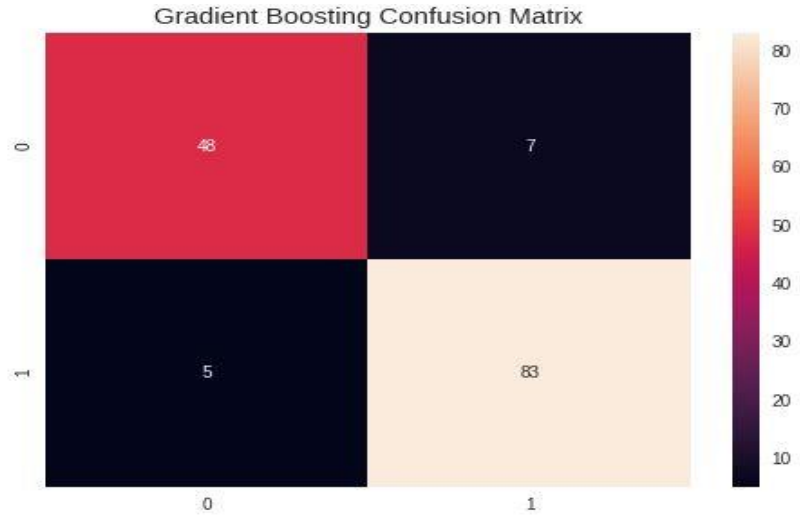
Şekil 4.21: Adaboost(Karar Ağacı) hata matrisi.



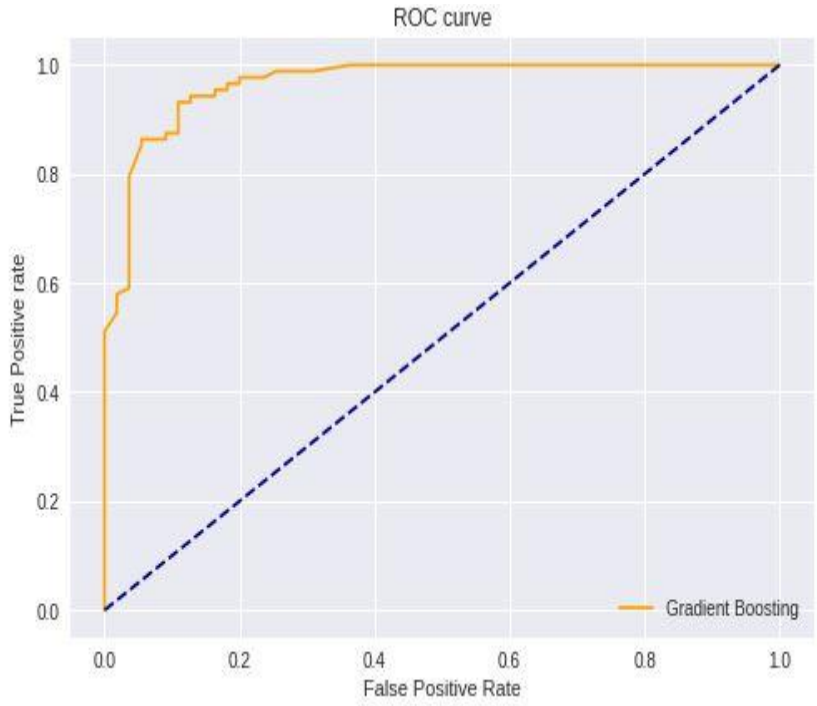
Şekil 4.22: Adaboost(Karar Ağacı) ROC eğrisi.

4.2.4.2. Gradient Boosting

Gradient Boosting sınıflandırıcısı ile; 0.92 kesinlik, 0.94 duyarlılık, 0.93 f-skor, %91.61 doğruluk, %90.80 AUC skor elde edilmiştir. Şekil 4.23'te, Gradient Boosting topluluk yönteminin hata matrisi verilmiştir. Şekil 4.24'te, Gradient Boosting topluluk yönteminin ROC eğrisi gösterilmiştir.



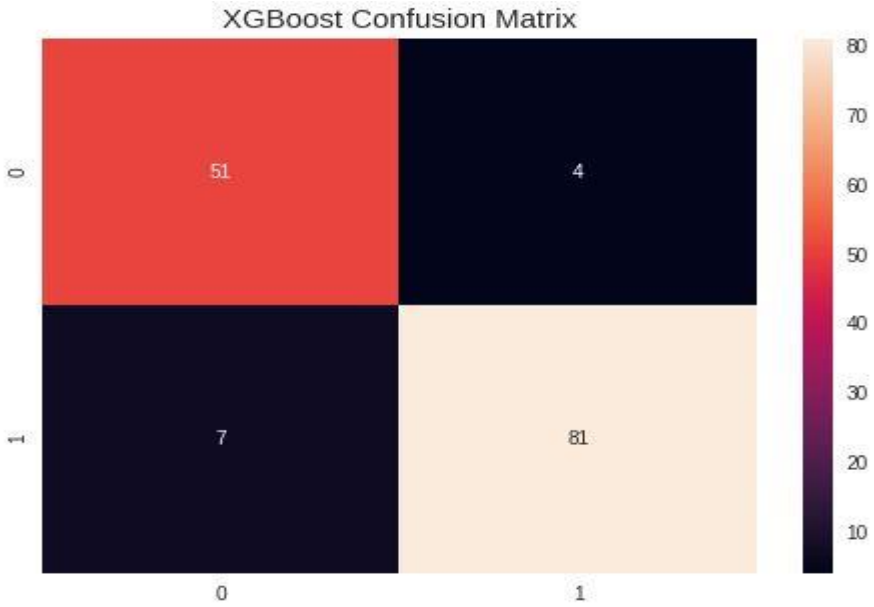
Şekil 4.23: Gradient Boosting hata matrisi.



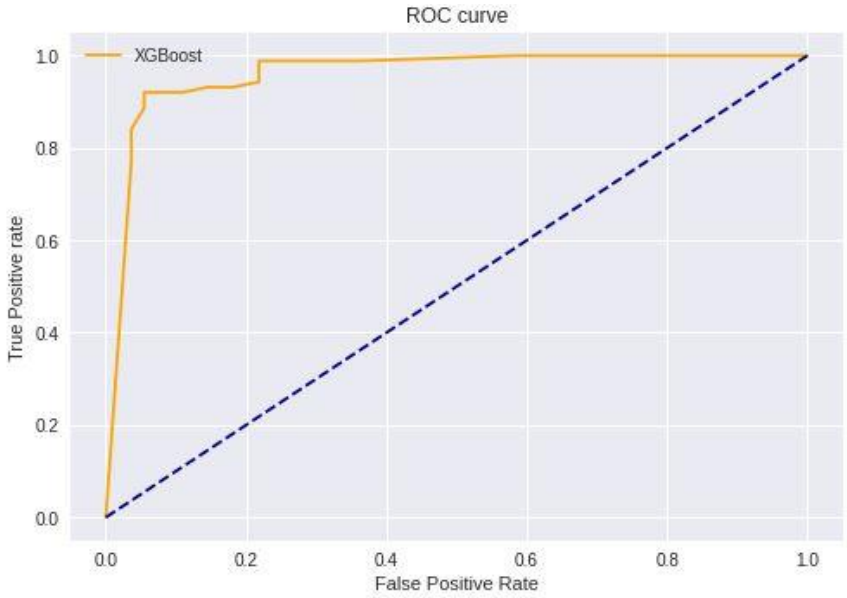
Şekil 4.24: Gradient Boosting ROC eğrisi.

4.2.4.3. XGBoost

XGBoost sınıflandırıcısı ile; 0.95 kesinlik, 0.92 duyarlılık, 0.95 f-skor, %92.31 doğruluk, %92.39 AUC skor elde edilmiştir. Şekil 4.25'te, XGBoost topluluk yönteminin hata matrisi verilmiştir. XGBoost topluluk yönteminin ROC eğrisi, Şekil 4.26'da verilmiştir.



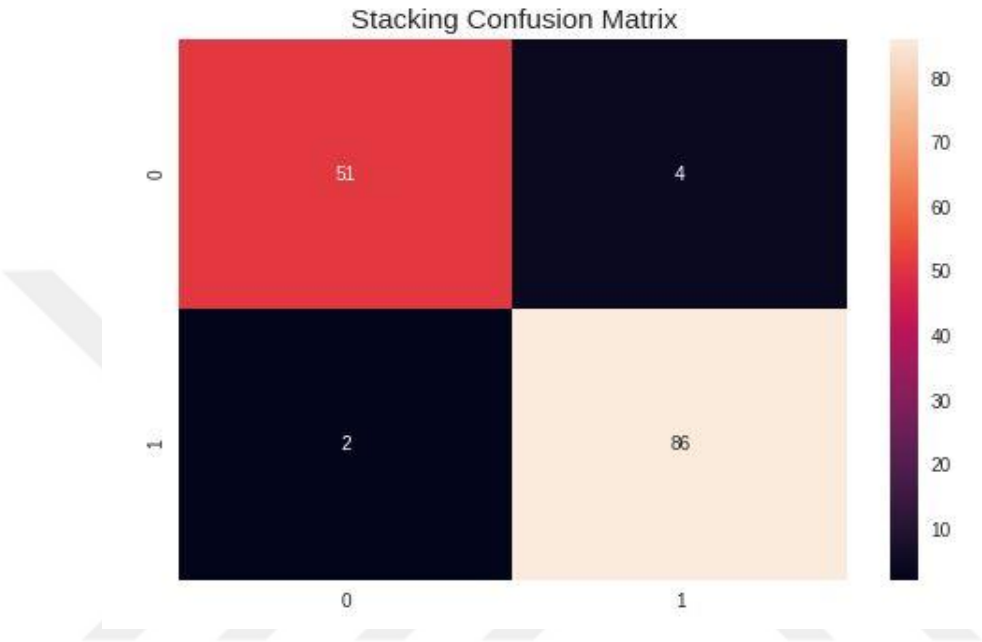
Şekil 4.25: XGBoost hata matrisi.



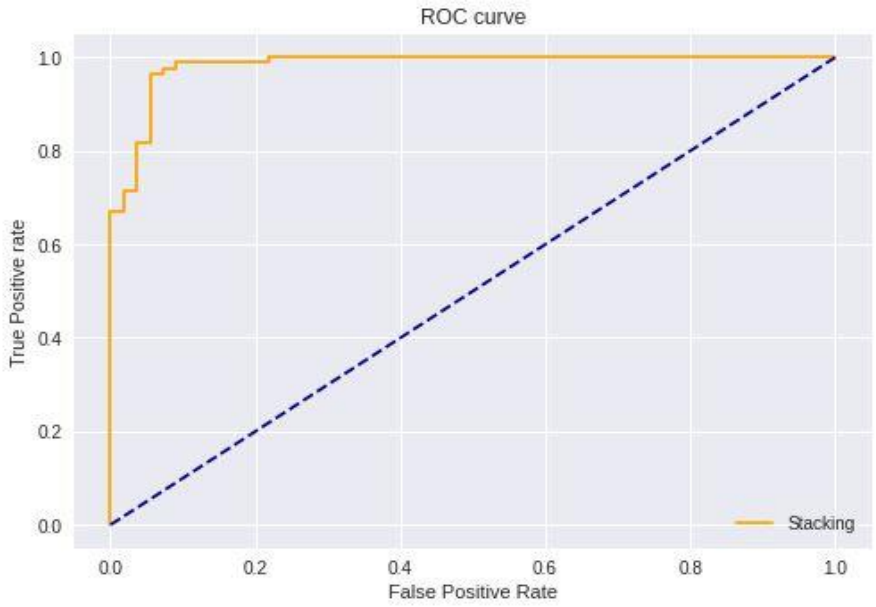
Şekil 4.26: XGBoost ROC eğrisi.

4.2.5. Yığma

Stacking sınıflandırıcısı ile; 0.96 kesinlik, 0.98 duyarlılık, 0.96 f-skor, %95.80 doğruluk, %95.23 AUC skor elde edilmiştir. Şekil 4.27'de, Stacking topluluk yönteminin hata matrisi verilmiştir. Stacking topluluk yönteminin ROC eğrisi, Şekil 4.28'de gösterilmiştir.



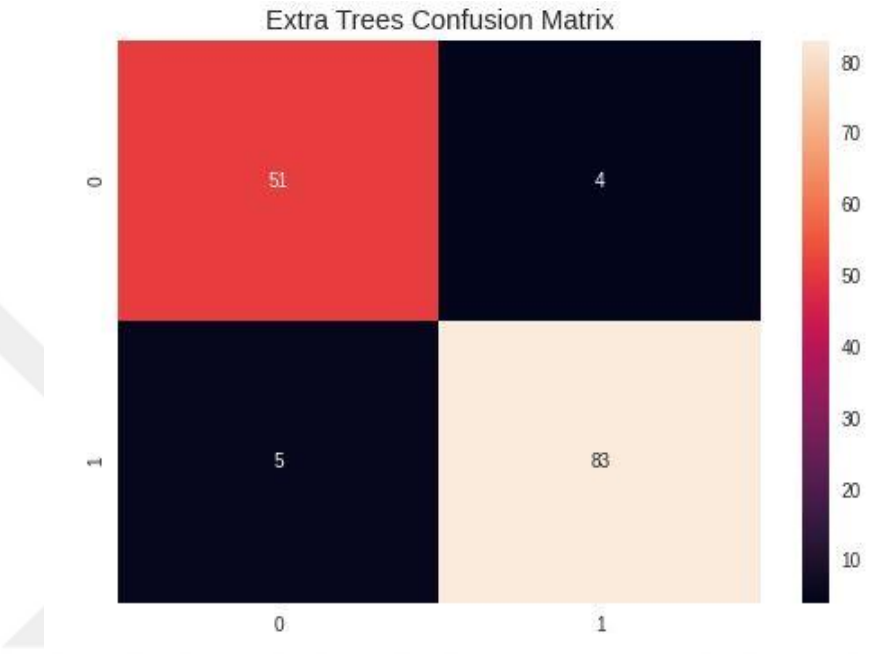
Şekil 4.27: Stacking hata matrisi.



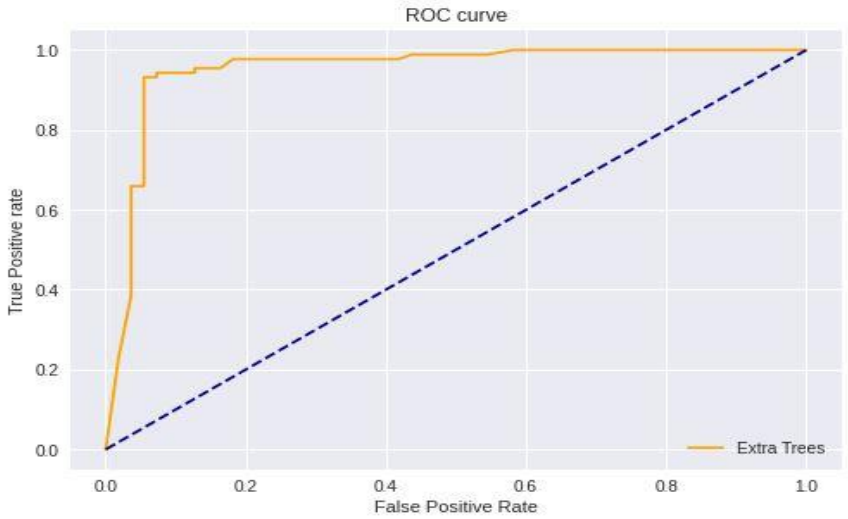
Şekil 4.28: Stacking ROC eğrisi.

4.2.6. Ekstra Ağaçlar

Extra Trees sınıflandırıcısı ile; 0.95 kesinlik, 0.94 duyarlılık, 0.95 f-skor, %93.71 doğruluk, %93.52 AUC skor elde edilmiştir. Extra Trees topluluk yönteminin hata matrisi, Şekil 4.29'da verilmiştir. Extra Trees topluluk yönteminin ROC eğrisi, Şekil 4.30'da gösterilmiştir.



Şekil 4.29: Extra Trees hata matrisi.



Şekil 4.30: Extra Trees ROC eğrisi.

4.3. MODELLERİN PERFORMANS KARŞILAŞTIRMALARI

SVC, KNN, Gaussian Naive Bayes ve Karar Ağaçları bireysel sınıflandırıcılarının; veri kümesine Standardizasyon ve PCA ön işleme işlemi uygulandığında, model başarım ölçütlerine göre elde edilen kesinlik, duyarlılık, f-skor, doğruluk oranları Tablo 4.1'de yer almaktadır.

Tablo 4.1: Bireysel sınıflandırıcıların Standardizasyon+PCA ön işleme işlemi sonrası model başarım oranları.

Sınıflandırıcı	Kesinlik	Duyarlılık	F-Skor	Doğruluk
SVC	0.94	0.95	0.95	%93.71
KNN	0.94	0.93	0.94	%92.31
Gaussian Naive Bayes	0.91	0.95	0.92	%91.61
Karar Ağacı	0.90	0.88	0.89	%86.01

Tablo 4.1'de yer alan doğruluk yüzdelerine göre; veri kümesine Standardizasyon ve PCA ön işleme işlemi uygulandığında, en yüksek doğruluk yüzdesi %93.71 ile SVC yöntemi tarafından elde edilmiştir. Bu çalışmada uygulanan topluluk yöntemleri, model başarım ölçütlerine göre Tablo 4.2'de karşılaştırılmıştır.

Tablo 4.2: Topluluk öğrenme yöntemlerinin model başarım ölçütlerine göre karşılaştırılması.

Topluluk Yöntemleri	Kesinlik	Duyarlılık	F-Skor	Doğruluk	AUC
Rassal Orman	0.95	0.93	0.95	%93.01	%92.95
Hard Voting	0.97	0.94	0.96	%94.41	%94.43
Soft Voting	0.97	0.95	0.96	%95.10	%95
Bagging (SVC)	0.94	0.97	0.95	%94.41	%93.75
Bagging (GaussianNB)	0.91	0.95	0.92	%91.61	%90.45
Bagging (Karar Ağacı)	0.94	0.92	0.94	%91.61	%91.48
Adaboost (SVC)	0.96	0.97	0.96	%95.10	%94.66
Adaboost (GaussianNB)	0.96	0.97	0.96	%95.10	%94.66
Adaboost (Karar Ağacı)	0.93	0.88	0.92	%88.11	%88.30
Gradient Boosting	0.92	0.94	0.93	%91.61	%90.80
XGBoost	0.95	0.92	0.95	%92.31	%92.39
Stacking	0.96	0.98	0.96	%95.80	%95.23
Extra Trees	0.95	0.94	0.95	%93.71	%93.52

Tablo 4.2'de yer alan model başarım ölçütlerinin değerleri karşılaştırıldığında en yüksek doğruluk, Stacking yöntemi ile %95.80 oranında elde edilmiştir.

2. TARTIŞMA VE SONUÇ

Göğüs kanseri; hastalığın ilk evrelerinde teşhis edilip hastanın tedavisi gerçekleştirildiğinde, düşük ölüm oranına sahip bir kanser türüdür. Göğüs kanseri teşhisinde makine öğrenmesi yöntemlerinin kullanılması ile hastalığın geleneksel yöntemlere göre daha hızlı ve daha doğru bir şekilde teşhis edilerek göğüs kanseri hastalarının ölüm oranlarının azaltılması amaçlanmaktadır.

Bu çalışmada; SVC, KNN, Gaussian Naive Bayes ve Karar Ağacı algoritmaları bireysel sınıflandırıcı olarak kullanılmıştır. Bireysel sınıflandırıcıların doğruluk oranını arttımak için; rassal orman, oylama, torbalama, yükseltme, yığma ve ekstra ağaçlar topluluk yöntemleri uygulanmıştır. Eğitim ve test aşaması için Wisconsin Diagnostic Breast Cancer veri kümesi kullanılmıştır. Bu tez çalışması için geliştirilen uygulama, Python dili ile yazılmıştır. Sklearn kütüphanesinden yararlanılmıştır. Kesinlik, duyarlılık, f-skor, doğruluk, AUC skor ve ROC eğrisi model başarım ölçütlerine göre çalışmada uygulanan yöntemlerin performansları karşılaştırılmıştır.

SVC bireysel sınıflandırıcısının, SVC tabanlı topluluk sınıflandırıcıları ile kesinlik, duyarlılık, f-skor ve doğruluk model başarım ölçütlerine göre karşılaştırılması Tablo 5.1'de verilmiştir.

Tablo 5.1: SVC bireysel sınıflandırıcısının SVC tabanlı topluluk sınıflandırıcıları ile model başarım ölçütlerine göre karşılaştırılması.

Sınıflandırıcı	Kesinlik	Duyarlılık	F-Skor	Doğruluk
SVC	0.94	0.95	0.95	%93.71
Bagging (SVC)	0.94	0.97	0.95	%94.41
Adaboost (SVC)	0.96	0.97	0.96	%95.10

Tablo 5.1'e göre; SVC tabanlı Bagging (SVC) ve Adaboost (SVC) topluluk sınıflandırıcılarının, SVC bireysel sınıflandırıcısının doğruluk oranını arttırdığı gözlenmiştir.

SVC bireysel sınıflandırıcısı ile %93.71 doğruluk elde edilirken; Bagging (SVC) ve Adaboost (SVC) topluluk sınıflandırıcıları ile sırasıyla %94.41 ve %95.10 doğruluk elde edilmiştir. Gaussian Naive Bayes bireysel sınıflandırıcısının, Gaussian Naive Bayes tabanlı topluluk sınıflandırıcıları ile kesinlik, duyarlılık, f-skor ve doğruluk model başarım ölçütlerine göre karşılaştırılması Tablo 5.2'de verilmiştir.

Tablo 5.2: Gaussian Naive Bayes bireysel sınıflandırıcısının Gaussian Naive Bayes tabanlı topluluk sınıflandırıcıları ile karşılaştırılması.

Sınıflandırıcı	Kesinlik	Duyarlılık	F-Skor	Doğruluk
GaussianNB	0.91	0.95	0.92	%91.61
Bagging (GaussianNB)	0.91	0.95	0.92	%91.61
Adaboost (GaussianNB)	0.96	0.97	0.96	%95.10

Tablo 5.2'de yer alan model başarım ölçütlerinin değerleri karşılaştırıldığında; Gaussian Naive Bayes tabanlı Adaboost (GaussianNB) topluluk sınıflandırıcısının, Gaussian Naive Bayes bireysel sınıflandırıcısının doğruluk oranını arttırdığı gözlenmiştir (%95.10). Karar Ağacı bireysel sınıflandırıcısının, Karar Ağacı tabanlı topluluk sınıflandırıcıları ile kesinlik, duyarlılık, f-skor ve doğruluk model başarım ölçütlerine göre karşılaştırılması Tablo 5.3'te yer almaktadır.

Tablo 5.3: Karar Ağacı bireysel sınıflandırıcısının Karar Ağacı tabanlı topluluk sınıflandırıcıları ile karşılaştırılması.

Sınıflandırıcı	Kesinlik	Duyarlılık	F-Skor	Doğruluk
Karar Ağacı	0.90	0.88	0.89	%86.01
Rassal Orman	0.95	0.93	0.95	%93.01
Bagging (Karar Ağacı)	0.94	0.92	0.94	%91.61
Adaboost (Karar Ağacı)	0.93	0.88	0.92	%88.11
Gradient Boosting	0.92	0.94	0.93	%91.61
XGBoost	0.95	0.92	0.95	%92.31
Extra Trees	0.95	0.94	0.95	%93.71

Tablo 5.3'te yer alan model başarım ölçütlerinin değerleri karşılaştırıldığında; Karar Ağacı tabanlı topluluk sınıflandırıcılarının, Karar Ağacı bireysel sınıflandırıcısının doğruluk oranını arttırdığı gözlenmiştir. Karar Ağacı bireysel sınıflandırıcısı ile %86.01 doğruluk elde edilirken, Karar Ağacı tabanlı topluluk sınıflandırıcıları (Rassal Orman, Bagging(Karar Ağacı), Adaboost(Karar Ağacı), Gradient Boosting, XGBoost, Extra Trees) ile sırasıyla; %93.01, %91.61, %88.11, %91.61, %92.31 ve %93.71 doğruluk elde edilmiştir.

Bu çalışmada uygulanan bireysel sınıflandırıcılar karşılaştırıldığında; en yüksek doğruluk, veri kümesine Standardizasyon ve PCA işlemi uygulanması sonrasında SVC yöntemi ile elde edilmiştir (%93.71). Topluluk öğrenme yöntemleri arasında en yüksek doğruluk oranı, Stacking yöntemi ile %95.80 oranında elde edilmiştir.

Gelecekteki çalışmalarda, bu çalışmada kullanılan Wisconsin Diagnostic Breast Cancer veri kümesinden farklı veri setleri de kullanılarak daha iyi sonuçlar elde edilebilir. Bu çalışmada uygulanan yöntemler, diğer kanser türlerinin tespitinde de kullanılarak hekimlerin yararlanabileceği yardımcı bir araç geliştirilebilir.

KAYNAKLAR

Abdar, M., & Makarenkov, V. (2019). CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. *Measurement*, *146*, 557-570.

Abdar, M., Zomorodi-Moghadam, M., Zhou, X., Gururajan, R., Tao, X., Barua, P. D., & Gururajan, R. (2020). A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recognition Letters*, *132*, 123-131.

Agarap, A. F. M. (2018). On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In *Proceedings of the 2nd international conference on machine learning and soft computing*, 5-9.

Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform*, *4*(124), 3.

Avhad, K., Kore, G., Kaul, R., Jethwa, S., & Mundhe, P. (2020). Prediction of Breast Cancer Using Ensemble Learning. *International Research Journal of Engineering and Technology (IRJET)*, *7*(5), 7228-7231.

Azmi, M. S. B. M., & Cob, Z. C. (2010). Breast cancer prediction based on backpropagation algorithm. In *2010 IEEE Student Conference on Research and Development (SCOReD)*, 164-168.

Bernardi, D., et al. (2012). On the diagnostic accuracy of stereotactic vacuum-assisted biopsy of nonpalpable breast abnormalities. Results in a consecutive series of 769 procedures performed at the Trento Department of Breast Diagnosis. *Tumori Journal*, *98*(1), 113-118.

Berg, W. A. (2010). Benefits of screening mammography. *Jama*, *303*(2), 168-169.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 144-152.

Cancer Research UK, 2020, *Chemotherapy for breast cancer*, https://www.cancerresearchuk.org/about-cancer/breast-cancer/treatment/chemotherapy, [Ziyaret tarihi: 30 Nisan 2021].

El Bouchefry, K., & de Souza, R. S. (2020). Learning in Big Data: Introduction to Machine Learning. In *Knowledge Discovery in Big Data from Astronomy and Earth Observation*, 225-249.

Freund Y., Schapire R.E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory*, 23-37.

Friedman, J. (2001). Greedy boosting approximation: a gradient boosting machine. *Ann. Stat*. 29, 1189–1232.

Gayathri, B. M., & Sumathi, C. P. (2016). Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 1-5.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*(1), 3-42.

Ghosh, S., Mondal, S., & Ghosh, B. (2014). A comparative study of breast cancer detection based on SVM and MLP BPN classifier. In *2014 First International Conference on Automation, Control, Energy and Systems (ACES)*, 1-4.

Gu, D., Su, K., & Zhao, H. (2020). A case-based ensemble learning system for explainable breast cancer recurrence prediction. *Artificial Intelligence in Medicine*, *107*, 101858.

Hasan, H., & Tahir, N. M. (2010). Feature selection of breast cancer based on principal component analysis. In *2010 6th International Colloquium on Signal Processing & its Applications*, 1-4.

Islam, M. M., Iqbal, H., Haque, M. R., & Hasan, M. K. (2017). Prediction of breast cancer using support vector machine and K-Nearest neighbors. In *2017 IEEE Region 10 Humanitarian Technology Conference*, 226-229.

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, *1*(5), 1-14.

John Hopkins Medicine, *Breast Cancer Recurrence*, https://www.hopkinsmedicine.org/breast_center/breast_cancers_other_conditions/breast_cancer_recurrence/, [Ziyaret tarihi: 2 Mayıs 2021].

Li, H., & Zhang, Y. (2009). An algorithm of soft fault diagnosis for analog circuit based on the optimized SVM by GA. In *2009 9th International Conference on Electronic Measurement & Instruments*, 4-1023.

Liu, Y., Ren, L., Cao, X., & Tong, Y. (2020). Breast tumors recognition based on edge feature extraction using support vector machine. *Biomedical Signal Processing and Control*, *58*, 101825.

Łukasiewicz, E., Ziemiecka, A., Jakubowski, W., Vojinovic, J., Bogucevska, M., & Dobruch-Sobczak, K. (2017). Fine-needle versus core-needle biopsy–which one to choose in preoperative assessment of focal lesions in the breasts? Literature review. *Journal of ultrasonography*, *17*(71), 267.

Manikandan, G., Karthikeyan, B., Rajendiran, P., Harish, R., Prathyusha, T., & Sethu, V. (2019). Breast Cancer Prediction Using Ensemble Techniques. *SCOPUS IJPHRD CITATION SCORE*, *10*(7), 183.

Moon, W. K., Lee, Y. W., Ke, H. H., Lee, S. H., Huang, C. S., & Chang, R. F. (2020). Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Computer methods and programs in biomedicine*, *190*, 105361.

Mushtaq, M. S., & Mellouk, A. (2017). Methodologies for Subjective Video Streaming QoE Assessment. In *Quality of Experience Paradigm in Multimedia Services: Application to OTT Video Streaming and VoIP Services*, 27-57.

National Breast Cancer Foundation, *Radiation Therapy*, https://www.nationalbreastcancer.org/breast-cancer-radiation-therapy, [Ziyaret tarihi: 27 Nisan 2021].

National Cancer Institute, 2017, *Cancer Statistic*, https://www.cancer.gov/about-cancer/understanding/statistics, [Ziyaret tarihi: 10 Mart 2021].

Nematzadeh, Z., Ibrahim, R., & Selamat, A. (2015). Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. In *2015 10th Asian Control Conference*, 1-6.

Ojha, U., & Goel, S. (2017). A study on prediction of breast cancer recurrence using data mining techniques. In *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*, 527-530.

Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). Machine learning classification techniques for breast cancer diagnosis. In *IOP Conference Series: Materials Science and Engineering*, *495*(1), 012033.

Osareh, A., & Shadgar, B. (2010). Machine learning techniques to diagnose breast cancer. In *2010 5th international symposium on health informatics and bioinformatics*, 114-120.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

Pharoah, P. D., Day, N. E., Duffy, S., Easton, D. F., & Ponder, B. A. (1997). Family history and the risk of breast cancer: a systematic review and meta-analysis. *International journal of cancer*, *71*(5), 800-809.

Purnami, S. W., Rahayu, S. P., & Embong, A. (2008). Feature selection and classification of breast cancer diagnosis based on support vector machines. In *2008 International Symposium on Information Technology*, *1*, 1-6.

Senie, R. T., Rosen, P. P., Lesser, M. L., & Kinne, D. W. (1981). Breast self-examination and medical examination related to breast cancer stage. *American Journal of Public Health*, *71*(6), 583-590.

Sikov, W. M., Boughey, F. J. C., Al-Hilli, Z., & Chen, W, 2020, *General principles of neoadjuvant management of breast cancer*, https://www.uptodate.com/contents/general-principles-of-neoadjuvant-management-of-breast-cancer, [Ziyaret tarihi: 30 Nisan 2021].

Singletary, S. E., et al. (2002). Revision of the American Joint Committee on Cancer staging system for breast cancer. *Journal of clinical oncology*, *20*(17), 3628-3636.

Smith W. M., 2021, *What Is Targeted Therapy for Cancer?*, https://www.webmd.com/cancer/guide/targeted-therapy-cancer, [Ziyaret tarihi: 1 Mayıs 2021].

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. Advance online publication. https://doi.org/10.3322/caac.21660

Takkar, N., Kochhar, S., Garg, P., Pandey, A. K., Dalal, U. R., & Handa, U. (2017). Screening methods (clinical breast examination and mammography) to detect breast cancer in women aged 40–49 years. *Journal of mid-life health*, *8*(1), 2.

University of Rochester Medical Center, *Breast Conserving Surgery*, https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=92&contentid=P07789, [Ziyaret tarihi: 4 Mayıs 2021].

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1992). Breast cancer Wisconsin (diagnostic) data set. *UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/]*.

Wolberg, W. H., Street, W. N., Heisey, D. M., & Mangasarian, O. L. (1995). Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*, *26*(7), 792-796.

Yang, Y. (2016). *Temporal Data Mining via Unsupervised Ensemble Learning*. Elsevier.