



**SINIFLANDIRMA PROBLEMLERİNDE TOPLULUK ÖĞRENME
YÖNTEMLERİNİN İNCELENMESİ VE KÜÇÜK HÜCRELİ
DIŞI AKCİĞER KANSERİ VERİLERİ ÜZERİNE
BİR UYGULAMASI**

Mehmet KIVRAK

BİYOİSTATİSTİK VE TIP BİLİŞİMİ ANABİLİM DALI

**Tez Danışmanı
Prof. Dr. Cemil ÇOLAK**

Doktora Tezi-2021

**T.C.
İNÖNÜ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ**

**SINIFLANDIRMA PROBLEMLERİNDE TOPLULUK ÖĞRENME
YÖNTEMLERİNİN İNCELENMESİ VE KÜÇÜK HÜCRELİ DIŞI AKCİĞER
KANSERİ VERİLERİ ÜZERİNE BİR UYGULAMASI**

Mehmet KIVRAK

**Biyoistatistik ve Tıp Bilişimi Anabilim Dalı
Doktora Tezi**

**Tez Danışmanı
Prof. Dr. Cemil ÇOLAK**

**MALATYA
2021**

İÇİNDEKİLER

ÖZET.....	vi
ABSTRACT.....	vii
SİMGELER VE KISALTMALAR DİZİNİ.....	viii
ŞEKİLLER DİZİNİ.....	x
TABLolar DİZİNİ.....	xii
1. GİRİŞ.....	1
2. GENEL BİLGİLER.....	6
2.1. Topluluk Öğrenme.....	6
2.2. Tarihsel Gelişimi.....	8
2.3. Topluluk Öğrenme Yöntemleri.....	10
2.4. Sınıflayıcılar.....	15
2.5. Performans Metrikleri.....	18
2.6. Yeniden Örnekleme (Resampling) Yöntemleri.....	21
3. MATERYAL VE METOT.....	26
3.1. Örneklem Büyüklüğü.....	27
3.2. Veri Tabanı Bilgi Keşfi Süreci.....	27
4. BULGULAR.....	30
4.1. Tanımlayıcı İstatistikler.....	30
4.2. Bireysel Sınıflandırıcıların Performans Metrikleri Değerlendirmesi.....	35
4.3. Bagging Topluluk Öğrenme Yöntemi İle Sınıflayıcıların Performans Metriklerinin Değerlendirilmesi.....	38
4.4. Boosting Topluluk Öğrenme Yöntemi İle Sınıflayıcıların Performans Metriklerinin Değerlendirilmesi.....	40
4.5. Voting Topluluk Öğrenme Yöntemi İle Sınıflayıcıların Performans Metriklerinin Değerlendirilmesi.....	42
4.6. Stacking Topluluk Öğrenme Yöntemi İle Sınıflayıcıların Performans Metriklerinin Değerlendirilmesi.....	45
4.7. Genel Değerlendirme.....	49
5. TARTIŞMA.....	54
6. SONUÇ VE ÖNERİLER.....	58
KAYNAKLAR.....	59

EKLER.....	65
Ek-1. Özgeçmiş.....	65
Etik Kurul Almama Gerekçesi.....	66



TEŐEKKÜR

Doktora eđitimim süresince, desteđini hiçbir zaman esirgemeyen, bilgi ve birikimlerini her zaman benimle paylaŐan, tez sürecinde yol gösterici olan deđerli danıŐmanım Sayın Prof. Dr. Cemil OLAK'a, doktora eđitimi boyunca ve tez sürecinde her zaman beni destekleyen ve deneyimlerini paylaŐan Sayın Prof. Dr. Saim YOLOĐLU'na sonsuz teŐekkürlerimi sunarım.

Desteklerini hiçbir zaman esirgemeyen Sayın, ArŐ. Gör. Ahmet Kadir ARSLAN'a, Biyoistatistik ve Tıp BiliŐimi Anabilim dalında görev yapan öđretim elemanlarına, her zaman yanımda olan ve destekleyen eŐim Pakize KIVRAK'a sonsuz teŐekkürlerimi sunarım.

Mehmet KIVRAK

ÖZET

Sınıflandırma Problemlerinde Topluluk Öğrenme Yöntemlerinin İncelenmesi ve Küçük Hücreli Dışı Akciğer Kanseri Verileri Üzerine Bir Uygulaması

Amaç: Bu çalışmanın amacı, yapay zekâya dayalı topluluk öğrenme yöntemlerinin tıpta kullanılabilirliğini araştırılmak ve geliştirmektir. KHDAK ölüm durumunu sınıflandırmak için çeşitli klinik değişkenleri içeren veri seti kullanılarak, KHDAK değişkeni; yanıt/çıktı/hedef ve veri setinde yer alan ölçümler ile diğer faktörler açıklayıcı/tahminleyici/bağımsız değişkenler olmak üzere, topluluk öğrenme yöntemlerinin KHDAK değişkenini sınıflandırmadaki performansı incelenecek ve en iyi model belirlenecektir.

Materyal ve Metot: Bu çalışma, adresli açık kaynak erişimli web sitesinde yer alan KHDAK hastaları ölüm durumunu sınıflandırmak için çeşitli klinik değişkenleri içeren veri setini kullanarak yapılmıştır. Veri seti toplam 181 inoperabl evre I-IIIB KHDAK hasta kaydını içermektedir. Veri setinin yaklaşık % 55.2'si radyoterapi yada kemoterapi alan hastalardan oluşurken % 44.8'i radikal tedavi gören hastalardan oluşmaktadır. Makine öğrenme yöntemlerinden olan SMO, K-NN, random forest ve XGBoost gibi bireysel sınıflandırıcılar ve performansları ile topluluk öğrenme yöntemlerinden voting, bagging, boosting ve stacking yöntemleri kullanılmıştır.

Bulgular: Genel değerlendirmeye göre doğruluk, duyarlılık, kesinlik, özgüllük ve ROC eğrisi altında kalan alan metriklerinde en yüksek performansı boosting topluluk öğrenme yöntemi sağlamıştır. XGBoost ile en yüksek sınıflandırma performansını sağlayan boosting topluluk öğrenme yöntemi 0.982 doğruluk değeri, 0.971 duyarlılık değeri, 0.989 kesinlik değeri, 0.989 özgüllük değeri ve 0.998 ROC eğrisi altında kalan alan değerlerini elde etmiştir.

Sonuç: Temel sınıflandırıcılara göre topluluk öğrenme sınıflandırıcıları KHDAK ölüm durumunu sınıflandırmada daha iyi sonuçlar vermiştir. Prevelansı yüksek kanser hastalıklarında, sınıflandırma problemleri için topluluk öğrenme yöntemlerinin kullanılması başarılı sonuçların elde edilmesi açısından önerilmektedir.

Anahtar Kelimeler: KHDAK, Makine Öğrenme Yöntemleri, Sınıflandırıcılar, Topluluk Öğrenme Yöntemleri.

ABSTRACT

Examination of Ensemble Learning Methods in Classification Problems and An Application on Non-Small Cell Lung Cancer Data

Aim: The aim of this study is to investigate and improve the usability of AI-based community learning methods in medicine. Using the data set including various clinical variables to classify NSCLC death status, NSCLC variable; The performance of ensemble learning methods in classifying NSCLC variable will be examined, and the best model will be determined.

Material and Methods: This study was conducted using the data set containing various clinical variables to classify the death status of NSCLC patients on the addressed open source website. The data set includes a total of 181 inoperable stage I-IIIB NSCLC patients. Approximately 55.2 % of the data set consisted of patients receiving radiotherapy or chemotherapy, while 44.8 % consisted of patients receiving radical treatment. Individual classifiers such as SMO, K-NN, random forest and XGBoost, which are machine learning methods, and their performances, and voting, bagging, boosting and stacking methods from ensemble learning methods were used.

Results: According to the general evaluation, the boosting ensemble learning method provided the highest performance in the metrics of accuracy, sensitivity, precision, specificity and the area under the ROC curve. The boosting ensemble learning method, which provides the highest classification performance with XGBoost, achieved 0.982 accuracy value, 0.971 sensitivity value, 0.989 precision value, 0.989 specificity value and 0.998 ROC curve.

Conclusion: Ensemble learning classifiers gave better results in classifying NSCLC mortality according to the basic classifiers. It is recommended to use ensemble learning methods for classification problems in cancer patients with high prevalence in order to achieve successful results.

Key Words: NSCLC, Machine Learning Methods, Classifiers, Ensemble Learning Methods.

SİMGELER VE KISALTMALAR DİZİNİ

WHO	: Dünya Sağlık Örgütü
KHDAK	: Küçük Hücreli Dışı Akciğer Kanseri
PET	: Pozitron Emisyon Tomografi
MR	: Manyetik Rezonans
TNM	: Tümör, Lenf Nodu, Metastaz
VTBK	: Veri Tabanından Bilgi Keşfi
K-NN	: K- En Yakın Komşu
SVM	: Destek Vektör Makineleri
Adaboost	: Adaptiv Boosting
OOB	: Out of Bag
SMO	: Ardışık Minimal Optimizasyon
XGBoost	: Aşırı Gradyan Artırma
AC	: Doğruluk
TP	: Gerçek Pozitif Oranı
FP	: Gerçek Negatif Oranı
ROC	: Receiver Operating Characteristic
WHO-PS	: Dünya Sağlık Örgütü Performans Durumu
FEV1s%	: Solunum Testleri
RT	: Radyasyon Tedavisi
GTV	: Primer Tümör ve Metastatik Lenf Nodlarının Toplamı
OPN	: Osteopontin
CA-9	: Karbonik Anhidraz-9
IL-6	: İnterlökin-6
IL-8	: İnterlökin-8
CRP	: C-Reaktif Protein
CEA	: Tümör Yüğü Karsinoembriyonik Antijen
Cayfra 21-1	: Sitokeratin 21-1
α2M	: Alpha-2-Macroglobulin

sIL2R	: Serumda Çözülür IL2 Respektörü
TLR-4	: Toll Benzeri Respektör 4
VEGF	: Vasküler Endotel Büyüme Faktörü
LOF	: Yerel Aykırı Değer Faktörü
AUC	: ROC Eğrisi Altında Kalan Alan
RIPPER	: Repeated Incremental Pruning to Produce Error Reduction
DS	: Decision Stamp
GBT	: Gradian Boosted Tree
MLP	: Multilayer Perceptron



ŞEKİLLER DİZİNİ

Şekil No	Sayfa No
Şekil 1.1. Veri Madenciliği Bilgi Keşfi Süreçleri.....	3
Şekil 1.2. Sentetik Üç Gauss Verisi.....	4
Şekil 2.1. Çeşitlilik ve Doğruluk İlişkisi.....	7
Şekil 2.2. Boosting Topluluk Öğrenme İşlem Adımları.....	11
Şekil 2.3. Bagging Topluluk Öğrenme İşlem Adımları.....	13
Şekil 2.4. Voting Topluluk Öğrenme İşlem Adımları.....	14
Şekil 2.5. Stacking Topluluk Öğrenme.....	15
Şekil 2.6. Random Forest Sınıflayıcısı.....	16
Şekil 2.7. SVM Sınıflayıcısı.....	17
Şekil 2.8. K-NN Sınıflayıcısı.....	18
Şekil 2.9. XGBoost Sınıflayıcısı.....	18
Şekil 2.10. ROC Eğrisi Grafiği.....	21
Şekil 2.11. k-Kat Çapraz Geçerlilik.....	23
Şekil 2.12. Jackknife Yeniden Örnekleme.....	24
Şekil 2.13. Bootstrap Yeniden Örnekleme.....	25
Şekil 3.1. Her Bir Grup için Önerilen Örneklem Büyüklüğü.....	27
Şekil 3.2. Çok Değişkenli Normal Dağılım.....	29
Şekil 4.1. Akciğer Kanseri Durumu.....	30
Şekil 4.2. Cinsiyete Göre Akciğer Kanseri Durumu.....	31
Şekil 4.3. Lymphnodes Değişkenine Göre Akciğer Kanseri Durumu.....	31
Şekil 4.4. WHO-PS Değişkenine Göre Akciğer Kanseri Durumu.....	32
Şekil 4.5. OPN Değişkenine Ait Histoğram Grafiği.....	33
Şekil 4.6. SIL2R Değişkenine Ait Histoğram Grafiği.....	33

Şekil No	Sayfa No
Şekil 4.7. Yaş Değişkenine Ait Histoğram Grafiği.....	34
Şekil 4.8. Fev1s% Değişkenine Ait Histoğram Grafiği.....	35
Şekil 4.9. Total Dose (2) Değişkenine Ait Histoğram Grafiği.....	35
Şekil 4.10. Bireysel Sınıflandırıcıların Performans Metrikleri.....	37
Şekil 4.11. Bagging Topluluk Öğrenme Yöntemlerinde Sınıflandırıcıların Performans Metrikleri.....	40
Şekil 4.12. Boosting Topluluk Öğrenme Yöntemlerinde Sınıflandırıcıların Performans Metrikleri.....	42
Şekil 4.13. Voting Topluluk Öğrenme Yöntemlerinde Sınıflandırıcıların Performans Metrikleri.....	45
Şekil 4.14. Stacking Topluluk Öğrenme Yöntemlerinde Sınıflandırıcıların Performans Metrikleri.....	49
Şekil 4.15. XGBoost Sınıflayıcısına Ait Pseudo Kodu.....	51
Şekil 4.16. Genel Performans Metrikleri Değerlendirmesi.....	53

TABLolar DİZİNİ

Tablo No	Sayfa No
Tablo 1.1. TNM Sınıflamasına Göre KHDAK Evrelemesi.....	2
Tablo 2.1. Confusion Matrisi.....	19
Tablo 3.1. Çalışmada Kullanılan Değişkenler.....	26
Tablo 4.1. Akciğer Kanseri Sağkalım Durumu.....	30
Tablo 4.2. Kategorik Değişkenlere Göre Akciğer Kanseri Durumu.....	32
Tablo 4.3. Normal Dağılım Gösteren Sürekli Değişkenlere Göre Akciğer Kanseri Durumu.....	34
Tablo 4.4. Normal Dağılım Göstermeyen Sürekli Değişkenlere Göre Akciğer Kanseri Durumu (Man-Whitney U)....	35
Tablo 4.5. Random Forest Sınıflandırıcısının Performans Metriği.....	36
Tablo 4.6. SMO Sınıflandırıcısının Performans Metriği.....	36
Tablo 4.7. K-NN Sınıflandırıcısının Performans Metriği.....	37
Tablo 4.8. XGBoost Sınıflandırıcısının Performans Metriği.....	37
Tablo 4.9. Random Forest Sınıflandırıcısının Bagging Topluluk Öğrenme Yöntemine Göre Performans Metriği.....	38
Tablo 4.10. SMO Sınıflandırıcısının Bagging Topluluk Öğrenme Yöntemine Göre Performans Metriği.....	38
Tablo 4.11. K-NN Sınıflandırıcısının Bagging Topluluk Öğrenme Yöntemine Göre Performans Metriği.....	39
Tablo 4.12. XGBoost Sınıflandırıcısının Bagging Topluluk Öğrenme Yöntemine Göre Performans Metriği.....	39
Tablo 4.13. Random Forest Sınıflandırıcısının Boosting Topluluk Öğrenme Yöntemine Göre Performans Metriği.....	40
Tablo 4.14. SMO Sınıflandırıcısının Boosting Topluluk Öğrenme Yöntemine Göre Performans Metriği.....	41

Tablo No	Sayfa No
Tablo 4.15. K-NN Sınıflandırıcısının Boosting Topluluk Öğrenme Yöntemine Göre Performans Metriği.....	41
Tablo 4.16. XGBoost Sınıflandırıcısının Boosting Topluluk Öğrenme Yöntemine Göre Performans Metriği.....	41
Tablo 4.17. Voting Topluluk Öğrenme Yöntemine Göre Model 1'in Performans Metriği.....	43
Tablo 4.18. Voting Topluluk Öğrenme Yöntemine Göre Model 2'nin Performans Metriği.....	43
Tablo 4.19. Voting Topluluk Öğrenme Yöntemine Göre Model 3'nin Performans Metriği.....	43
Tablo 4.20. Voting Topluluk Öğrenme Yöntemine Göre Model 4'ün Performans Metriği.....	44
Tablo 4.21. Voting Topluluk Öğrenme Yöntemine Göre Model 5'in Performans Metriği.....	44
Tablo 4.22. Voting Topluluk Öğrenme Yöntemine Göre Model 6'nın Performans Metriği.....	44
Tablo 4.23. Voting Topluluk Öğrenme Yöntemine Göre Model 7'nin Performans Metriği.....	45
Tablo 4.24. Stacking Topluluk Öğrenme Yöntemine Göre Model 1'in Performans Metriği.....	46
Tablo 4.25. Stacking Topluluk Öğrenme Yöntemine Göre Model 2'nin Performans Metriği.....	46
Tablo 4.26. Stacking Topluluk Öğrenme Yöntemine Göre Model 3'nin Performans Metriği.....	47
Tablo 4.27. Stacking Topluluk Öğrenme Yöntemine Göre Model 4'ün Performans Metriği.....	47
Tablo 4.28. Stacking Topluluk Öğrenme Yöntemine Göre Model 5'in Performans Metriği.....	48

Tablo No	Sayfa No
Tablo 4.29. Stacking Topluluk Öğrenme Yöntemine Göre Model 6'nın Performans Metriği.....	48
Tablo 4.30. Stacking Topluluk Öğrenme Yöntemine Göre Model 7'nin Performans Metriği.....	48
Tablo 4.31. Genel Değerlendirme.....	50
Tablo 4.32. En İyi Modellere Ait Değişken Önemlilikleri.....	52



1. GİRİŞ

Akciğer kanseri, tüm dünyada olduğu gibi ülkemizde de en sık görülen kanser türlerinin başında yer almaktadır. Özellikle sigara kullanımına bağlı olarak erkeklerde ve kadınlarda hastalığın ortaya çıkması günden güne artmaktadır (1). Sigara ve pasif sigara maruziyeti, hava kirliliği, cinsiyet, mesleki nedenler, genetik faktörler, kronik akciğer rahatsızlıkları ve radyoterapi başlıca risk faktörleridir. Dünya Sağlık Örgütü'nün (DSÖ) 2004 sınıflamasında, "malign epitelyal tümörler" başlığı altında sekiz ayrı tümör kategorisi ve çok sayıda varyant bulunmaktadır. Tedavi ve prognoz açısından akciğer kanseri, küçük hücreli (KHAK) ve küçük hücreli dışı akciğer kanseri (KHDAK) olarak sınıflandırılır. Küçük hücreli akciğer kanseri radyoterapi yada kemoterapi ile tedavi edilirken KHDAK tedavisi cerrahi müdahaledir. KHDAK vakalarında, yaş, cinsiyet, kilo kaybı, performans durumu, tümör evresi gibi etkenler başlıca prognostik faktörlerdir. Özellikle cerrahi müdahale uygulanan hastalarda 5 yıllık sağ kalım oranı evre IA için % 67, evre IIA için % 55, evre IIIA için % 23 olarak bilinmektedir. Evre IIIB olgularda ise bu oran sadece % 3-7'dir (2).

Akciğer kanserinde tümör tanısının konulması ve evrelendirilmesi için pozitron emisyon tomografisi (PET), manyetik rezonans (MR) ve spiral toraks tomografi (Toraks-BT) kullanılan görüntüleme yöntemleridir (3). KHDAK tanısı, olguların çoğunda ileri evrelerde konmaktadır. Belirtiler tümörün büyümesine bağlı olduğu kadar, metastatik hastalığa veya non- metastatik paraneoplastik sendromlara da bağlı olabilir (4). Ayrıca genomik ve proteomikten türetilmiş fakat bunlarla sınırlı olmayan yeni prognostik parametrelerin araştırılması da gerekmektedir. Proteinler de içeren biyobelirteçlerin analizi hızlı bir şekilde gelişmektedir ve normal biyolojik süreçlerin, patojenik süreçlerin veya terapötik bir müdahaleye farmakolojik yanıtların öngörülmesine veya tanımlanmasına izin verir. Onkoproteinler tümör hücreleri tarafından veya varlıklarına yanıt olarak üretilir ve kanser hastalarının kan dolaşımına salınabilir. Akciğer kanseri hastalarında doku örneklemesi genellikle mümkün olmadığından, venipunktur ile kan örneği toplanması, güvenli ve uygulanması kolay alternatif bir yöntemdir. Hastadan alınan kan biyobelirteçleri, tümör çevresindeki farklılıkları yansıtır. Ayrıca hastalık prognozu ve tedaviye yanıt ile bağlantılıdır (5). Kanserin patolojik tanısında bronkoskopi, transtorasik iğne aspirasyonu ve balgam sitolojisi gibi primer tümöre

yönelik yöntemler kullanıldığı gibi torasentez, kapalı plevra biyopsisi, torakoskopi, lenf bezi ve cilt biyopsisi gibi metastaz bölgesine yönelik yöntemler de kullanılır (6). KHDAK evrelemesinde tümörün özellikleri, lenf nodu ve metastaz (TNM) sistemi kullanılmaktadır. T primer tümörü, N bölgesel lenf bezlerini ve M uzak metastazı tanımlar. TNM sınıflamasına göre evreleme tablo 1.1'de verilmiştir (7).

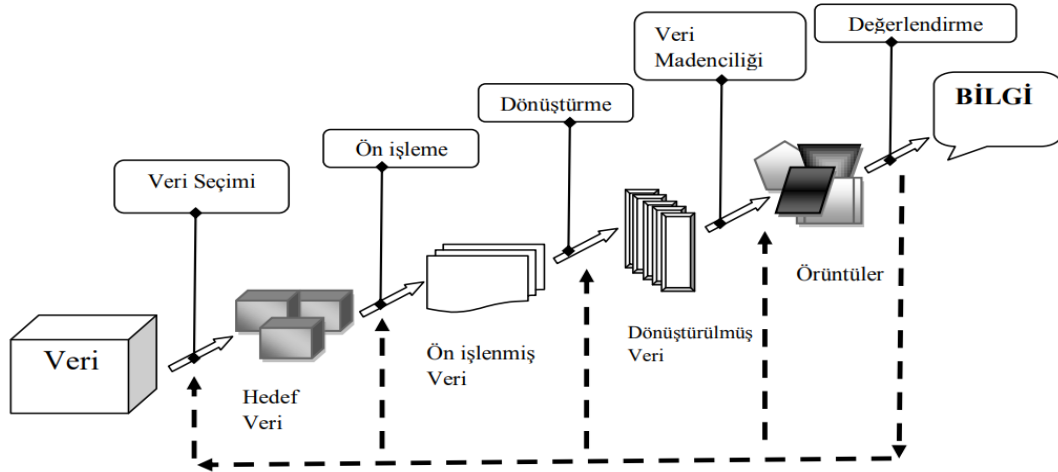
Tablo 1.1. TNM Sınıflamasına Göre KHDAK Evrelemesi (7).

EVRE 0	TisN0M0
EVRE IA	T1N0M0
EVRE IB	T2N0M0
EVRE IIA	T1N1M0
EVRE IIB	T2N1M0 T3N0M0
EVRE IIIA	T3N1M0 T1-3N2M0
EVRE IIIB	T4N0-3M0 T1-4N3M0
EVRE IV	Herhangi bir T M1 Herhangi bir N M1

KHDAK hastaları için tanı veya tedavi gecikmeleri önemli bir sorundur. Özellikle bu gecikmelerin tümör evresi ve prognoza ciddi etkisi olduğu bilinmektedir (2). KHDAK hastaları için çok faktörlü karara dayalı radyasyon onkolojisi yoluyla geliştirilmiş karar destek sistemleri, yapay zeka ve makine öğrenme algoritmaları sayesinde, doğru tedavi yöntemleri belirlenerek kişiye özel bir tedavi süreci oluşturur. Böylece hastaya aşırı yada gereğinden az tedaviden sakınılmış olunur (5). Makine öğrenme yöntemleri, hastalıkların teşhisinde ve klinik karar destek sistemlerinde son yıllarda kullanılan veri madenciliği yöntemlerindedir. İstatistik analiz yöntemleri ve yapay zeka algoritmalarının birlikte kullanılarak veri içerisindeki gizli bilgilerin açığa çıkarılması ve verinin nitelikli bilgiye dönüştürülmesi sürecidir (8). Günümüzde birçok sağlık alanında kullanılan bu yöntemler, veri tabanlarında gizli kalmış örüntülerin otomatik olarak elde edilmesi ve veri analiz sürecinin verimliliğine önem vermektedir (9).

Birçok araştırmacı, veri madenciliği yöntemlerini; veriden tabanından bilgi bulma keşfi (VTBK) olarak da tanımlamaktadır. VTBK, belirli süreçler ile basamaklandırılmıştır. Bu bilgi keşfi süreçleri aşağıdaki gibidir:

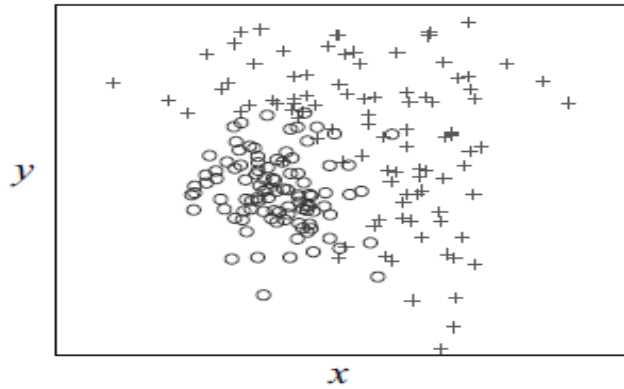
- 1) Veri Temizleme (Data Cleaning): Verilerin gereksizlikten ve tutarsızlıktan temizlenmesi sürecidir.
- 2) Veri Entegrasyonu (Data Integration): Büyük miktarlardaki verilerin birleştirilmesi işlemidir.
- 3) Veri Seçimi (Data Selection): Bu süreçte veri ambarlarından amaca uygun veriler alınır.
- 4) Veri Dönüştürme (Data Transformation): Verilerin analiz edilebilecek ve özetlenebilecek forma dönüştürüldüğü süreçtir.
- 5) Veri Madenciliği (Data Mining): Veri temizleme ve önışlemeden geçen verilere, veri madenciliği yöntemlerinin uygulanmasıdır. Bu yöntemler: kümeleme, sınıflama, ilişkilendirme, vb. olarak sıralanabilir.
- 6) Veriye Kalıp Biçmek (Data Pattern Evaluation): Veriye çeşitli ölçüm ve doğru yaklaşımlarla yeni bir değer biçmek olarak tanımlanabilir.
- 7) Bilgi Keşfi (Knowledge Discovery): Bu süreç, araştırmacının kullanacağı tutarlı verilerin keşfi sürecidir (10).



Şekil 1.1. Veri Madenciliği Bilgi Keşfi Süreçleri (10).

İstatistiksel makine öğrenme algoritması, veri madenciliğinin temel görevinden olan örüntü tanıma ve veri kümelerinden iyi modeller oluşturmak için tasarlanmış işlemler bütünüdür (11). Geliştirilen istatistiksel yöntemlerden biri olan topluluk öğrenme, bir optimal tahmin modelini üretmek için birkaç temel modeli birleştiren

algoritmadır (12). Benzer istatistiksel makine öğrenme algoritmaları için çok çeşitli uygulamalar ve bir algoritma belirli bir probleme uygulandığında değerlendirilmesi gereken çeşitli performans ölçümleri vardır. Performans ölçümlerinin karşılaştırılması, lojistik regresyon, naive bayes (NB) sınıflandırıcı, K- En Yakın Komşu (K-NN), Karar Ağaçları ve Destek Vektör Makineleri (SVM) gibi temel makine öğrenme algoritmalarıyla yapılır (13). İstatistiksel makine öğrenmede veri seti, genellikle özellik vektörlerinden oluşur; burada her özellik vektörü, bir özellik seti kullanarak her bir nesneyi açıklamaktadır. Örneğin, Şekil 1.2'de gösterilen sentetik üç Gauss verisine bakıldığında, her bir nesne, x ve y koordinatında tanımlı özelliklerine göre şekillenen bir veri noktasıdır. Bir veri kümesinin özelliklerinin sayısına boyut veya boyutluluk denir; örneğin, şekil 1.2'deki veri setinin boyutsallığı üçtür. Veri setinden model üretme sürecine öğrenme yada eğitim denilmektedir. Türlerine göre en yaygın kullanılan öğrenme yöntemleri, denetimli ve denetimsiz öğrenmedir. Denetimli öğrenmede amaç, görülmeyen yada belli olmayan durumlarda, bir hedef özelliğinin değerini tahmin etmektir. Denetimli öğrenmede öğrenilen modele kestirici denir. Örneğin, üç Gauss veri noktasının şeklini tahmin etmek istiyorsak, "çapraz" ve "daire" etiketleri olarak adlandırır ve etiket bilgilerinin bilinmediği bir örneğin etiketini tahmin ederiz. Eğer etiket (label) bilgisi kategorik ölçekte veri ise sınıflama, sayısal ölçekte veri ise regresyon algoritması kullanılmaktadır. Her iki durumda da, eğitim süreci etiket bilgisini içeren veri setleri üzerinden yürütülür. Denetimsiz öğrenme yöntemleri etiket bilgisine dayanmaz. Denetimsiz öğrenmede amaç veri yapısındaki gizli bilgileri keşfetmektir. Örneğin denetimsiz öğrenme yöntemlerinden kümeleme yöntemi, verilerin benzer özelliklerine göre bir araya gelmesini sağlamaktadır (14).



Şekil 1.2. Sentetik Üç Gauss Verisi (14).

Modellerin performansını deęerlendirmek ve tahmin etmek, daha sonra kullanıcının bir modelin kabul edilebilir olup olmadıęına karar vermesine veya bir dizi aday arasından en uygun modeli seęmesine izin vermek için, modelin genelleme hatası ya da kestirim hatasına bakmak gereklidir. Ancak genelleme hatasını doęrudan tahmin etmek mümkün olmadıęından etiket bilgisinin doęru bilindięi test verileri üzerinden tahminler yapılarak, test hatasının genelleme hatası olarak kabul edilmesi saęlanır. Bu işlem sürecine test (testing) süreci denir. Test sürecinden önce, öęrenilmiş bir modelin genellikle parametreleri ayarlama gibi yapılandırılması gerekir ve bu işlem aynı zamanda öęrenme performansını deęerlendirmek için bilinen temel doęruluk etiketleriyle verilerin kullanımını içerir; Buna doęrulama (validation) denir ve veriler doęrulama verileridir. Genel olarak, test verileri eğitim ve doęrulama verileri ile örtüşmemelidir; aksi takdirde, model performansı aşırı iyimser olabilir (14). Makine öęrenme modellerini deęerlendirmek için kullanılan dięer yaygın performans ölçütleri özgüllük, duyarlılık, kesinlik, ROC eęrisi, vs.'dir (11).

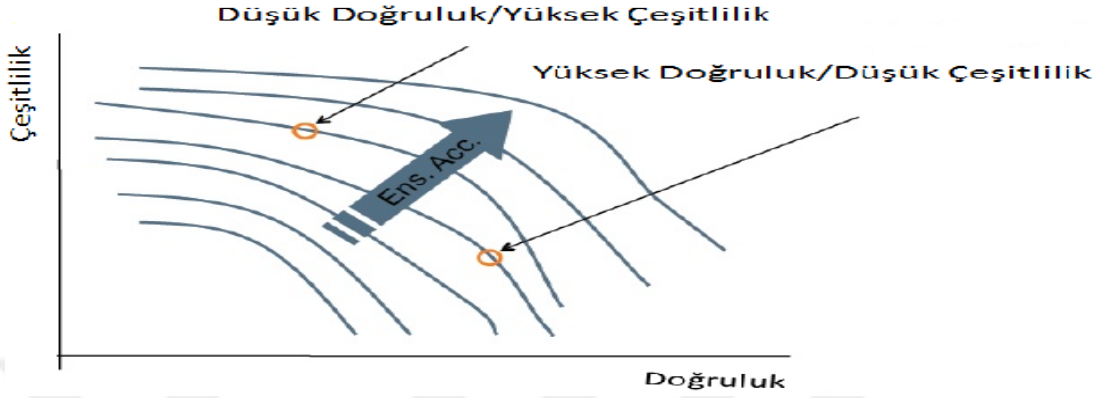
Bu çalışmanın amacı, yapay zekâya dayalı topluluk öęrenme yöntemlerinin tıpta kullanılabilirliğini araştırılmak ve geliştirmektir. KHDAK ölüm durumunu sınıflandırmak için çeşitli klinik deęişkenleri içeren veri seti kullanılarak, KHDAK deęişkeni; yanıt/çıktı/hedef ve veri setinde yer alan ölçümler ile dięer faktörler açıklayıcı/tahminleyici/bağımsız deęişkenler olmak üzere, topluluk öęrenme yöntemlerinin KHDAK deęişkenini sınıflandırmadaki performansı incelenecek ve en iyi model belirlenecektir.

2. GENEL BİLGİLER

2.1. Topluluk Öğrenme (Ensemble Learning)

Bu tez çalışmasında, performans ölçütlerini değerlendirmek, karşılaştırmak ve optimize etmek için topluluk öğrenme algoritmaları kullanılmıştır. Birçok farklı model algoritması kullanarak tahmin üreten bu yöntemin son yıllarda veri madenciliği ve yapay zeka alanında yapılan çalışmalarda geniş bir uygulama alanının olması, sıklıkla başvurulan bir yöntem olarak kullanımını artırmaktadır (15). Doğruluğu artırıp varyansı azaltmak için geliştirilen topluluk öğrenme algoritmaları, özellik çıkarımı, güven tahmini, hata düzeltme, dengesiz veriler ve durağan olmayan dağılımlarda sapmayı öğrenme gibi çeşitli makine öğrenme sistemlerinde başarılı sonuçlar vermektedir. Topluluk öğrenme, çoklu modelleri birleştirme yöntemidir ve tipik olarak modelin kestirim performansını geliştirerek verilecek karara olan güveni artırmaktadır (16). Topluluk öğrenme yöntemlerinde ilk olarak veri seti, eğitim aşamasında bir dizi model oluşturmak için kullanılır (17). Topluluk sistemlerinin geliştirilmesi için önem arz eden bu aşamada, farklı hatalar öğrenilir. Çünkü topluluk sınıflandırıcılarının benzer sonuçlar üretmesi, sınıflandırıcıların birleştirilmesine anlamlı bir etki yapmaz (16). Ayırma fonksiyonu, bu aşamada modellerin ne kadar eğitim verisi ile çalıştığını belirler. Yani bu süreçte, modelin bütün bir eğitim veri setiyle mi yoksa eğitim veri setinin farklı alt kümeleriyle mi çalışacağını belirlenir. İdeal topluluk öğrenme yöntemleri çeşitli modellere sahip olsa bile, ayırma fonksiyonu elde edilen bilgileri kararlı hale getirmek için girdi verilerini yapay olarak değiştirerek çeşitliliği artırmaktadır (18). Topluluk sınıflayıcıların eğitimi için çok sayıda algoritma geliştirilmiştir. Torbalama (bagging), artırma (boosting), oylama (voting) ve istifleme (stacking) algoritmaları bilinen en yaygın yaklaşımlardır. Ayırma fonksiyonu ile kullanılan bu türden farklı yaklaşımlar, topluluk sistemlerinin gelişimine katkı sunmuştur. Örneğin eğitim veri setinin bootstrapp yöntemiyle tekrarlı olarak eğitilmesi, "bagging" yönteminin gelişmesine sebep olmuştur. Benzer şekilde, her sınıflandırıcıyı eğitmek için veri setinin farklı alt kümelerinin kullanımı, rastgele alt uzay metodu (random subspace methods) yönteminin gelişmesine katkı sunmuştur. Çeşitliliğin olmaması, model performansını olumsuz etkilese bile topluluk sistemlerinin doğruluğu ile doğrudan bir ilişkisi bulunamamıştır. Önemli olan topluluk sınıflandırıcı sonuçlarının bağımsız ve negatif yönlü korelasyona sahip olmasıdır (17). Sınıflayıcılar genelde düşük biasta ve yüksek varyansa sahiptirler. Varyansın düşmesini sağlayan etkiyi ise ortalama

parametresi yapmaktadır. Sınıflandırma hatasının yanlılık (bias) ve varyans olmak üzere iki temel bileşeni vardır. Topluluk öğrenme yöntemleri, sabit yanlılığa sahip birkaç sınıflandırıcı oluşturarak ortalamalar ile ortalamaların varyansını düşürecek sonuçları birleştirecektir (16). Çeşitlilik ve doğruluk ilişkisi şekil 2.1'de verilmiştir.



Şekil 2.1. Çeşitlilik ve Doğruluk İlişkisi

Yanlılık modelin problem çözümünü içermediğini gösterir. Modelin zayıf kaldığı bu duruma eksik öğrenme (underfitting) denir. Yüksek yanlılığa sahip bir modelin, eğitim verisini eksik öğrenme olasılığı çoğu zaman yüksek çıkar (18, 19). Yanlılık, araştırmanın herhangi bir aşamasında (planlama, veri derleme, analiz ve rapor yazımı), ya da herhangi bir aşamanın herhangi bir işleminde ortaya çıkabilecek ve araştırma sonucunu gerçek durumundan (doğruluktan) saptırabilecek sistematik işlemler olarak tanımlanabilir. Yanlılık örnekleme değişkenliği dışındaki nedenlere bağlı olarak ortaya çıktığından, örnek hacmi artırılrsa bile ortadan kalkmaz. Model varsayımlarının geçerliliği ve rastgelelik ile yanlılık azaltılabilir. Varyans genelde rastgele hatalardan kaynaklanmaktadır. Önceden bilinmeyen, araştırma sürecinde ortaya çıkabilen hatalar olarak tanımlanabilir (20). Varyans, modelin tahmin ettiği verilerin, gerçek veriler etrafında nasıl saçıldığını ölçer. Eğer varyans yüksek ise model fazla geneldir, buna aşırı öğrenme (overfitting) denir (19). Örnek büyüklüğünün artırılması, kullanılan yöntemin iyileştirilmesi ve uygun analiz yöntemlerinin kullanılması varyansı azaltabilir (20). Sınıflandırma performansını önemli ölçüde artıran diğer bir işlem birleştirme algoritmasının seçimidir. Sınıflayıcılar için en uygun farklı birleştirme yöntemleri vardır. Torbalama (bagging), artırma (boosting), oylama (voting) ve istifleme (stacking) topluluk öğrenme yöntemleri bunlardandır (14, 16, 21).

2.2. Tarihsel Gelişimi

Topluluk öğrenme yönteminin klinik çalışmalarda sınıflandırma için kullanımı, makine öğrenme yöntemlerinin klinik çalışmalarda kullanılmaya başlaması ile ortaya çıkmıştır. Topluluk öğrenme yöntemlerinin yakın zamanda ortaya çıkması ve etkili bir yöntem olması, halen güncelliğini korumasını sağlamaktadır. Topluluk öğrenme yöntemlerinin temeli, Sheela ve Dasarath 1979'da özellik uzayını çoklu sınıflandırıcı kullanarak bölmesine dayanır (22). Wittner ve Denker 1988'de çok katmanlı sinir ağlarını tartışarak, sınıflandırma stratejilerini araştırdılar (23).

1990 yılında Schapire, yapmış olduğu bir çalışmada "boosting" adını verdiği yöntemle topluluk sınıflandırıcılarından elde edilen hatanın tek bir sınıflandırıcıdan elde edilen hatadan daha düşük olacağını ve ikili sınıflandırmalar için güçlü bir sınıflandırıcı olacağını belirtmiştir (24).

Hansen ve Salamon 1990 yılında sinir ağları ile ilgili yapmış oldukları çalışmada, sınıflama performansının geliştirilmesinde topluluk öğrenme yöntemlerini önermişlerdir (25).

Xu ve arkadaşları 1992 yılında, birden fazla sınıflayıcı ve uygulamalarını, el yazısı tanımlama için ele aldılar. Çeşitli sınıflandırıcıların sonuçlarına göre çoklu sınıflandırıcıların voting, bayesçi yaklaşım ve Demster -Shafer gibi üç şekilde birleştirilebileceğini belirttiler. Bu sayede bireysel sınıflandırıcıların performansının önemli ölçüde iyileştirilebileceğini buldular (26).

Luo ve arkadaşları 2012 yılında göğüs kanserinin mamografi ile teşhisi konulu çalışmalarında, topluluk öğrenme yöntemlerinden yararlanmışlardır. Modelde karar ağacı ve destek vektör makinesi sınıflayıcıların birleşiminden oluşan modeli önermişlerdir. Topluluk modelin bireysel sınıflayıcılara göre daha yüksek performans sağladığını tespit etmişlerdir (27).

Nagi ve arkadaşları 2013 yılında kanserli vakalarda mikrodizi verilerinin sınıflandırılması konulu çalışmalarında, bireysel sınıflandırıcılar ve topluluk öğrenme yöntemlerinden faydalanarak sınıflama performanslarını ölçmüşlerdir (28)

Onan ve arkadaşları 2015 yılında göğüs kanserinin sınıflandırılması konulu yapmış oldukları çalışmada, topluluk öğrenme yöntemlerini kullanarak farklı sınıflandırıcılar ile topluluk öğrenme algoritmalarında modele dahil etmişlerdir. Topluluk

öğrenme algoritmalarının bireysel sınıflandırıcılara göre daha başarılı sonuçlar verdiğini gözlemlemişlerdir (29).

Tabibu ve arkadaşları 2019 yılında Pan-renal hücreli karsinomların sınıflandırılmasına ilişkin yapmış oldukları çalışmalarında, histopatolojik görüntülerin makine öğrenme yöntemleri ve konvolüsyonel sinir ağları yöntemlerini uygulamışlardır. Çalışmalarında görüntülerin tanınmasında konvolüsyonel sinir ağları kullanmalarının yanısıra sınıflandırma performansının artırılması için yeni bir destek vektör makinesi modeli öne sürmüşlerdir (30).

Wang ve arkadaşları 2019 yılında stacking topluluk öğrenme yöntemiyle prostat kanserinin tespiti konulu çalışmalarında, stacking topluluk öğrenme yöntemlerinde ağaç tabanlı sınıflayıcıları kullanarak başarılı sonuçlar elde etmişlerdir (31).

KHDAK hastalıklarında topluluk öğrenme yöntemleri, çok yakın zamanlarda kullanılmaya başlanmıştır. Cai ve arkadaşları yapmış oldukları çalışmada, bireysel makine öğrenme algoritmaları ve topluluk öğrenme yöntemlerini kullanarak KHDAK, KHAK ve karsinoid akciğer kanseri türleri arasında klinik tanı ve tedaviyi kolaylaştıracak ortak bir DNA metilasyon belirteçleri panelinin olası varlığını ortaya koymuştur (32).

Bergquist ve arkadaşları sağlık hizmeti talep verilerinde topluluk yöntemleri ve makine öğrenimi ile akciğer kanseri şiddetinin sınıflandırılması konulu çalışmalarında, kanser kayıt verilerini kullanarak, kemoterapi alan akciğer kanseri hastalarını erken ve geç evre kanser olarak sınıflandırmak için bir araç geliştirmişlerdir (33).

Tan ve arkadaşları kanser sınıflandırması için gen ekspresyon verileri üzerinde topluluk öğrenme yöntemleri konulu yapmış oldukları çalışmada, klasik makine öğrenme yöntemleri ile topluluk öğrenme yöntemleri kullanılarak yedi kanserli mikrodizi verisi üzerinde sınıflandırma işlemi gerçekleştirerek bu yöntemlerin sınıflandırma / tahmin performansını karşılaştırdılar (34).

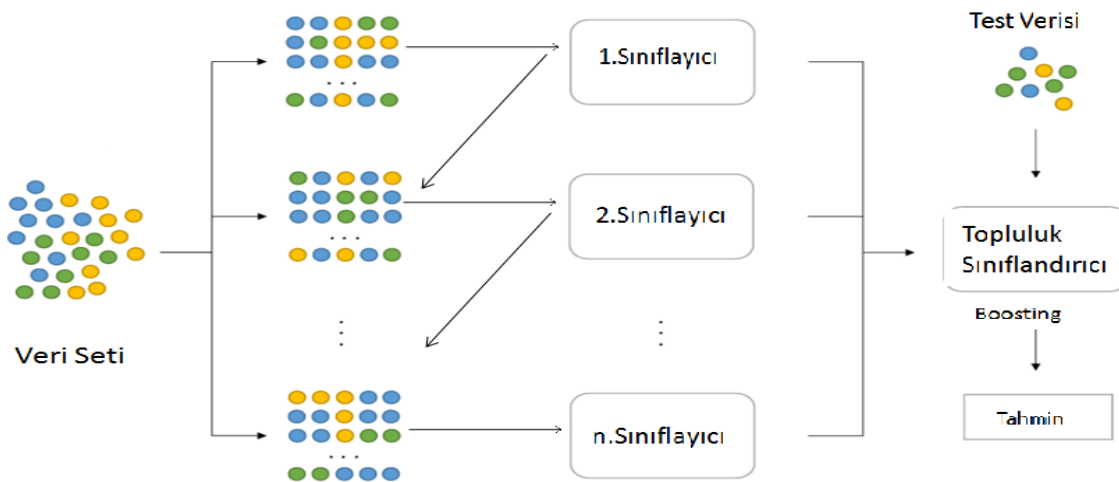
Sayfari ve arkadaşları topluluk öğrenme yöntemleri kullanarak akciğer kanserinde sağkalım'ı tahmin etme konulu çalışmalarında, bireysel sınıflandırıcılar ile topluluk öğrenme yöntemleri kullanılarak performans karşılaştırması gerçekleştirdiler. Sonuçlar ampirik olarak, topluluk yöntemlerinin temel sınıflandırıcılarının performansını değerlendirebildiğini ve kanser sağkalımının analizi için uygun yöntemler olduğunu gösterdi (35).

Faisal ve arkadaşları akciğer kanserinin erken evre tahmini için makine öğrenimi sınıflandırıcılarının topluluk öğrenmeye dayalı bir değerlendirmesi konulu çalışmalarında, makine öğrenimi ve topluluk öğrenme yöntemlerini kullanarak akciğer kanserine yatkın hastaların tahmini, tedavileri ile ilgili karar verme süreçlerinde yardımcı olması bakımından belirleyici bir çalışma olmuştur (36).

2.3. Topluluk Öğrenme Yöntemleri

Boosting ve Adaboost Topluluk Öğrenme

Boosting algoritması, düşük eğitim hatası ile zayıf sınıflandırıcılardan güçlü sınıflandırıcılar elde etme yaklaşımıdır. Boosting basit çoğunluk oyu yöntemini kullanarak zayıf sınıflandırıcılar topluluğunu birleştirmektedir (16). Genel olarak boosting prosedürü oldukça basittir. Bu prosedürde her bir ardışık sınıflandırıcı, daha önce yanlış sınıflandırılmış örnekler üzerine yoğunlaşır. Örneğin ikili sınıflandırmalar için tasarlanmış boosting algoritmasında, herhangi bir D dağılımı üzerinde çalışan zayıf sınıflandırıcı için; eğitim setinin X uzayından seçildiğini ve doğruluk fonksiyonunun f olduğunu düşünelim. X uzayının tek seferde X_1, X_2 ve X_3 olmak üzere üç zayıf sınıflandırıcı kümesi oluştuğu düşünülürse, X_1 sınıflandırıcısı, h_1 eğitim kümesinin rastgele bir alt kümesinde bagging algoritmasına benzer bir şekilde eğitilir. X_2 sınıflandırıcısı, h_2 orjinal veri setinin farklı bir altkümesi ile eğitilir ve yarısı h_1 tarafından doğru, diğer yarısı ise yanlış olarak sınıflandırılır. X_3 sınıflandırıcısı ise h_1 ve h_2 tarafından eğitilmeyen veriler ile eğitilir. Son olarak bu üç sınıflandırıcı çoğunluk oyu (majority voting) ile birleştirilir (14). Boosting topluluk öğrenme adımları şekil 2.2'de verilmiştir.



Şekil 2.2. Boosting Topluluk Öğrenme İşlem Adımları

Adaptiv artırma (adaboosting) algoritması, en çok kullanılan boosting algoritmalarından biridir. Freund ve Schapire tarafından önerilen bu yaklaşım boosting yönteminin geliştirilmiş halidir. Kestirim hızının diğer algoritmalara göre yüksek ve birçok modele uyarlanabiliyor olması kullanımını artırmaktadır. Regresyon problemlerinde de uygulanabilen algoritmanın, AM1, AM2 ve AR algoritmaları en yaygın kullanılanlarıdır (16). Adaboosting algoritmasının çalışma süreci de boosting algoritmasına benzer şekilde, düşük eğitim hatası ile zayıf sınıflandırıcılardan güçlü sınıflandırıcılar elde etme yaklaşımıdır. Yani artık (residuals) optimizasyonuna dayalı performans artırmaya çalışılır.

Algoritma adımları:

1.Adım: Veri setindeki N adet eğitim örneği $\{(x_1, y_1), \dots, (x_N, y_N)\}$ şeklinde verilmiş olsun. $x_i, y_i \in [0,1]$ için x_i ; her bir örneğin sınıf etiketi, y_i ise kullanılan algoritmanın verdiği karardır. Yapılan risk tahmini y_i değeri, 0 ve 1 aralığında reel değer olabilmektedir.

2.Adım: I; toplam iterasyon sayısı olmak üzere, $t = 1, 2, \dots, I$ için;

- Her bir iterasyon için tüm örnekler ele alınarak normalizasyon gerçekleştirilir: $w_i \rightarrow \frac{w_i}{\sum_{t,a} w_{t,a}}$
- Her bir j özniteliği için, h_j sınıflandırıcısı eğitilerek hata oranı, w_t ağırlığına göre hesaplanır: $\epsilon_j = \sum_i w_i |h_j(x_i) - y_i|$
- En düşük ϵ_j hatasına sahip h_j sınıflandırıcı seçilir.
- Ağırlıklar güncellenir: $w_{t+1,i} = w_{t,i} \beta_t^{1-e_i}$
- $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$ olarak hesaplanır.

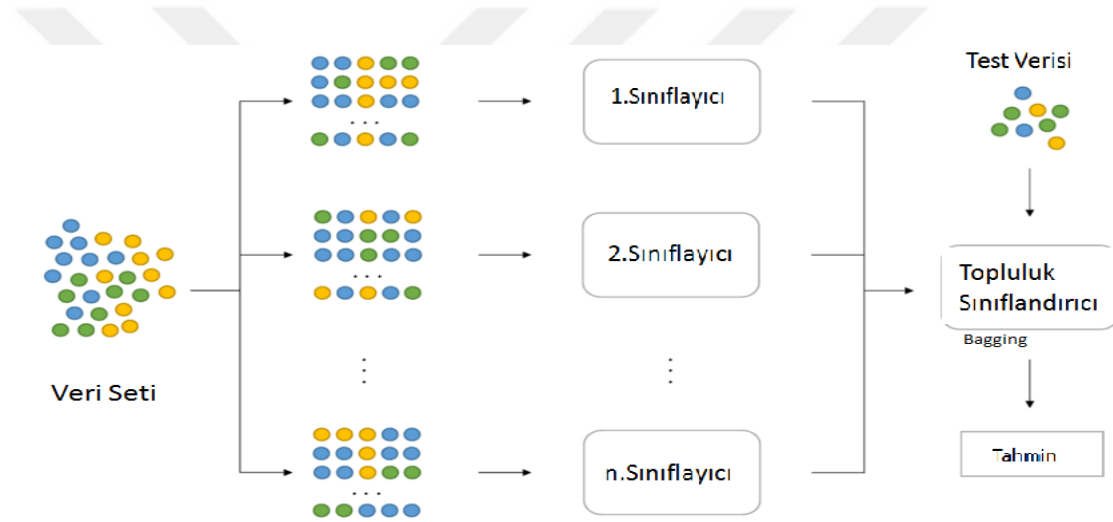
Burada h_j düşük hata değeri ile sınıflandırma yaptıysa $\epsilon_j = 0$ aksi halde 1 olur.

3. Adım: $\alpha_t = \log \frac{1}{1-\beta_t}$ olarak belirlendiğinde, $h(x)$ sınıflandırıcısının son durumu;

$$h(x) = \begin{cases} 1, & \sum_{t=1}^I \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^I \alpha_t \\ 0, & \text{diğer durumlar} \end{cases} \quad (2.1)$$

Bagging Topluluk Öğrenme

Bagging, “bootstrap aggregation” ifadesinin kısaltılmış halidir. Yeniden örneklem yöntemi ile elde edilen sınıflandırıcıların bir araya getirilmesi ile oluşturulan bir algoritmadır. Bu yöntemde veri seti içerisinde yerini koyacak şekilde tekrar tekrar örnekler çekilerek yeni sınıflandırıcılar oluşturulur ve bu sınıflandırıcılardan topluluk öğrenme algoritması geliştirilir. Ortaya çıkan sınıflandırıcıların hepsi modelleme için kullanılarak her birine bir oylama yapılır ve tekil sınıflandırıcılar çoğunluk oyu için birleştirilmeden önce küçük setler ile eğitilir (37). Bagging algoritmasının farklı bir oluşumu, rastgele orman (random forest) algoritmasıdır. Bu algoritma bagging yöntemi ile eğitilmiş karar ağaçları topluluğudur (15). Bagging topluluk algoritması şekil 2.3’de verilmiştir.



Şekil 2.3. Bagging Topluluk Öğrenme İşlem Adımları

Algoritma adımları:

1.Adım: Veri setindeki N adet eğitim örneği $\{(x_1, y_1), \dots, (x_N, y_N)\}$ şeklinde verilmiş olsun. $x_i, y_i \in [-1, +1]$ için x_i ; her bir örneğin sınıf etiketi, y_i ise kullanılan algoritmanın verdiği karardır. Yapılan risk tahmini y_i değeri, -1 ve 1 aralığında reel değer olabilmektedir.

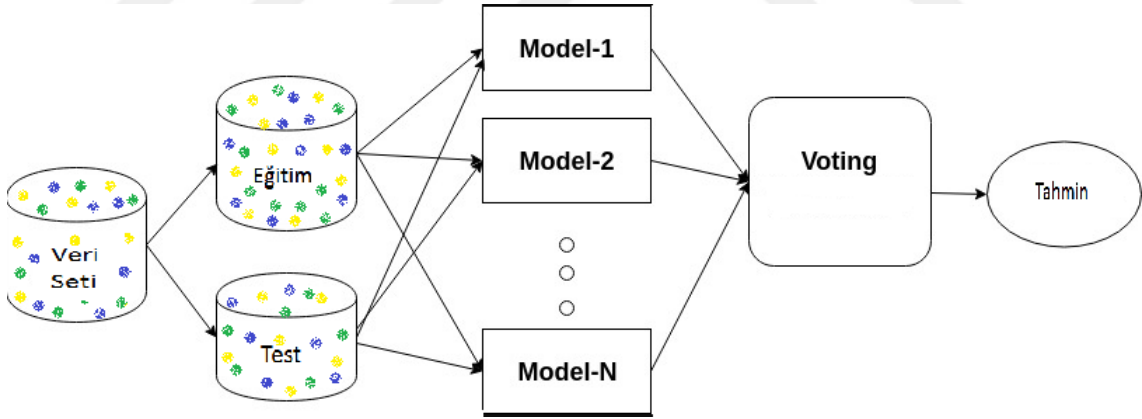
2.Adım: I; toplam iterasyon sayısı olmak üzere, $t = 1, 2, \dots, I$ için;

- b. Bootstrap örnek N_b olmak üzere, N kümesinden x örnek çekilerek yer değiştirilir.
- Her bir N_b bootstrap örneğinden, H_t sınıflayıcısı öğrenilir.

- H_1, H_2, \dots, H_t Sınıflayıcısından çoğunluk oyu yöntemiyle son sınıflama sonucu öğrenilir. $h(x) = \text{sign}(\sum_{t=1}^I H_t(x))$ (16).

Voting Topluluk Öğrenme

Voting yöntemi, nominal sonuçlar için en yaygın kullanılan birleştirme yöntemidir (14). Çoklu makine öğrenme algoritmalarından gelen tahminleri birleştirmenin en basit yollarından biridir. Bu yöntem ile her sınıflandırıcının farklı özelliklerinden yararlanmak için paralel eğitilmiş ve değerlendirilmiş farklı gruplardan oluşan birleştirme işlemi yapılmaktadır (38). Sınıflandırma tahminlerinde farklı türden sınıflayıcıların sınıflandırılması doğruluk performansını artırdığından, voting yönteminde çeşitliliğin sağlanması durumunda doğruluk oranı artacaktır (16). Boosting ve Bagging yöntemlerinde benzer sınıflayıcıdan fazla sayıda sınıflayıcı, veri setinin farklı alt kümelerini eğitmektedir. Voting yönteminde ise aynı veri setinin, farklı türdeki sınıflayıcılar tarafından sınıflandırılması söz konusudur (21). Voting yönteminde en sık kullanılan birleştirme kuralı, Çoğunluk Oyu (majority voting)'dir (16). Sürekli sonuçlar türeten sınıflandırıcılar için ortalama ve ağırlıklı ortalama alma kuralları da kullanılmaktadır (14). Voting topluluk algoritması şekil 2.4'de verilmiştir.



Şekil 2.4. Voting Topluluk Öğrenme İşlem Adımları

Ortalama Alma

Ortalama alma, en güncel kullanılan birleştirme metotları arasında yer almaktadır.

$$H(x) = \frac{1}{T} \sum_{i=1}^T h_i(x) \quad (2.2)$$

olarak hesaplanmaktadır (14).

Ağırlıklı Ortalama Alma

Ağırlıklı ortalama alma, her bir sınıflandırıcı sonucunu önem düzeyine göre farklı ağırlıklarla oranlayan birleştirme yöntemidir.

$$H(x) = \frac{1}{T} \sum_{i=1}^T w_i h_i(x) \quad (2.3)$$

Olarak hesaplanmaktadır. $w_i \geq 0$ ve $\sum_{i=1}^T w_i = 1$ (14).

Majority Voting

Majority voting en yaygın kullanılan voting yöntemidir. Burada her sınıflandırıcı bir sınıf etiketine oy verir ve oyların yarısından fazlasını alan etiket çıktı etiketi olur. Sınıf etiketlerinden hiçbirisi oyların yarısından fazlasını alamaz ise birleşik sınıflandırıcı tahmin yapmaz (14).

$$H(x) = \begin{cases} c_j, & \sum_{i=1}^T h_i^j(x) > \frac{1}{2} \sum_{k=1}^l \sum_{i=1}^T h_i^k(x) \\ \text{null, diğer durumlarda} & \end{cases} \quad (2.4)$$

Sınıflandırıcı bulguları bağımsız ise majority voting yöntemi en iyi birleştirme kuralı olarak ifade edilmektedir. T sınıflandırıcı sayısı olmak üzere her sınıflandırıcının doğru sınıflama olasılığı p olduğu düşünülürse, topluluk yönteminin karar vermek için en az $\frac{T}{2} + 1$ sınıflandırıcının doğru etiketi seçmesi gerekmektedir. Topluluk yöntemlerinin doğruluğu binom dağılım göstermektedir (16). $k \geq \left(\frac{T}{2}\right) + 1$ gerçekleşme olasılığı, doğru sınıfları döndüren T sınıflandırıcıya bağlıdır. Her sınıflandırıcının bir p olasılıkla başarı oranı olduğu düşünüldüğünde topluluk öğrenme başarısı;

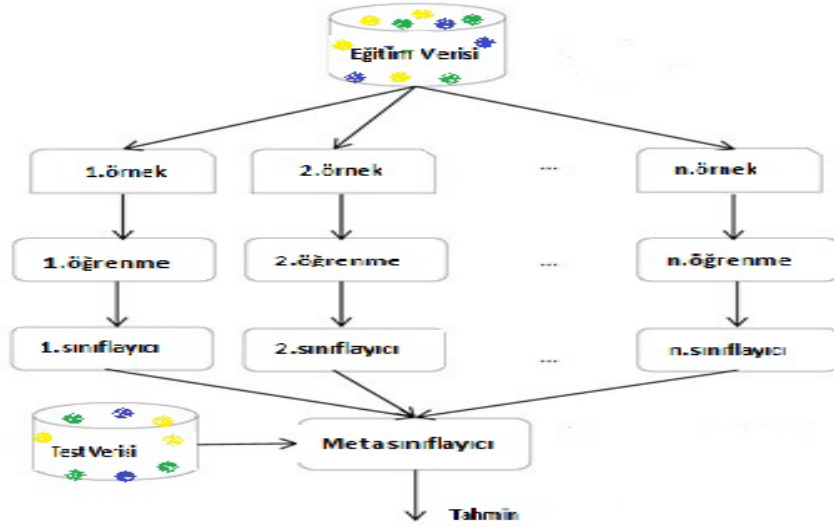
$$p_{ens.} = \sum_{k=\frac{T}{2}+1}^T \binom{T}{k} p^k (1-p)^{T-k} \quad (2.5)$$

Eşitliği ile ifade edilir (16).

Stacking Topluluk Öğrenme

Farklı türden sınıflayıcıların tahminleri ile yüksek doğruluk performansı sağlamak amacıyla sınıflayıcı tahminlerini meta sınıflayıcı için girdi kabul eden bir yöntemdir. Bu yöntemde, eğitim veri setinden farklı sınıflayıcılar ile tahminler elde edilirken, sonrasında ise elde edilen tahminler meta sınıflayıcısında birleştirilir ve topluluk öğrenme modeli

tahmin edilir (39). Stacking öğrenme algoritması şekil 2.5'de gösterilmiştir.

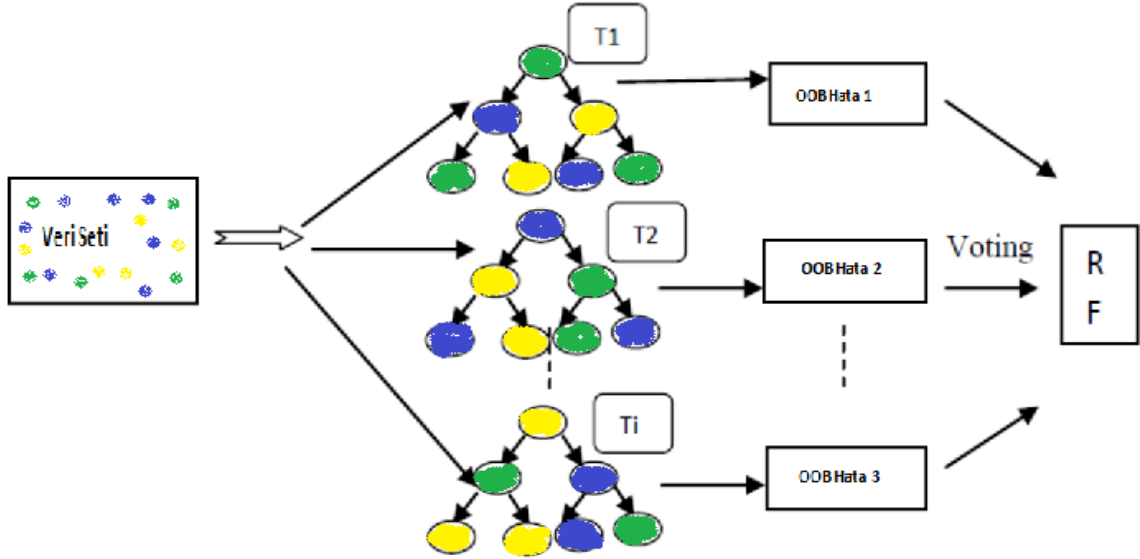


Şekil 2.5. Stacking Topluluk Öğrenme

2.4. Sınıflayıcılar

Random Forest Sınıflayıcısı

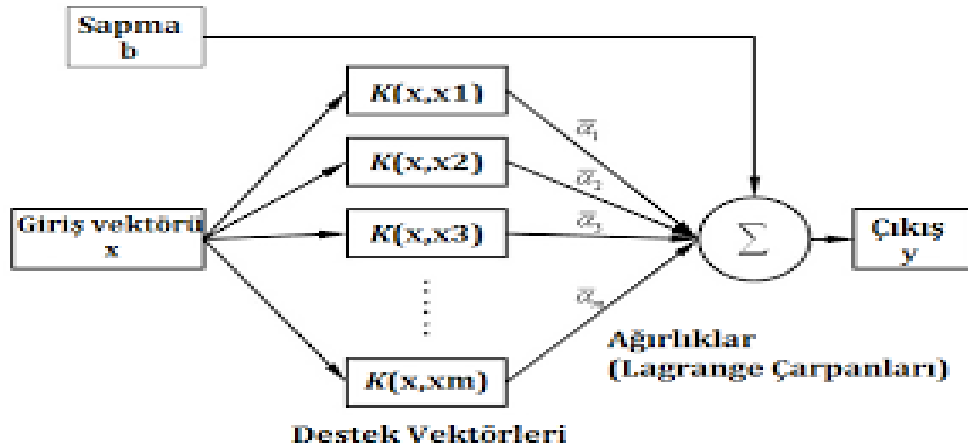
Random forest algoritması, çok sayıda karar ağacından oluşan orman sınıflayıcısıdır ve bu yöntemle sınıflama veya regresyon ağaçları kurulabilmekte ve kümeleme yapılabilir. Sınıflama değişkeni kategorik ise sınıflama tabanlı, sürekli ise regresyon tabanlı ağaçlar kurulmaktadır. Bu yöntemde sınıflama ve regresyon ağaçları (CART) algoritması ile ağaçlar oluşturulur ve ağaçlar budanmaz. CART algoritması veri setinin hangi değişkenden başlayarak dallara ayrılacağına "bilgi kazancını" kullanarak karar verir. Ayrıca dallara ayrılmak için seçilen değişkenin uygun test kriterleri (cut off değeri) "gini katsayısı" ile belirlenir. Genelde veri seti test ve öğrenme olarak iki bölüme ayrılır. Veri seti ayrılmamış olsa da veri seti bir bütün olarak ele alınarak model kurulmakta ve modelin performansı modelin iç hatası olarak da bilinen Out Of Bag (OOB) hata oranı ile ölçülebilmektedir. Oluşturulan her ağaca önceden hesaplanan OOB hata oranına göre bir ağırlık verilir. En düşük hata oranına sahip ağaç en yüksek ağırlığı, en yüksek hata oranına sahip ağaç en düşük ağırlığı alır. Her ağaç belirlenen ağırlığa göre yaptığı sınıf tahmini için bir oy verme işlemine tabi tutulur. Ağaçların oyu oluştuktan sonra random forest algoritmasında bu ağırlıklı oylar toplanır (38). Rastgele orman modelinin hiper parametreleri, minimum kazanç, minimum yaprak boyutu, sp için minimum boyut ve grid search optimizasyon algoritması ile kalibre edilen ön budama sayısıdır. Random forest sınıflayıcısı şekil 2.6'da verilmiştir.



Şekil 2.6. Random Forest Sınıflayıcısı

Ardışık Minimal Optimizasyon (SMO) Sınıflayıcısı

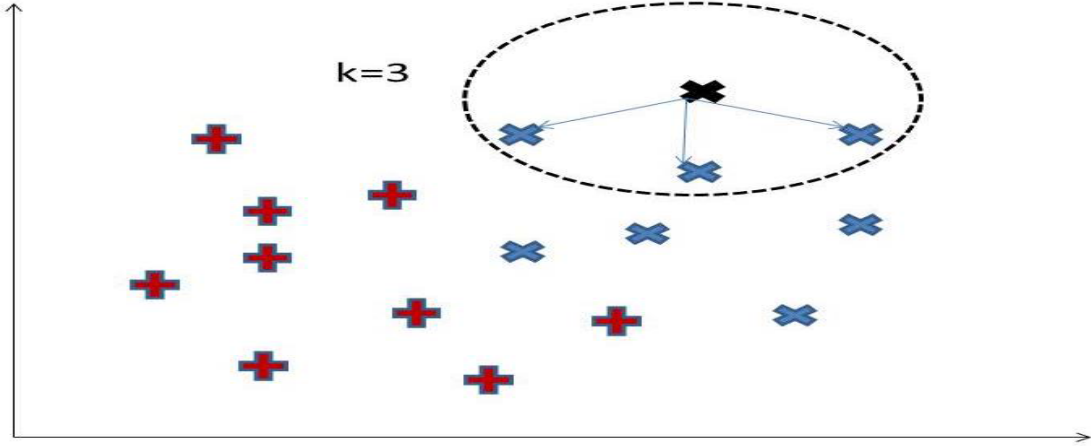
SMO sınıflayıcısı, esasında bir çeşit SVM algoritmasıdır. İki Lagrange çarpanının yer aldığı en küçük optimizasyon probleminin her aşamada destek vektör makinesini çalıştırarak model tahmini yapar (40). SVM, en iyi düzlemleri belirlemek için farklı çekirdek fonksiyonları ile sınıflandırma ve tahmin problemlerinde hedef değişkene ilişkin doğrusal ve doğrusal olmayan tahminler üretir. En uygun çekirdek fonksiyonunun belirlenmesi modelin doğruluk performansı için önemli bir kriterdir (41, 42). Radyal, doğrusal, laplace gibi çekirdek fonksiyonları SVM algoritmasında kullanılmaktadır. SVM, optimizasyon tekniği sayesinde büyük ve karmaşık veri setlerinde optimal çözümler sunmaktadır (43, 44). SVM sınıflayıcısının hiper parametreleri kernel tipi, L pos, L neg ve epsilondur. SVM sınıflayıcısı şekil 2.7'de gösterilmiştir.



Şekil 2.7. SVM Sınıflayıcısı

K-NN (K-En Yakın Komşu) Sınıflayıcısı

K-NN sınıflayıcısı, yaygın olarak kullanılan bir makine öğrenimi yöntemidir. K-NN, denetimli öğrenme paradigmasının sınıflandırıcıları gibi, $h(x)$, bilindiği durumda $x \in X$ veri noktalarını içeren $P \subseteq X$ eğitim verisine ihtiyaç duyar. Sınıflandırıcının P bilgisini kullanarak yeni örnek sınıf etiketini tahmin etmesi beklenir. K-NN, iyi performansının yanı sıra basit kullanımı sayesinde makine öğrenme yaklaşımlarında yaygın olarak kullanılmaktadır. Ayrıca parametrik olmayan bir sınıflandırıcı olan K-NN, veri dağıtımı hakkında önceden yapılmış varsayımlara bağlı değildir. Bununla birlikte, başarılı bir sınıflandırma için K-NN sınıflandırıcısında; k değeri, komşuları elde etmek için kullanılan mesafe ölçüsü ve örneklem gibi önemli üç faktör gereklidir (45). K-NN modelinin hiperparametreleri k , ölçüm türü, karma ölçülerdir ve bir grid search algoritması ile optimize edilmiştir. K-NN sınıflayıcısı şekil 2.8'de gösterilmiştir.



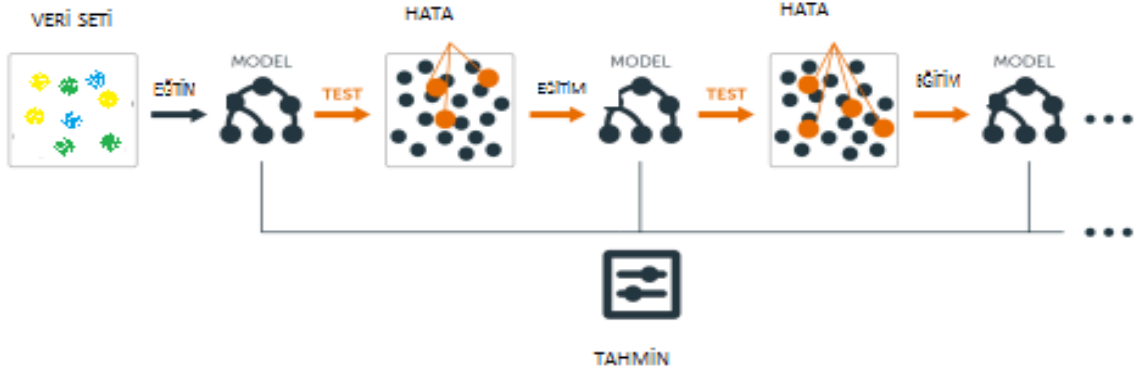
Şekil 2.8. K-NN Sınıflayıcısı

Aşırı Gradyan Artırma (XGBoost) Sınıflayıcısı

Chen ve Guestrin'in tarafından geliştirilen XGBoost sınıflandırıcısı, sınıflandırma ve regresyon problemlerinde kullanılan uçtan uca bir ağaç sisteminde ölçeklendirilebilir makine öğrenme tekniğidir. XGBoost, her biri $K_E^i \setminus i \in 1 \dots K$ düğümlerine sahip CART topluluğu kullanır. Nihai tahmin, her ağaç için tahmin puanlarının toplamıdır.

$$L(\emptyset) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (2.6)$$

XGBoost modelinin hiper parametreleri, grid search optimizasyon algoritması tarafından ayarlanan maksimum derinlik, bölme sayısı ve öğrenme hızıdır. XGBoost sınıflayıcı algoritması şekil 2.9'da verilmiştir.



Şekil 2.9. XGBoost Sınıflayıcısı

2.5. Performans Metrikleri

Yeni geliştirilen bir testin her zaman güvenilir, geçerli ve kullanışlı olması istenilir. Bir testin ölçmek istediği özelliği doğru ölçmesi **geçerlilik**, ölçüm işleminin tekrarlanması durumunda benzer sonuçların elde edilmesi ise **güvenilirlik** olarak tanımlanır (46). Geçerliliğin, kapsam, mantıksal, bir referansa göre ve yapı geçerliliği olmak üzere 4 çeşidi vardır. Yaygın olarak kullanılan geçerlilik yönteminden bir referansa göre geçerlilikte, performansı standart olarak ölçtüğü düşünülen test ile benzer özelliği ölçebileceği düşünülen bir yada daha fazla testin skorları ile karşılaştırılması süreci yer alır. Ölçmek istediği özelliği kesin bir şekilde ölçen teste **altın standart test** denirken, yüksek geçerliliğe sahip ancak ölçmek istediği özelliği her zaman ölçemeyen teste de **referans test** denir. En yaygın iki referans geçerliliği, uyum ve kestirim geçerliliğidir. Uyum geçerliliğinde, bir geçerlilik katsayısı elde edilir ve bu katsayı, veriler sayısal ölçümle sınıflandırıldığında, **korelasyon katsayılarından**, iki kategorili niteliksel veri türünde sınıflandırıldığında ise **duyarlılık ve seçicilik** istatistiklerinden hesaplanır. Kestirim geçerliliğinde ise, standart ölçümü kestiren test ya da değişkenlerden elde edilen bir skoru kullanarak standardın kestirilme derecesi olarak tanımlanır (46).

Karışıklık (Confusion) matrisi, niteliksel testlerin geçerliliğini bulmak adına bir sınıflandırma sistemi tarafından yapılan gerçek ve tahmini sınıflandırmalar hakkında bilgi verir. Model performansı genellikle matristeki veriler kullanılarak hesaplanır (47). Tablo 2.1’de verilmiştir.

Tablo 2.1. Confusion Matrisi (47).

Geliştirilen Yeni Test	Kesin Tanı		Toplam
	Hasta(+)	Sağlam(-)	
Hasta(+)	a	b	a+b
Sağlam(-)	c	d	c+d
Toplam	a+c	b+d	a+b+c+d

Tablo 2.1'de bahsedilen gözeler için;

- a gözü, Gerçekten hasta olup, test sonucunda da hasta olanların sayısını verir ve **doğru pozitif** olarak adlandırılır.
- b gözü, gerçekte sağlam olup, testin yanlışlıkla hasta dediği gözlemlerin sayısını verir ve yanlış pozitif olarak adlandırılır.
- c gözü, gerçekte hasta olup, testin yanlışlıkla sağlam dediği gözlem sayısını verir ve yanlış negatif olarak adlandırılır.
- d gözü, gerçekte sağlam olup, test sonucunda da sağlam olanların sonucunu verir ve doğru negatif olarak adlandırılır (46).

Doğruluk (AC), uyumlu gözlerdeki değerlerin toplam gözlem sayısına bölünmesi olarak tanımlanır ve eşitlik (2.7) ile gösterilir (47).

$$AC = \frac{a+d}{a+b+c+d} \quad (2.7)$$

AC değerinin 1'e yakın olması istenilen bir durumdur. AC değeri 0,50 ve 0,50'nin altında çıkması durumunda, geliştirilen test ile yapılan sınıflandırmanın şans eseri ortaya çıktığı söylenir (46).

Gerçek pozitif oranı (TP), hasta olanların gerçekten hasta olma olasılığını verir ve eşitlik (2.8) ile gösterilir (47). Bu değer aynı zamanda **kesinlik** olarak ifade edilir.

$$TP = \frac{a}{a+b} \quad (2.8)$$

Gerçek negatif oranı (FP), sağlam olanların gerçekten sağlam olma olasılığını verir ve eşitlik (2.9) ile gösterilir (47).

$$FP = \frac{d}{c+d} \quad (2.9)$$

Duyarlılık, testin gerçek hastalar içinden hastaları ayırma yeteneğidir ve eşitlik (2.10) ile gösterilir (48).

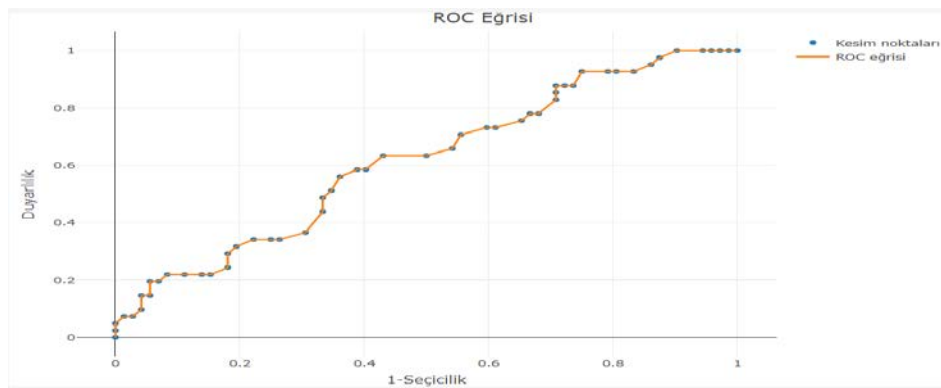
$$\text{Duyarlılık} = \frac{a}{a+c} \quad (2.10)$$

Özgüllük, testin gerçek sağlamlar içinden sağlamları ayırma yeteneğidir ve eşitlik (2.11) ile gösterilir (48).

$$\text{Özgüllük} = \frac{d}{b+d} \quad (2.11)$$

Geçerliliğin iki önemli ölçüsü olan duyarlılık ve seçiciliğin 1 ya da 1'e yakın olması istenilen bir durumdur (46). Klinik çalışmalarda, çeşitli tanı yöntemleri ve laboratuvar testlerinin sonuçlarına dayanarak hasta ve sağlıklı bireylerin ayrımının yapılması amaçlanabilir. Bu durumda bir testin/yöntemin, hasta bireyleri sağlıklılardan ne kadar doğrulukla ayırt edebildiğinin bilinmesi çok önemlidir. Tıbbi karar verme sürecinde, testin ayırt ediciliğini belirlemek ve tanı koyabilmek için en uygun kesim noktasının belirlenmesi gerekir. Bunun için kullanılan yöntemlerden biri de işlem karakteristik (Receiver Operating Characteristic [ROC] Curve) eğrisi yöntemidir (11).

ROC Eğrisi yaklaşımı, bir hastalığı teşhis etmede kullanılan tanı testlerinin performans değerlendirmesinde ve kesim noktalarının belirlenmesinde kullanılır. ROC eğrisi çizmek için, gözlemlerin altın standart test ile kesin olarak belirlenmesi gereklidir. Test sonuçları sürekli sayısal veri türünden elde edildiğinde, verideki tüm değerler kesim noktası olarak alınıp, her değer için duyarlılık ve seçicilik değerleri yer alır. ROC eğrisinin dikey ekseninde duyarlılık ve yatay ekseninde 1-seçicilik değerleri bulunur (46).



Şekil 2.10. ROC Eğrisi Grafiği (11).

En iyi kesim noktasının belirlenmesinde eğer duyarlılık ya da seçiciliğin belli bir değere eşit olması yönünde bir beklenti yoksa duyarlılık + seçicilik değerinin en büyük olduğu nokta kesim noktası olarak belirlenir. Bazen de eğrinin sol üst köşesine en yakın olan nokta kesim noktası olarak seçilir. ROC eğrisi altında kalan alanın alabileceği en büyük değer 1'dir. Alanın 1'e yakın olması testin bireyleri kesin olarak ayırt edebildiğini gösterir (46).

2.6. Yeniden Örnekleme (Resampling) Yöntemleri

Örnekleme yöntemi, genel anlamda minimum örneklem hatası ile kitleyi en iyi şekilde temsil edecek şekilde örnek seçme işlemidir. Klasik parametrik testler, örneklem sayısı az olduğunda güvenilir sonuçlar vermemekte ve aynı zamanda parametrik yöntemlerin varsayımlarını da bozabilmektedir. Klasik parametrik testler, gözlemlenen istatistikleri teorik örnekleme dağılımlarıyla karşılaştırırlar. Yeniden örnekleme yönteminde ise, teorik dağılımlardan farklı olarak çıkarım yapmak için aynı örneklem içinde tekrarlanan örnekleme işlemi yapılmaktadır. Son zamanlarda verilerin parametrik yöntemlerle değerlendirilmesinin uygun olmadığı durumlarda yeniden örnekleme yöntemleri kullanılmaya başlanmıştır. Kullanım amacına yönelik en az 4 farklı yeniden örnekleme yöntemleri bulunmaktadır. Bunlar randomizasyon kesinlik testi (randomization exact test), çapraz geçerlilik (cross validation), jackknife ve bootstrap yöntemleridir (49).

Randomize Kesinlik Testi

Fisher tarafından geliştirilen randomize kesinlik testi, permütasyon testi olarak da bilinmektedir. Gözlemlenen verilerin tüm olası yeniden düzenlemeleri altında, H_0 hipotezi altındaki test istatistiği dağılımının olası değerleri hesaplanarak yapılmaktadır (49).

Algoritma Adımları

1.Adım: Orijinal örneklem üzerinden test istatistiği hesaplanır

2.Adım: Rastgele örnekler ile örneklem çerçevesi hazırlanır,

3.Adım: N adet yeniden örneklendirilmiş bir grup oluşturarak örneklem çerçevesi değiştirilmeden örnekler bu gruplara eşit dağıtılır ve yeniden örneklendirilmiş B adet örnekleri oluşturmak için işlem tekrarlanır ($B = N!/(n_1! n_2!)$).

4. *Adım*: 3. adımda oluşturulan her yeniden örneklendirilmiş örnek (B adet) için eklenti (plug-in) bağımsız örneklem t_{yeni} istatistiği hesaplanır,

5. *Adım*: 4. Aşamada hesaplanan t_{yeni} değeri ile 1. aşamadaki t_{orj} değeri karşılaştırılır (50). Çift yönlü p değeri,

$$p = \frac{|t_{yeni}| \geq |t_{orj}|}{\binom{N}{n_2}} = \frac{|t_{yeni}| \geq |t_{orj}|}{B} \quad (2.12)$$

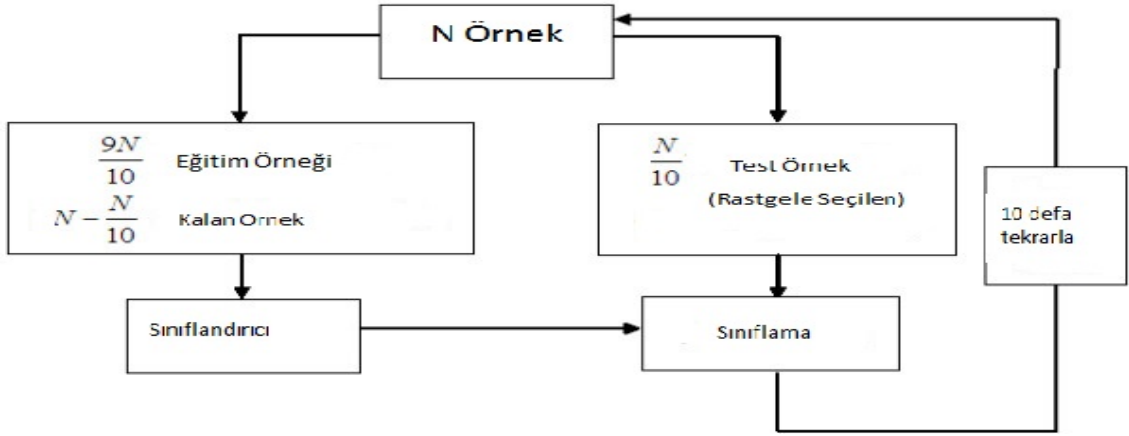
Çapraz Geçerlilik (Cross Validation)

Çapraz geçerlilik, bir veri örneğindeki makine öğrenimi modellerini değerlendirmek için kullanılan yeniden örnekleme yöntemidir. Çapraz doğrulamada, modelin daha önce görmediği yeni verileri tahmin etme yeteneğini test etmek, aşırı uyum veya seçim yanlılığı gibi sorunları ortadan kaldırmak genel amaçtır (51).

Algoritma Adımları

1. *Adım*: Veri seti rastgele eğitim ve test setlerine bölünür
2. *Adım*: Model eğitim setinde eğitilir
3. *Adım*: Model test setinde doğrulanır
4. *Adım*: Doğrulama sonucu kaydedilir.

Prosedürün, belirli bir veri örneğini bölmek için grup sayısını ifade eden k adında tek bir parametresi vardır. En yaygın kullanımı, 10 kat çapraz geçerlilik yöntemidir. 10 kat çapraz geçerlilikte, modeller on farklı kez eğitilir ve test edilir ve daha sonra, ortalama performans ölçütleri (yani, doğruluk, kesinlik vb.) sürecin sonunda tahmin edilir (52). Şekil 2.11'de verilmiştir.



Şekil 2.11. k-Kat Çapraz Geçerlilik

Jackknife Yeniden Örnekleme Yöntemi

Kitle parametrelerinin tahmin edilmesinde, dar güven aralıkları ve örneklem hatasının en aza indirilmesi amacına yönelik geliştirilmiştir. Bu yöntemde, veri setinde her bir gözlem değerini bir kez dışarıda bırakarak geride kalan gözlemlerden örneklem istatistiklerini hesaplamaya dayanmaktadır. Bu yeniden örnekleme yöntemine birini dışarıda bırakma (leave one out) yöntemi de denmektedir (49). Normal dağılım varsayımı gerektirmektedir.

Algoritma Adımları

1. Adım: Orijinal örneklem üzerinden test istatistiği hesaplanır

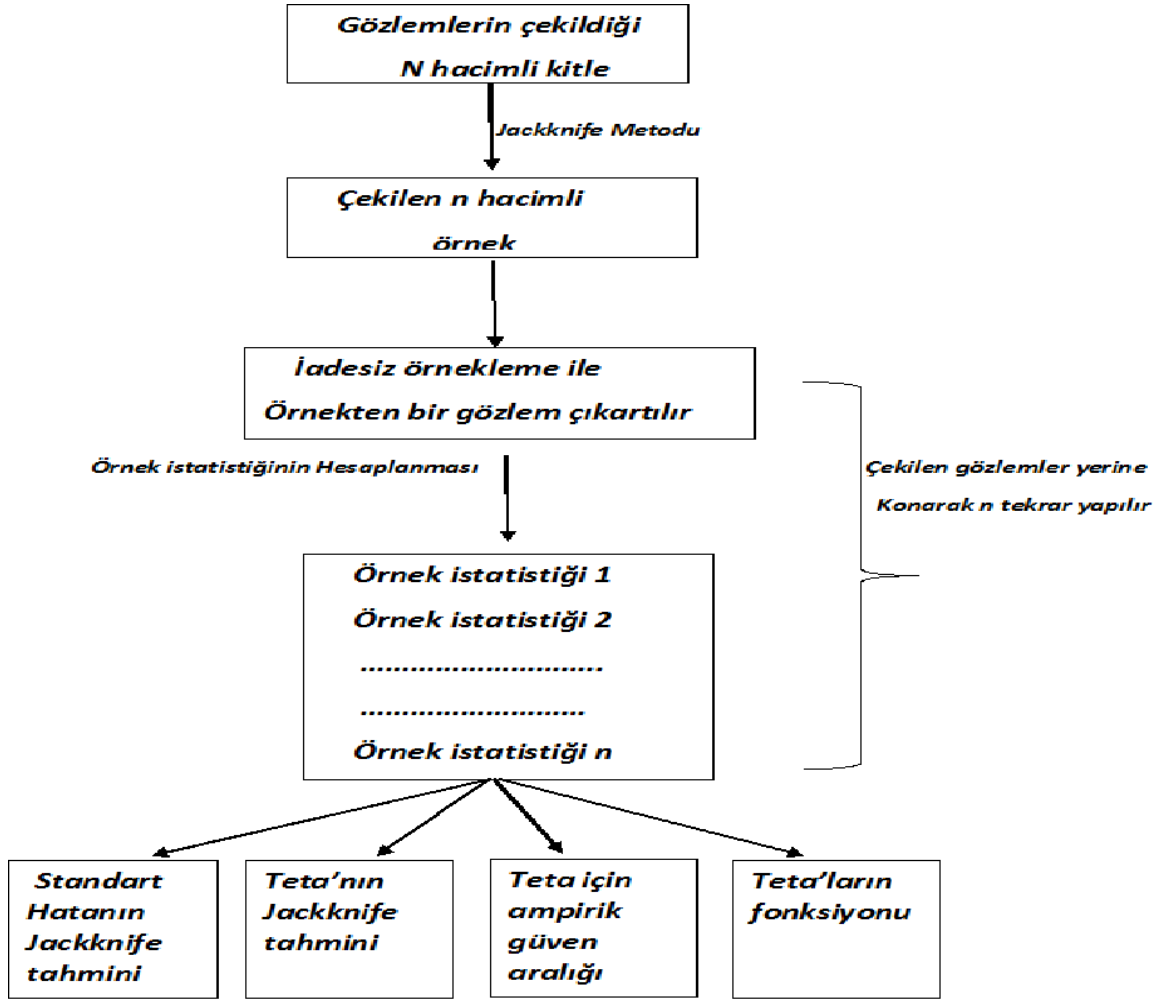
2. Adım: Örnekleme çerçevesini hazırlanır. Jackknife için tek gruplu örneklerle, bu gözlemlenen örnekleri kullanılmasını içerir.

3. Adım: Örnek çerçevesinden 1 adet örnek çıkartılarak M adet (N-1) yeniden örneklendirilmiş jackknife örneği oluşturulur ve B adet örneği oluşturmak için işlem tekrarlanır ($B = \binom{N}{N-1}$).

4. Adım: 3. adımda oluşturulan her yeniden örneklendirilmiş örnek (B adet) için eklenti (plug-in) bağımsız örneklem t_{yeni} istatistiği hesaplanır,

5. Adım: B değeri için standart sapma hesaplanır.

Jackknife yöntemine ait şekil 2.12’de verilmiştir



Şekil 2.12. Jackknife Yeniden Örnekleme (53).

Bootstrap Yeniden Örnekleme Yöntemi

Bu yöntemde jackknife yöntemi gibi, dar güven aralıkları ve örneklem hatasının en aza indirilmesi amacıyla geliştirilmiştir. Bu yöntemde herhangi büyüklükteki mevcut veri setindeki gözlemlerin, şansa bağlı yer değiştirilerek yeniden örneklenmesi ile yeni veri setleri oluşturulmasına dayanmaktadır. Bu yöntemde normal dağılım varsayımı gerekmez (49).

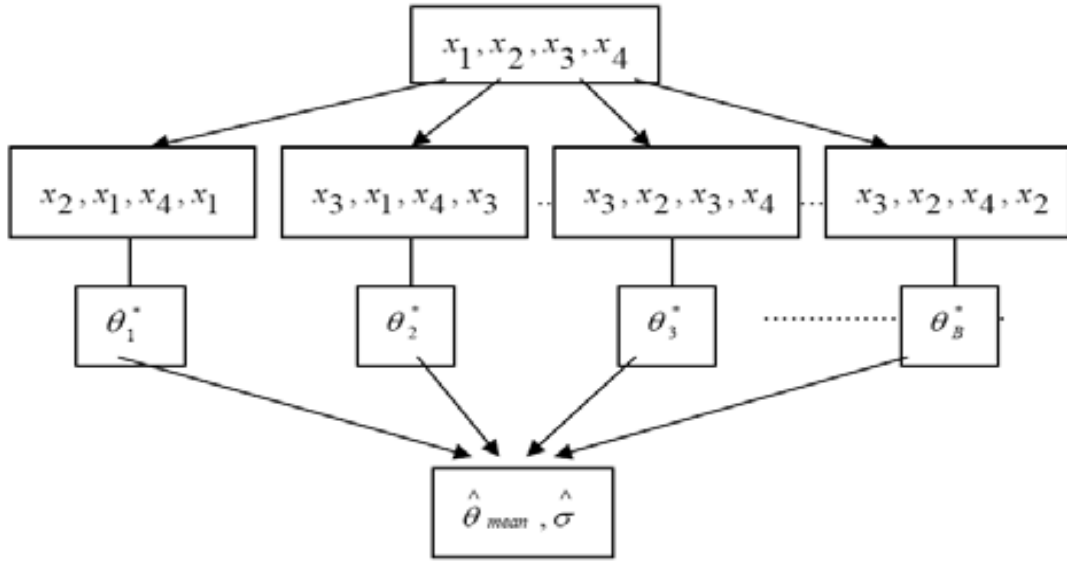
Algoritma Adımları

1. Adım: Orijinal örneklem üzerinden test istatistiği hesaplanır
2. Adım: Rastgele örnekler ile yerine koyma yöntemiyle örneklem çerçevesi hazırlanır.
3. Adım: Bootstrap tahmincisi hesaplanır ($\hat{\theta}$).

4. *Adım*: B adet bootstrap yeniden örnek tahmincisini ($\hat{\theta}_1, \dots \dots \hat{\theta}_B$) elde etmek için 1. ve 2. adımlar birden fazla kez tekrarlanır.

5. *Adım*: B değeri için standart sapma hesaplanır.

Bootstrap yeniden örnekleme yöntemine ait şekil 2.13'de verilmiştir.



Şekil 2.13. Bootstrap Yeniden Örnekleme (53).

3. MATERYAL VE METOT

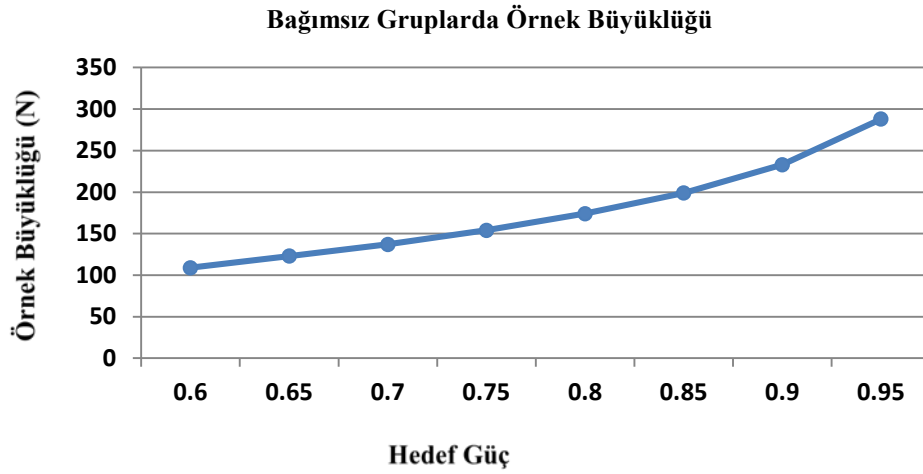
Bu çalışma, <https://www.cancerdata.org/resource/doi:10.17195/candat.2016.04.1> adresli açık kaynak erişimli web sitesinde yer alan KHDAK hastaları ölüm durumunu sınıflandırmak için çeşitli klinik değişkenleri içeren veri setini kullanarak yapılmıştır (5). Veri seti, KHDAK'da hipoksi, iltihaplanma, bağışıklık tepkisi ve tümör yüküyle ilgili kan biyobelirteçlerinin prognostik ve klinik değerleri yönünden incelenecektir. Veri seti toplam 181 inoperabl evre I-IIIB KHDAK hasta kaydını içermektedir. Veri setinin yaklaşık % 55.2'si radyoterapi yada kemoterapi alan hastalardan oluşurken % 44.8'i radikal tedavi gören hastalardan oluşmaktadır.

Tablo 3.1. Çalışmada Kullanılan Değişkenler

	Değişken	Tipi	Açıklaması	Rolü
Demografik	Status	Kategorik	Öldü/Hayatta	Bağımlı/Hedef
	Survival	Nümerik	Beklenen Yaşam Süresi	Bağımsız/Tahminleyici
	Age	Nümerik	Yaş(Yıl)	Bağımsız/Tahminleyici
	Stage	Kategorik	Hastalık Seviyesi (I-IIIB)	Bağımsız/Tahminleyici
	Histology	Kategorik	Hastalık Geçmişi (Adeno/NOS/SCC/Tanımsız)	Bağımsız/Tahminleyici
	Gender	Kategorik	Cinsiyet (1:Bayan,2:Erkek)	Bağımsız/Tahminleyici
Hastalık Öyküsü	WHO-PS	Kategorik	Dünya Sağlık Örgütü Performans Durumu(0:Tamamen aktif,1:Fiziksel olarak yorucu faaliyette sınırlı, 2:ezici ve her türlü öz bakım yeteneğine sahip, ancak herhangi bir iş aktivitesinde bulunamıyor,3:Sadece sınırlı kişisel bakım yeteneğine sahip,4:Tamamen devre dışı. Kişisel bakım yapılamaz.	Bağımsız/Tahminleyici
	FEV1s%	Nümerik	Solunum Testleri: Zorlu Vital Kapasite Manevrası	Bağımsız/Tahminleyici
	Lymph nodes	Nümerik	PET taramasıyla tanımlanan pozitif lenf düğüm sayısı	Bağımsız/Tahminleyici
	RT Protocol	Kategorik	Standart ışın radyasyon tedavisi	Bağımsız/Tahminleyici
	Total dose(1st)	Nümerik	Kemo-Radyoterapide uygulanan 1.toplam doz miktarı	Bağımsız/Tahminleyici
	Total dose(2st)	Nümerik	Kemo-Radyoterapide uygulanan 2.toplam doz miktarı	Bağımsız/Tahminleyici
	GTV	Nümerik	Primer tümör ve metastatik lenf nodlarının toplamı	Bağımsız/Tahminleyici
Kan Biyobelirteçleri	OPN	Nümerik	Osteopontin	Bağımsız/Tahminleyici
	CA-9	Nümerik	Karbonik Anhidraz-9	Bağımsız/Tahminleyici
	IL-6	Nümerik	İnterlökin-6	Bağımsız/Tahminleyici
	IL-8	Nümerik	İnterlökin-8	Bağımsız/Tahminleyici
	CRP	Nümerik	C-Reaktif Protein	Bağımsız/Tahminleyici
	CEA	Nümerik	Tümör yükü karsinoembriyonik antijen	Bağımsız/Tahminleyici
	Cyfra 21-1	Nümerik	Sitokeratin 21-1	Bağımsız/Tahminleyici
	α2M	Nümerik	Alpha-2-Macroglobulin	Bağımsız/Tahminleyici
	sIL2R	Nümerik	Serumda çözümlü IL2 Respektörü	Bağımsız/Tahminleyici
	TLR-4	Nümerik	Toll benzeri Respektör 4	Bağımsız/Tahminleyici
VEGF	Nümerik	Vasküler Endotel Büyüme Faktörü	Bağımsız/Tahminleyici	

3.1. Örneklem Büyüklüğü

Bir bilimsel araştırmaya başlamadan önce çalışmanın planlama aşamasında klinik ve istatistiksel olarak anlamlı sonuçlar elde edebilmek en az kaç birim (denek, hasta, deney hayvanı vb.) gerektiğinin belirlenmesi (teorik güç analizi) ve tamamlanan bir araştırmanın gerçekleşen gücünü ortaya koymak (deneysel güç analizi) oldukça önemlidir. Örneklem büyüklüğünün hesaplanmasında <http://biostatapps.inonu.edu.tr/> adresindeki güç analizinden faydalanmıştır (54). Referans çalışmada, I-IIB evre KHDAK ortalama beklenen yaşam süreleri, 1.91 ± 1.42 (Ortalama \pm SS) olarak alınmıştır. I tip hata (alfa) 0.05 ve II. tip hata (Beta) 0.20 düzeyinde, testin gücü % 80, etki büyüklüğü 0.3 ve alternatif hipotez (H_1) iki yönlü iken bu test kullanılarak anlamlı bir fark bulunabilmesi için gerekli olan minimum örneklem büyüklüğü 174 olarak hesaplanmıştır. Güç analizi dışında, çok değişkenli istatistiksel analiz yöntemleri için farklı kaynaklarda değişken sayısının 5 katı ya da değişken sayısı 6 ve üzeri olduğu durumda $n = 104 + k$ (k: bağımsız değişken sayısı) olması gerektiği belirtilmiştir (20, 46). Bu durumda değişken sayısının 5 katı için her grupta olması gereken örnek büyüklüğü (5×24) 120 iken değişken sayısı 6 ve üzeri olduğu durumda ise yine 130 ($104 + 24$) çıkmıştır.



Şekil 3.1. Her Bir Grup için Önerilen Örneklem Büyüklüğü

3.2. Veri Tabanı Bilgi Keşfi Süreci

Veri tabanı bilgi keşfi sürecinde, veri seçimi (Bağımlı: Status; Bağımsız: Tablo 3.1'deki faktörler), veri ön işleme (yerel aykırı değer faktörü (LOF) ile aykırı / aşırı gözlem ve rastgele orman ile kayıp değer analizi), veri dönüşümü, istatistiksel analizler, veri madenciliği (makine öğrenme yöntemlerinden olan SMO, K-NN, random forest ve XGBoost gibi bireysel sınıflandırıcılar ve performansları ile topluluk öğrenme

yöntemlerinden voting, bagging, boosting ve stacking yöntemleri), değerlendirme (performans ölçütleri) ve sonuçların yorumlanması çalışma boyunca gerçekleştirilir (55).

Veri Seçimi

Veri seti, demografik (status, survival, age, cinsiyet ve stage) ve hastalık öyküsü (histology, WHO-PS, FEV1s%, lymph nodes, RT protocol, Total dose(1st), Total dose(2st) ve GTV) değişkenleri dışında günlük klinik uygulamada ölçülebilen ve tedavi sonuçlarıyla ilişkili olduğu gösterilen kan biyobelirteçlerinden; hipoksi [osteopontin (OPN) ve karbonik anhidraz IX (CA-IX)] süreçleriyle ilişkili olanlar ve inflamasyon [interlökin 6 (IL-6), IL-8 ve C-reaktif protein (CRP)] ve tümör yükü [karsinoembriyonik antijen (CEA) ve sitokeratin fragmanı (CYFRA 21-1)] oluşmaktadır. Ek olarak immünolojik yanıtla ilgili olan alfa-2-makroglobulin (a2M), serum IL-2 reseptörü (sIL2R), geçiş ücreti benzeri reseptör 4 (TLR4) ve vasküler endotelial büyüme faktörü (VEGF) biyobelirteçleri de eklendi. Veri seti 1 bağımlı, 23 bağımsız değişkenden oluşmaktadır.

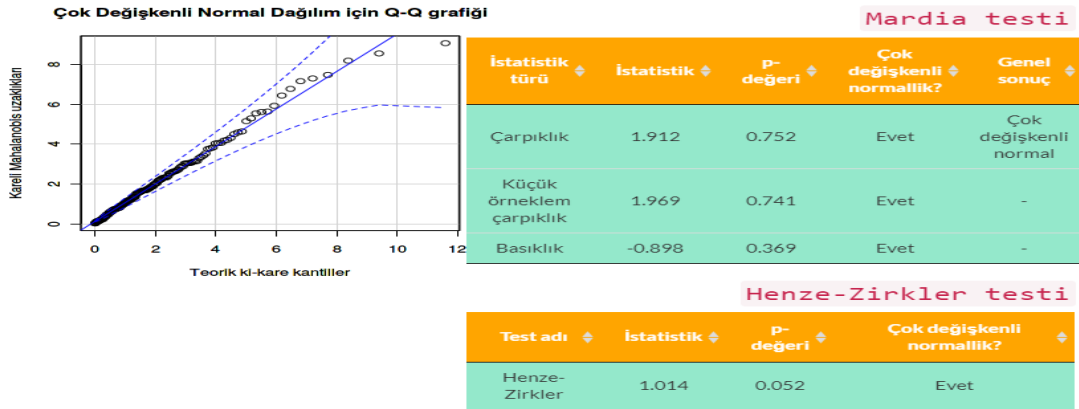
Veri Önleme

Veri setindeki değişkenlere öncelikle kayıp değer analizi uygulanmıştır. Kayıp değerlere random forest atama yöntemi ile rastgele değerler atanmıştır. Ardından veri setindeki aşırı ve aykırı değerlere bakılmıştır. Lokal aykırı faktör (LOF) ile her gözlem için kendine yakın nesne uzaklıkları hesaplanarak toplam 14 adet aykırı ya da uç değer veri setinden silinmiştir. Ayrıca çalışmada modele katkısı olmayan değişkenler veri setinden çıkartılarak özellik seçim işlemi gerçekleştirilmiştir. En uygun değişken seçim kriteri cezalandırılmış lojistik regresyon yöntemi olurken ceza kriteri olarak lasso seçilmiştir. Toplam 16 değişken önemli bulunurken 7 değişken analizden çıkartılmıştır.

Veri Dönüşümü

İstatistiksel çalışmalarda normallik varsayımı önemli bir kriterdir. Öncelikle veri setine çok değişkenli normallik testlerinden Mardia'nın çarpıklık basıklık ve Henze-Zirkler testleri uygulanmıştır. Veri setinin çok değişkenli normal dağılım göstermemesi dolayısıyla veri setindeki nicel değişkenlere veri dönüşümü uygulanmıştır. Her bir değişken için Pearson P test istatistiği hesaplanarak en düşük test istatistiğine sahip dönüşüm seçilmiştir. Survival ve GTV değişkenleri için Box-Cox dönüşümü, age değişkeni için Yeo-Johnson dönüşümü uygulanırken, FEV1s% ve Total dose(2st)

değişkeni için Lambert-W dönüşümü uygulanmıştır. Dönüşüm sonrası veri seti şekil 3.2'de olduğu gibi çok değişkenli normal dağılım göstermiştir.



Şekil 3.2. Çok Değişkenli Normal Dağılım

İstatistiksel Analiz

Nicel veriler, standart sapma ile aritmetik ortalama, min ve maksimum değerlerle medyan ve nitel veriler yüzdeye göre sayı olarak özetlendi. Verilerin çoklu normal dağılıma uygunluğu sonrasında, normal dağılım gösteren gruplarda, gruplar arasındaki farklılık bağımsız örneklerde t testi ile ve normal dağılım göstermeyen değişkenlerde Mann-Whitney U testi incelenmiştir. Kategorik değişkenler arasındaki frekans dağılımı ilişkisi Ki-kare testi ve Fisher kesin Ki-kare testi ile değerlendirilmiştir. İstatistiksel analizde, IBM SPSS sürüm 22 (56), RStudio sürüm 1.1.463 (57) ve Rapid Miner Studio sürüm 8.1.001 (58) kullanılmıştır.

Veri Madenciliği

Bu çalışmada, makine öğrenme yöntemlerinden olan SMO, K-NN, random forest ve XGBoost gibi bireysel sınıflandırıcılar ve performansları ile topluluk öğrenme yöntemlerinden voting, bagging, boosting ve stacking yöntemleri kullanılarak farklı sınıflandırıcıların performansları incelenmiştir.

Verilerin Değerlendirilmesi ve Yorumlanması

Veri setinde sınıflayıcıların ve topluluk yöntemlerinin performanslarının değerlendirilmesinde 10-katlı çapraz geçerlilik yönteminden yararlanılmıştır. Sınıflayıcıların ve topluluk yöntemlerinin sınıflandırma performanslarının değerlendirilmesinde doğruluk, duyarlılık, özgüllük, AUC (Roc Eğrisi altında kalan alan) ve kesinlik metrikleri ile değerlendirilmiştir (59).

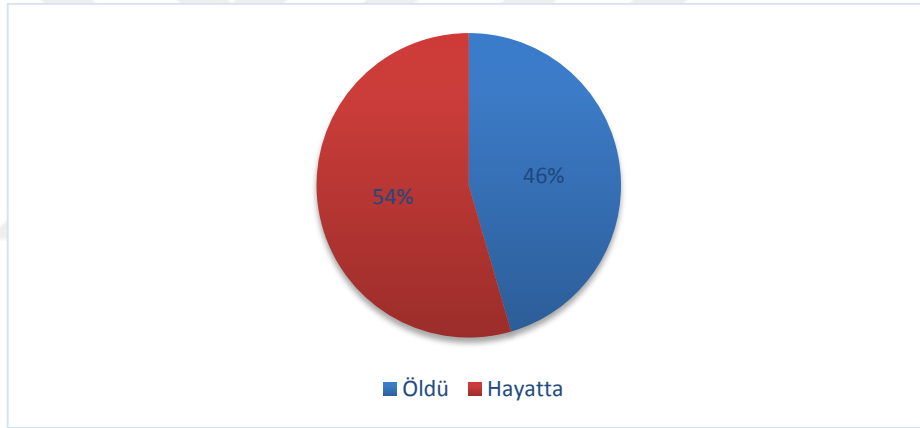
4. BULGULAR

4.1. Tanımlayıcı İstatistikler

Veri seti, 76 (% 46) akciğer kanserinden ölen hasta ile 91 (% 54) akciğer kanseri olan hastalardan oluşmaktadır. Durum dağılımları tablo 4.1 ve Şekil 4.1' de belirtilmiştir.

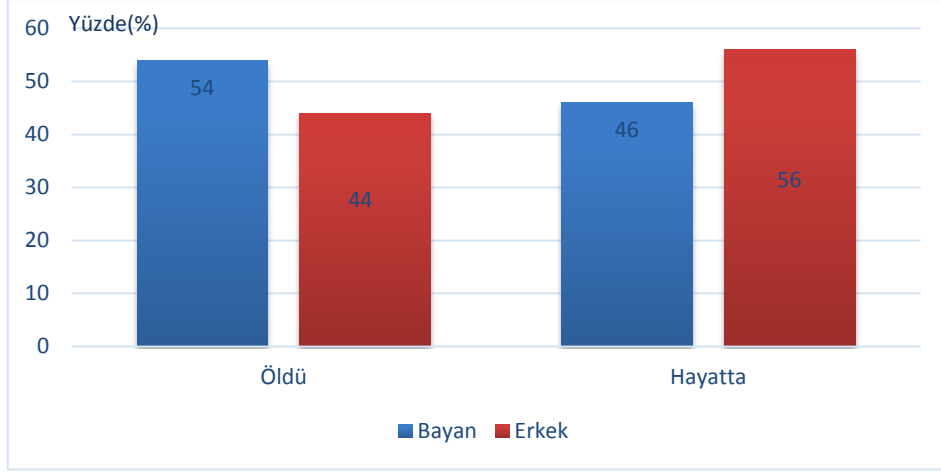
Tablo 4.1. Akciğer Kanseri Sağkalım Durumu

Akciğer Kanseri	Sayı	Yüzde (%)
Öldü	76	46
Hayatta	91	54
Toplam	167	100.0



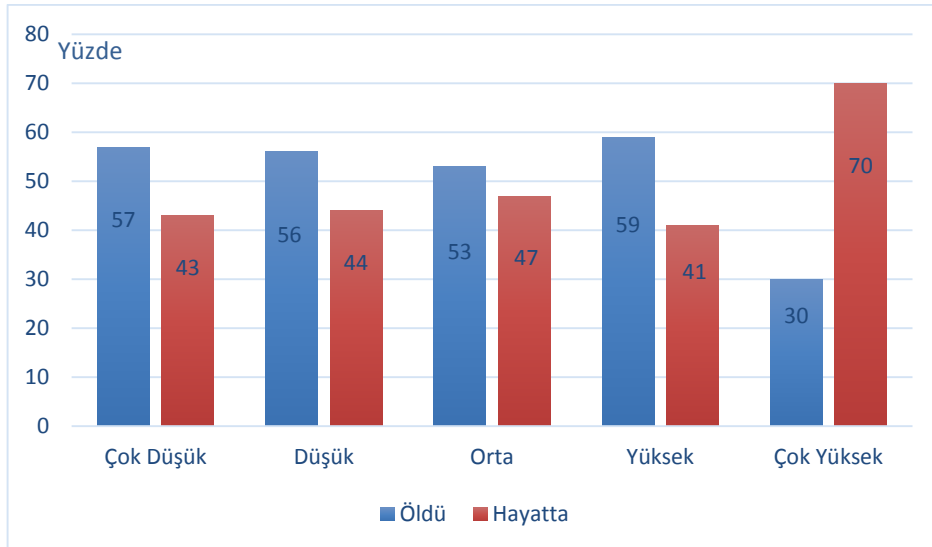
Şekil 4.1. Akciğer Kanseri Durumu

Şekil 4.2 ve tablo 4.2'de akciğer kanserinin cinsiyete göre dağılımına bakıldığında, bayanların % 54'ü kanser nedeniyle ölürken, % 61'i hayatta kalmıştır. Erkeklerin % 44'ü kanser nedeniyle ölürken, % 56'sı hayatta kalmıştır. Akciğer kanseri nedeniyle ölüm ve hayatta kalma oranlarında cinsiyete göre bir fark görülmemiştir ($p>0.05$).



Şekil 4.2. Cinsiyete Göre Akciğer Kanseri Durumu

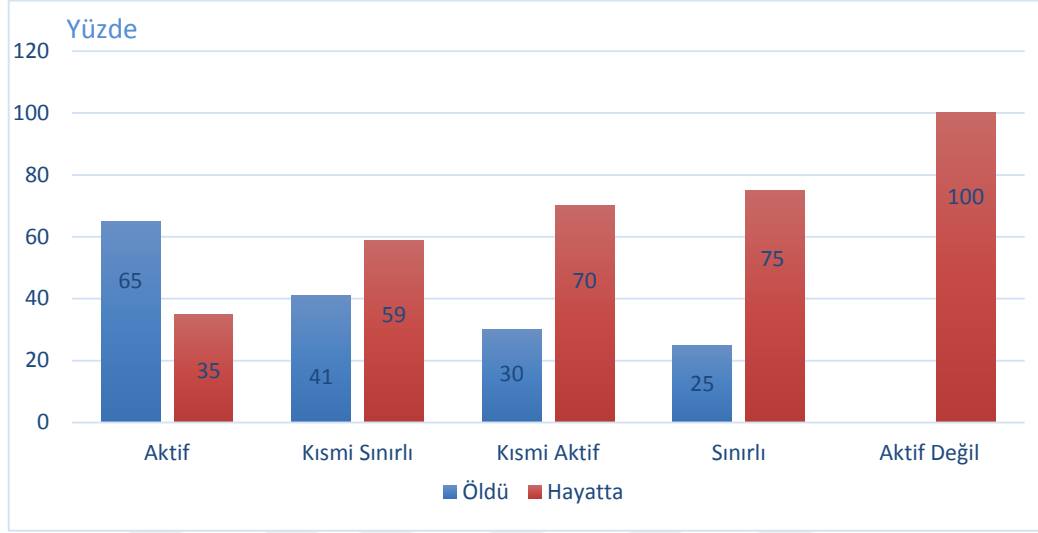
Şekil 4.3 ve tablo 4.2’de akciğer kanserinin lymphnodes değişkenine göre dağılımına bakıldığında, çok düşük olanların % 57’si ölürken % 43’ü hayatta kalmış, düşük olanların % 56’sı ölürken % 44’ü hayatta kalmış, orta olanların % 53’ü ölürken % 47’si hayatta kalmış, yüksek olanların % 59’u ölürken % 41’i hayatta kalmış ve çok yüksek olanların % 30’u ölürken % 70’i hayatta kalmıştır. Akciğer kanseri nedeniyle ölüm ve hayatta kalma oranlarında lymphnodes değişkenine göre bir fark görülmemiştir ($p>0.05$).



Şekil 4.3. Lymphnodes Değişkenine Göre Akciğer Kanseri Durumu

Şekil 4.4 ve tablo 4.2’ye göre, akciğer kanserinin WHO-PS değişkenine göre bakıldığında aktif olanların % 65’i ölürken % 35’i hayatta kalmış, kısmi sınırlı olanların % 41’i ölürken % 59’u hayatta kalmış, kısmi aktiflerin % 30’u ölürken % 70’si hayatta

kalmış, sınırlı olanların % 25'i ölürken % 75'i hayatta kalmış ve aktif olmayanların % 100'ü hayatta kalmıştır. Akciğer kanseri nedeniyle ölüm ve hayatta kalma oranlarında WHO-PS değişkenine göre bir fark görülmemiştir (p>0.05).



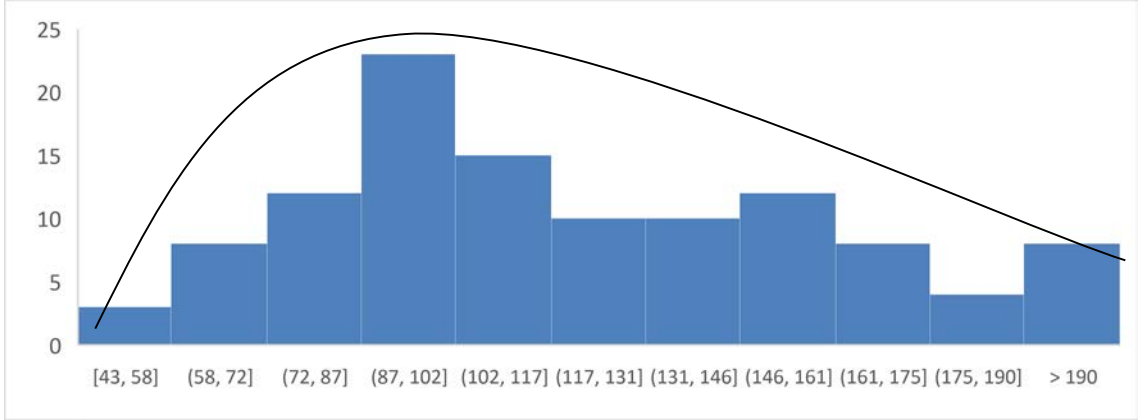
Şekil 4.4. WHO-PS Değişkenine Göre Akciğer Kanseri Durumu

Tablo 4.2. Kategorik Değişkenlere Göre Akciğer Kanseri Durumu

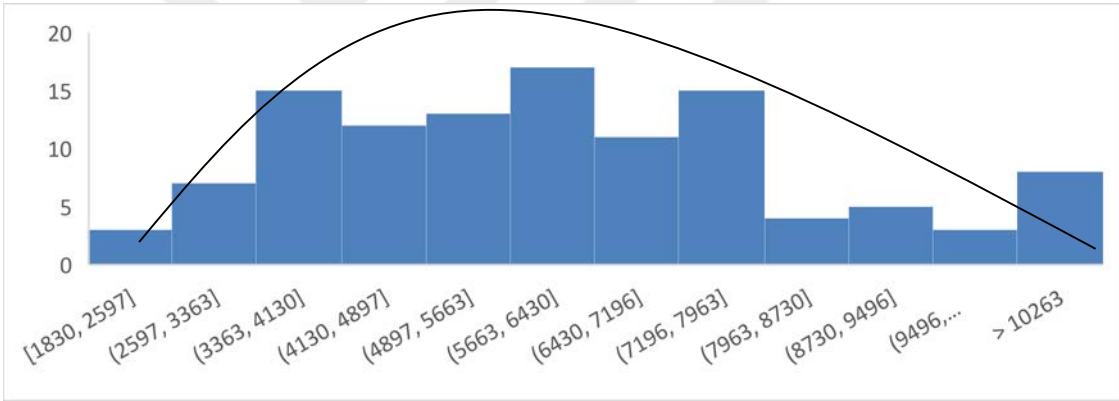
Cinsiyet (Sayı (yüzde))	Öldü	Hayatta	Toplam	X ²	p-değeri
Bayan	52 (54%)	61 (46%)	113 (100%)	0.036	0.849
Erkek	24 (44%)	30 (56%)	54 (100%)		
Toplam	76 (46%)	91 (54%)	167 (100%)		
Lymphnodes (Sayı (yüzde))				0.036	0.849
Çok Düşük	18 (57%)	24 (43%)	42 (100%)		
Düşük	15 (56%)	12 (44%)	27 (100%)		
Orta	17 (53%)	15 (47%)	32 (100%)		
Yüksek	13 (59%)	9 (41%)	22 (100%)		
Çok Yüksek	13 (30%)	31 (70%)	44 (100%)		
Toplam	76 (46%)	91 (54%)	167 (100%)		
WHO-PS (Sayı (yüzde))				0.036	0.849
Aktif	30 (65%)	16 (35%)	46 (100%)		
Kısmi Sınırlı	39 (41%)	57 (59%)	96 (100%)		
Kısmi Aktif	6 (30%)	14 (70%)	20 (100%)		
Sınırlı	1 (25%)	3 (75%)	4 (100%)		
Aktif Değil	0 (0.0%)	1 (100%)	1 (100%)		
Toplam	76 (46%)	91 (54%)	167 (100%)		

Tablo 4.3'deki karşılaştırma sonuçlarına göre akciğer kanseri nedeniyle ölen ve hayatta kalan hastaların OPN ve SIL2R değişkenleri açısından farklılık istatistiksel olarak

anlamli bulunmuştur ($p < 0.05$). OPN ve SIL2R deęişkenlerine ait histogram grafikleri Őekil 4.5 ve Őekil 4.6’da gsterilmiştir.



Őekil 4.5. OPN Deęişkenine Ait Histogram Grafięi



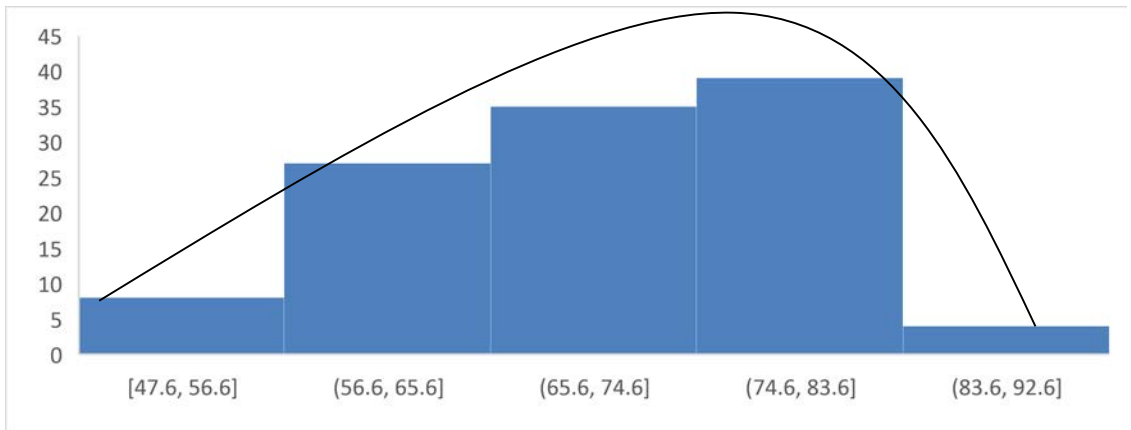
Őekil 4.6. SIL2R Deęişkenine Ait Histogram Grafięi

OPN ve SIL2R deęişkenleri hayatta kalanlarda daha yüksek deęerlerde gürülmüştür. Dięer deęişkenler ačíşından akcięer kanseri nedeniyle ölen ve hayatta kalan hastalar arasında istatistiksel olarak farklılık anlamli gürülmemiştir ($p > 0.05$).

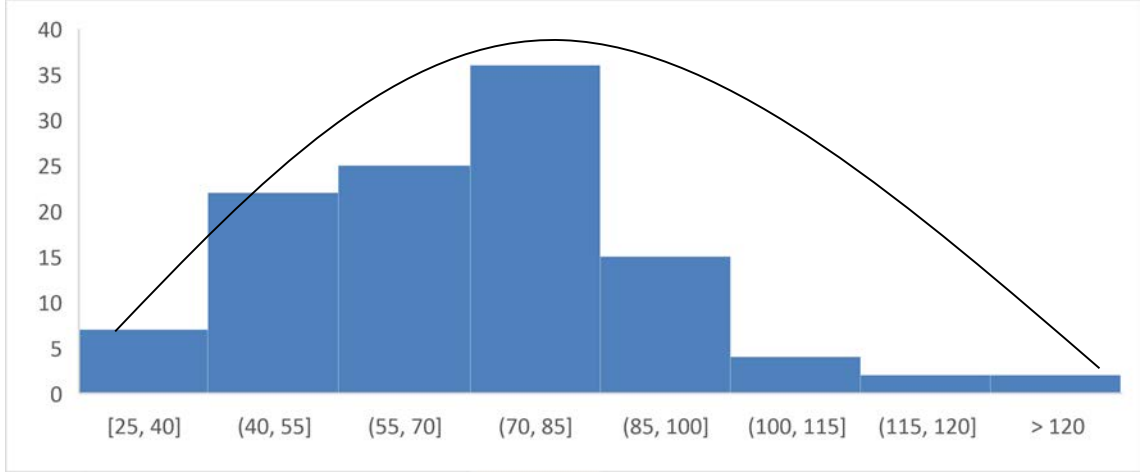
Tablo 4.3. Normal Dağılım Gösteren Sürekli Değişkenlere Göre Akciğer Kanseri Durumu

Değişkenler ($\bar{X} \pm SS$ / (Min.- Maks))	Öldü	Hayatta	t	P- değeri
OPN	102.8 \pm 51.6 /(16-489)	128.2 \pm 56.3 /(43-298)	3.042	0.003
CA9	291.7 \pm 197.3 /(48-985)	359.5 \pm 316.6 /(58-1546)	1.622	0.107
IL8	17.1 \pm 18.7 /(0.5-181)	18.6 \pm 14.2 /(2-65.1)	0.589	0.556
CEA	23.5 \pm 85.3 /(0.7-635)	18.2 \pm 40.95 /(1.5-279)	-0.517	0.606
SIL2R	5378.4 \pm 2220.8 /(1121- 10600)	6768.6 \pm 3353.2 /(2278- 20000)	3.091	0.002
TLR4	7.12 \pm 3.7 /(1.9-17.7)	7.26 \pm 4.63 /(1,4-29.9)	0.21	0.834
VEGF	101.5 \pm 82.3 /(17.8-504.8)	117.2 \pm 98.6 /(20.9-465.4)	1.01	0.272

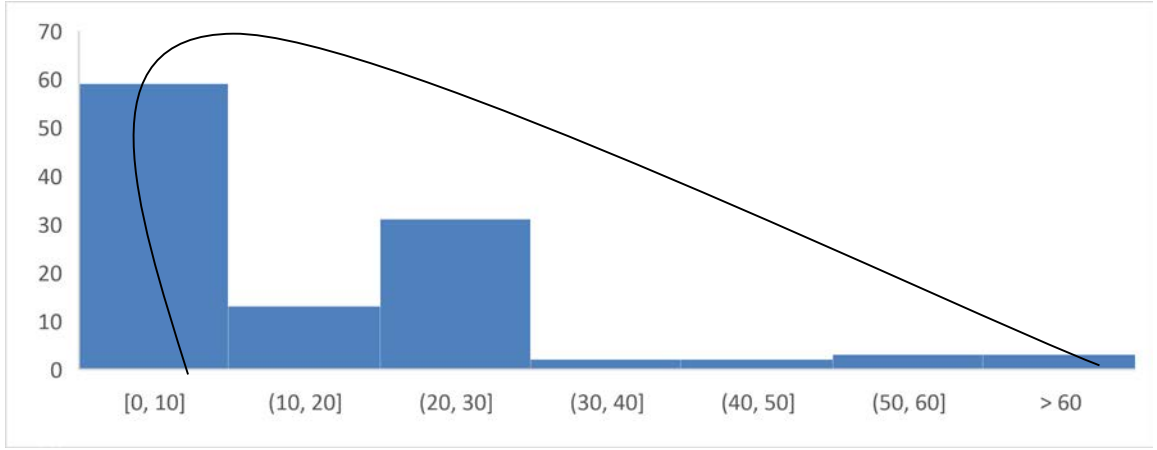
Tablo 4.4'deki karşılaştırma sonuçlarına göre akciğer kanseri nedeniyle ölen ve hayatta kalan hastaların yaş ve FEV1s% değişkenleri açısından farklılıkları istatistiksel olarak anlamlı değildir ($p > 0.05$). Yaş ve FEV1s% değişkenlerine ait histogram grafikleri şekil 4.7 ve 4.8'de gösterilmiştir. Total dose(2) değişkeni açısından akciğer kanseri nedeniyle ölen ve hayatta kalan hastalar arasındaki fark istatistiksel olarak anlamlıdır ($p < 0.05$). Total dose(2) değişkeni hayatta kalanlarda daha yüksek değerlerde görülmüştür. Total dose (2) değişkenine ait histogram grafiği şekil 4.9'da verilmiştir.



Şekil 4.7. Yaş Değişkenine Ait Histogram Grafiği



Şekil 4.8. Fev1s% Değişkenine Ait Histogram Grafiği



Şekil 4.9. Total Dose (2) Değişkenine Ait Histogram Grafiği

Tablo 4.4. Normal Dağılım Göstermeyen Sürekli Değişkenlere Göre Akciğer Kanseri Durumu (Man-Whitney U)

Değişkenler ($\bar{X} \pm SS$ / (Min.-Maks))	Öldü	Hayatta	χ^2	p-değeri
Age	68.3±51.6 /(16-489)	68.4±56.3 /(43-298)	3416.0	0.893
Fev1s%	73.1±197.3 /(48-985)	74.7±316.6 /(58-1546)	3424.5	0.914
Total Dose(2)	11.1±18.7 /(0,5-181)	17.5±14.2 /(2-65.1)	2694.5	0.009

4.2. Bireysel Sınıflayıcıların Performans Metrikleri Değerlendirmesi

KHDAK hastalarında ölüm durumunun tahminine yönelik 4 temel sınıflandırıcı belirlenmiştir. Bu sınıflandırıcılar, SMO, K-NN, Random Forest ve XGBoost şeklindedir. Tablo 4.5’de ilk olarak Random Forest sınıflandırıcısının akciğer kanseri hastalarında

ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 88.9, duyarlılık değeri % 81.6, özgüllük değeri % 94.4, kesinlik değeri % 85.6 ve AUC değeri % 98.9 olarak bulunmuştur. Random forest modeliyle ilgili hiperparametre değerleri, bir grid search için optimizasyon algoritmasına yerleştirilmiş, minimum kazanç için 0,1, minimum yaprak boyutu için 2, sp için minimum boyut için 4 ve ön budama sayısı için 3'dür.

Tablo 4.5. Random Forest Sınıflandırıcısının Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	88.9
Duyarlılık	81.6
Özgüllük	94.4
Kesinlik	85.6
AUC	98.9

Tablo 4.6'da SMO sınıflandırıcısının KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 91.5, duyarlılık değeri % 89.5, özgüllük değeri % 93.4, kesinlik değeri % 91.9 ve AUC değeri % 96.9 olarak bulunmuştur. SMO sınıflayıcısına ait hiperparametre değerleri bir grid search için optimizasyon algoritmasına yerleştirilmiş, kernel tipi için radial, L pos için 1, L neg için 1, epsilon için 0 'dır.

Tablo 4.6. SMO Sınıflandırıcısının Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	91.5
Duyarlılık	89.5
Özgüllük	93.4
Kesinlik	91.9
AUC	96.9

Tablo 4.7'de K-NN sınıflandırıcısının KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 59.8, duyarlılık değeri % 54.8, özgüllük değeri % 63.8, kesinlik değeri % 54.7 ve AUC değeri % 84.6 olarak bulunmuştur. K-NN modelinin hiperparametreleri, k için 1, ölçü türü için karma ölçüler, bileşik ölçüler için karma öklid, grid search optimizasyon algoritmasıdır.

Tablo 4.7. K-NN Sınıflandırıcısının Performans Metriği

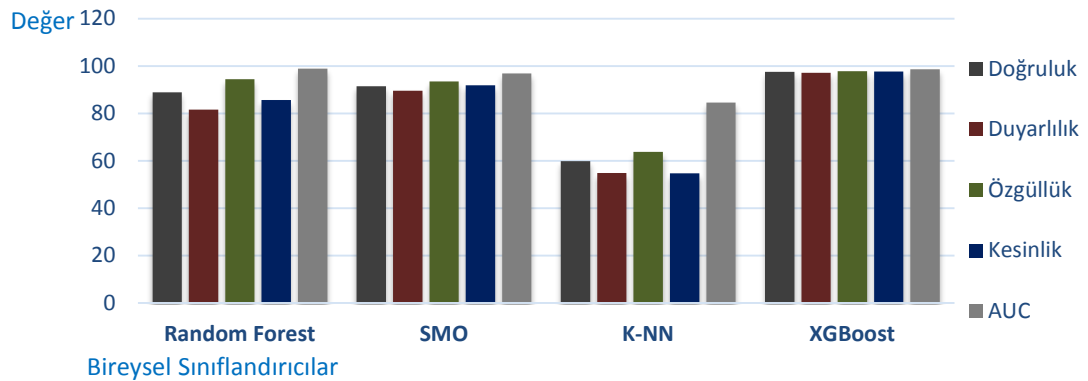
Performans Metriği	Değer (%)
Doğruluk	59.8
Duyarlılık	54.8
Özgüllük	63.8
Kesinlik	54.7
AUC	84.6

Tablo 4.8’de XGBoost sınıflandırıcısının KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 97.5, duyarlılık değeri % 97.1, özgüllük değeri % 97.8, kesinlik değeri % 97.6 ve AUC değeri % 98.6 olarak bulunmuştur. XGBoost modeli için hiperparametre grid search en uygun şekilde sokma algoritması tarafından ayarlanan maksimum derinlik için 5, bölme sayısı için 20 ve öğrenme hızı hiperparametreleri için 0.1 olan hiperparametrelere sahiptir.

Tablo 4.8. XGBoost Sınıflandırıcısının Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	97.5
Duyarlılık	97.1
Özgüllük	97.8
Kesinlik	97.6
AUC	98.6

Şekil 4.10’a göre bireysel sınıflandırıcıların performans metriklerinin değerlendirilmesinde, doğruluk, duyarlılık, özgüllük ve kesinlik değerlerinde XGBoost sınıflandırıcısı en yüksek değere sahipken ROC eğrisi altında kalan alan (AUC) değerinde random forest sınıflandırıcısı en yüksek değeri sağlamıştır. Tüm performans metriklerinde en düşük değer ise K-NN sınıflandırıcısında elde edilmiştir.

**Şekil 4.10.** Bireysel Sınıflandırıcıların Performans Metrikleri

4.3. Bagging Topluluk Öğrenme Yöntemi ile Sınıflayıcıların Performans Metriklerinin Değerlendirilmesi

KHDAK ölüm durumunun tahminine yönelik 4 temel sınıflandırıcı bagging topluluk öğrenme yönteminde ayrı ayrı modele dahil edilerek performans metrikleri hesaplanmıştır. Tablo 4.9’da ilk olarak Random Forest sınıflandırıcısının bagging topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 98.2, duyarlılık değeri % 97.1, özgüllük değeri % 98.9, kesinlik değeri % 98.9 ve AUC değeri % 99.4 olarak bulunmuştur.

Tablo 4.9. Random Forest Sınıflandırıcısının Bagging Topluluk Öğrenme Yöntemine Göre Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	98.2
Duyarlılık	97.1
Özgüllük	98.9
Kesinlik	98.9
AUC	99.4

Tablo 4.10’da SMO sınıflandırıcısının bagging topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 94.0, duyarlılık değeri % 90.9, özgüllük değeri % 96.7, kesinlik değeri % 95.9 ve AUC değeri % 98.0 olarak bulunmuştur.

Tablo 4.10. SMO Sınıflandırıcısının Bagging Topluluk Öğrenme Yöntemine Göre Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	94.0
Duyarlılık	90.9
Özgüllük	96.7
Kesinlik	95.9
AUC	98.0

Tablo 4.11’de K-NN sınıflandırıcısının bagging topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 59.8, duyarlılık değeri % 54.8, özgüllük değeri % 63.8, kesinlik değeri % 54.7 ve AUC değeri % 71.6 olarak bulunmuştur.

Tablo 4.11. K-NN Sınıflandırıcısının Bagging Topluluk Öğrenme Yöntemine Göre Performans Metriği

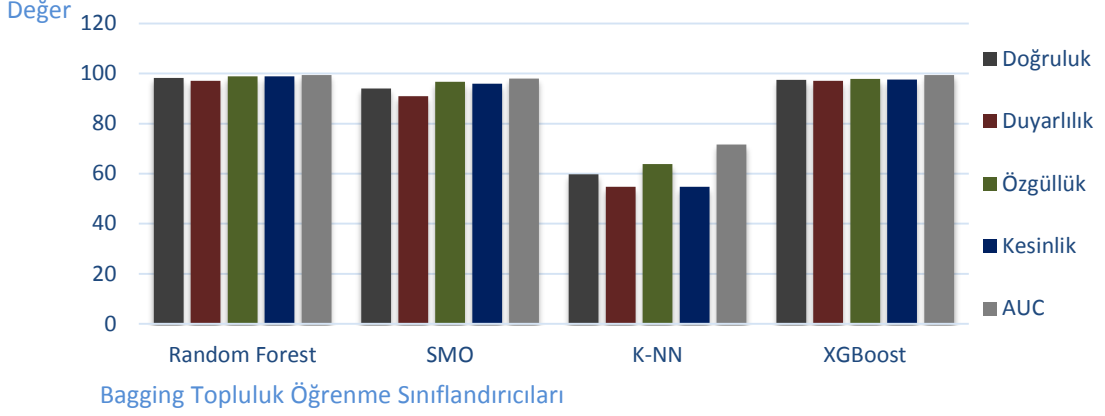
Performans Metriği	Değer (%)
Doğruluk	59.8
Duyarlılık	54.8
Özgüllük	63.8
Kesinlik	54.7
AUC	71.6

Tablo 4.12’de XGBoost sınıflandırıcısının bagging topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 97.5, duyarlılık değeri % 97.1, özgüllük değeri % 97.8, kesinlik değeri % 97.6 ve AUC değeri % 99.4 olarak bulunmuştur.

Tablo 4.12. XGBoost Sınıflandırıcısının Bagging Topluluk Öğrenme Yöntemine Göre Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	97.5
Duyarlılık	97.1
Özgüllük	97.8
Kesinlik	97.6
AUC	99.4

Şekil 4.11’e göre bagging topluluk öğrenme yöntemi için sınıflandırıcıların performans metriklerinin değerlendirildiğinde, doğruluk, özgüllük ve kesinlik değerlerinde random forest sınıflandırıcısı en yüksek değere sahipken özgüllük ve ROC eğrisi altında kalan alan (AUC) değerinde random forest ve XGBoost sınıflandırıcısı benzer en yüksek değeri sağlamıştır. Tüm performans metriklerinde en düşük değer ise K-NN sınıflandırıcısında elde edilmiştir.



Şekil 4.11. Bagging Topluluk Öğrenme Yöntemlerinde Sınıflandırıcıların Performans Metrikleri

4.4. Boosting Topluluk Öğrenme Yöntemi ile Sınıflayıcıların Performans Metriklerinin Değerlendirilmesi

KHDAK hastalarında ölüm durumunun tahminine yönelik 4 temel sınıflandırıcı boosting topluluk öğrenme yönteminde ayrı ayrı modele dahil edilerek performans metrikleri hesaplanmıştır. Tablo 4.13’de ilk olarak random forest sınıflandırıcısının boosting topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 98.1, duyarlılık değeri % 97.1, özgüllük değeri % 98.9, kesinlik değeri % 98.6 ve AUC değeri % 99.8 olarak bulunmuştur.

Tablo 4.13. Random Forest Sınıflandırıcısının Boosting Topluluk Öğrenme Yöntemine Göre Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	98.1
Duyarlılık	97.1
Özgüllük	98.9
Kesinlik	98.6
AUC	99.8

Tablo 4.14’de SMO sınıflandırıcısının boosting topluluk öğrenme yöntemi ile KHDAK kanseri hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 92.2, duyarlılık değeri % 89.6, özgüllük değeri % 94.6, kesinlik değeri % 93.6 ve AUC değeri % 97.4 olarak bulunmuştur.

Tablo 4.14. SMO Sınıflandırıcısının Boosting Topluluk Öğrenme Yöntemine Göre Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	92.2
Duyarlılık	89.6
Özgüllük	94.6
Kesinlik	93.6
AUC	97.4

Tablo 4.15’de K-NN sınıflandırıcısının boosting topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 59.8, duyarlılık değeri % 54.8, özgüllük değeri % 63.8, kesinlik değeri % 54.7 ve AUC değeri % 84.6 olarak bulunmuştur.

Tablo 4.15. K-NN Sınıflandırıcısının Boosting Topluluk Öğrenme Yöntemine Göre Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	59.8
Duyarlılık	54.8
Özgüllük	63.8
Kesinlik	54.7
AUC	84.6

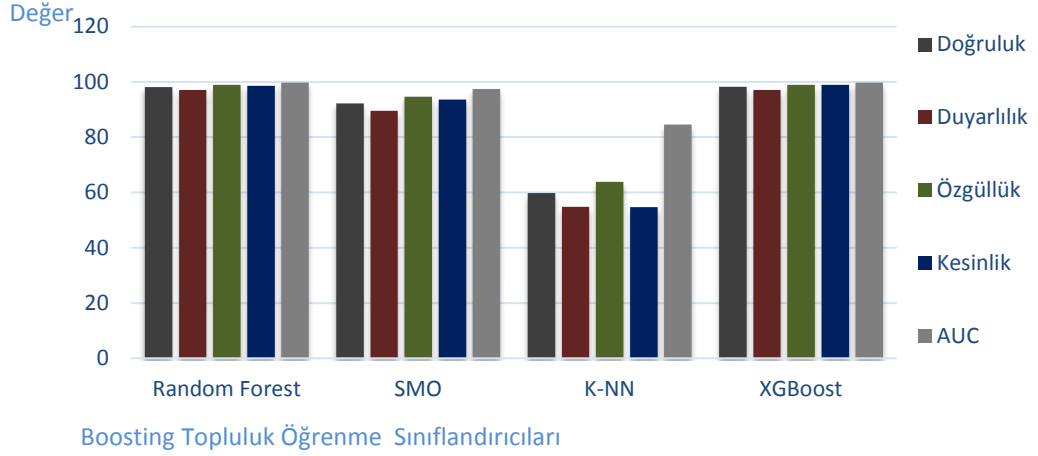
Tablo 4.16’da XGBoost sınıflandırıcısının boosting topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 98.2, duyarlılık değeri % 97.1, özgüllük değeri % 98.9, kesinlik değeri % 98.9 ve AUC değeri % 99.8 olarak bulunmuştur.

Tablo 4.16. XGBoost Sınıflandırıcısının Boosting Topluluk Öğrenme Yöntemine Göre Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	98.2
Duyarlılık	97.1
Özgüllük	98.9
Kesinlik	98.9
AUC	99.8

Şekil 4.12’e göre boosting topluluk öğrenme yöntemi için sınıflandırıcıların performans metriklerinin değerlendirildiğinde, doğruluk ve kesinlik değerlerinde XGBoost sınıflandırıcısı en yüksek değere sahipken duyarlılık, özgüllük ve ROC eğrisi

altında kalan alan (AUC) değerinde random forest ve XGBoost sınıflandırıcısı benzer en yüksek değeri sağlamıştır. Tüm performans metriklerinde en düşük değer ise K-NN sınıflandırıcısında elde edilmiştir.



Şekil 4.12. Boosting Topluluk Öğrenme Yöntemlerinde Sınıflandırıcıların Performans Metrikleri

4.5. Voting Topluluk Öğrenme Yöntemi ile Sınıflayıcıların Performans

Metriklerinin Değerlendirilmesi

Voting topluluk öğrenme yöntemine göre farklı sınıflandırıcılardan oluşturulan 7 modelin KHDAK hastalarında ölüm durumu sınıflandırma performansları ölçülmüştür. Model 1’de random forest ve SMO sınıflandırıcılarının birlikte voting topluluk öğrenme performansları değerlendirilmiştir. Model 2’de K-NN ve XGBoost sınıflandırıcılarının birlikte voting topluluk öğrenme performansları değerlendirilmiştir. Model 3’de random forest ve K-NN sınıflandırıcılarının birlikte voting topluluk öğrenme performansları değerlendirilmiştir. Model 4’de SMO ve XGBoost sınıflandırıcılarının birlikte voting topluluk öğrenme performansları değerlendirilmiştir. Model 5’de random forest, SMO ve K-NN sınıflandırıcılarının birlikte voting topluluk öğrenme performansları değerlendirilmiştir. Model 6’da XGBoost, SMO ve K-NN sınıflandırıcılarının birlikte voting topluluk öğrenme performansları değerlendirilmiştir. Model 7’de ise random forest, SMO, K-NN ve XGBoost sınıflandırıcılarının birlikte voting topluluk öğrenme performansları değerlendirilmiştir.

Tablo 4.17’de ilk olarak random forest ve SMO sınıflandırıcılarının model 1’de voting topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 94.5, duyarlılık değeri % 88.0, özgüllük değeri % 100, kesinlik değeri % 100 ve AUC değeri % 99.7 olarak bulunmuştur.

Tablo 4.17. Voting Topluluk Öğrenme Yöntemine Göre Model 1'in Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	94.5
Duyarlılık	88.0
Özgüllük	100.0
Kesinlik	100.0
AUC	99.7

Tablo 4.18'de K-NN ve XGBoost sınıflandırıcılarının model 2'de voting topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 78.4, duyarlılık değeri % 53.4, özgüllük değeri % 98.9, kesinlik değeri % 96.7 ve AUC değeri % 98.4 olarak bulunmuştur.

Tablo 4.18. Voting Topluluk Öğrenme Yöntemine Göre Model 2'nin Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	78.4
Duyarlılık	53.4
Özgüllük	98.9
Kesinlik	96.7
AUC	98.4

Tablo 4.19'da K-NN ve random forest sınıflandırıcılarının model 3'de voting topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 78.4, duyarlılık değeri % 53.4, özgüllük değeri % 98.9, kesinlik değeri % 97.5 ve AUC değeri % 98.6 olarak bulunmuştur.

Tablo 4.19. Voting Topluluk Öğrenme Yöntemine Göre Model 3'nin Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	78.4
Duyarlılık	53.4
Özgüllük	98.9
Kesinlik	97.5
AUC	98.6

Tablo 4.20'de SMO ve XGBoost sınıflandırıcılarının model 4'de voting topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 93.9, duyarlılık değeri % 88.0, özgüllük değeri % 98.9, kesinlik değeri % 98.8 ve AUC değeri % 99.7 olarak bulunmuştur.

Tablo 4.20. Voting Topluluk Öğrenme Yöntemine Göre Model 4'ün Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	93.9
Duyarlılık	88.0
Özgüllük	98.9
Kesinlik	98.8
AUC	99.7

Tablo 4.21'de random forest, SMO ve K-NN sınıflandırıcılarının model 5'de voting topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 95.1, duyarlılık değeri % 94.6, özgüllük değeri % 95.6, kesinlik değeri % 94.6 ve AUC değeri % 98.8 olarak bulunmuştur.

Tablo 4.21. Voting Topluluk Öğrenme Yöntemine Göre Model 5'in Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	95.1
Duyarlılık	94.6
Özgüllük	95.6
Kesinlik	94.6
AUC	98.8

Tablo 4.22'de SMO, K-NN ve XGBoost sınıflandırıcılarının model 6'da voting topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 95.7, duyarlılık değeri % 94.6, özgüllük değeri % 96.7, kesinlik değeri % 96.1 ve AUC değeri % 98.1 olarak bulunmuştur.

Tablo 4.22. Voting Topluluk Öğrenme Yöntemine Göre Model 6'nın Performans Metriği

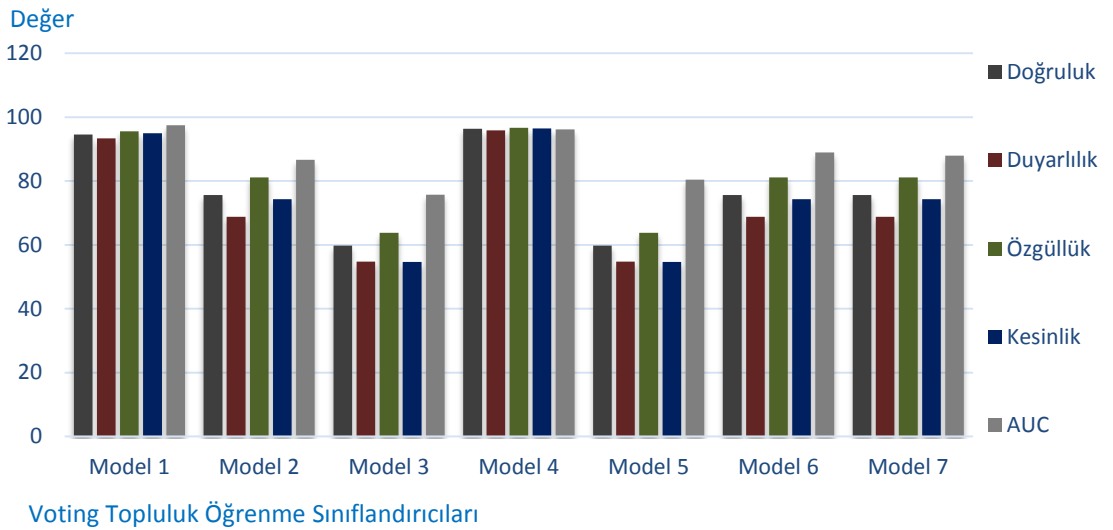
Performans Metriği	Değer (%)
Doğruluk	95.7
Duyarlılık	94.6
Özgüllük	96.7
Kesinlik	96.1
AUC	98.1

Tablo 4.23'de random forest, SMO, K-NN ve XGBoost sınıflandırıcılarının model 7'de voting topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 96.3, duyarlılık değeri % 93.2, özgüllük değeri % 98.9, kesinlik değeri % 98.8 ve AUC değeri % 99.0 olarak bulunmuştur.

Tablo 4.23. Voting Topluluk Öğrenme Yöntemine Göre Model 7'nin Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	96.3
Duyarlılık	93.2
Özgüllük	98.9
Kesinlik	98.8
AUC	99.0

Şekil 4.13'e göre voting topluluk öğrenme yöntemi için oluşturulan model sınıflandırıcıların performans metriklerinin değerlendirildiğinde, doğruluk değerinde en yüksek değere model 7 sahipken ve duyarlılık değerinde model 5 ve model 6 benzer en yüksek değere sahiptir. Kesinlik ve özgüllük değerlerinde en yüksek değere model 1 sahipken, ROC eğrisi altında kalan alan (AUC) değerinde ise model 1 ve model 4 en yüksek değerlere sahiptir.



Şekil 4.13. Voting Topluluk Öğrenme Yöntemlerinde Sınıflandırıcıların Performans Metrikleri

4.6. Stacking Topluluk Öğrenme Yöntemi ile Sınıflayıcıların Performans Metriklerinin Değerlendirilmesi

Stacking topluluk öğrenme yöntemine göre farklı sınıflandırıcılardan oluşturulan 7 modelin KHDAK hastalarında ölüm durumu sınıflandırma performansları ölçülmüştür. Model 1'de random forest ve SMO sınıflandırıcılarının birlikte stacking topluluk öğrenme performansları değerlendirilmiştir. Model 2'de K-NN ve XGBoost sınıflandırıcılarının birlikte stacking topluluk öğrenme performansları

değerlendirilmiştir. Model 3’de random forest ve K-NN sınıflandırıcılarının birlikte stacking topluluk öğrenme performansları değerlendirilmiştir. Model 4’de SMO ve XGBoost sınıflandırıcılarının birlikte stacking topluluk öğrenme performansları değerlendirilmiştir. Model 5’de random forest, SMO ve K-NN sınıflandırıcılarının birlikte stacking topluluk öğrenme performansları değerlendirilmiştir. Model 6’da XGBoost, SMO ve K-NN sınıflandırıcılarının birlikte stacking topluluk öğrenme performansları değerlendirilmiştir. Model 7’de ise random forest, SMO, K-NN ve XGBoost sınıflandırıcılarının birlikte stacking topluluk öğrenme performansları değerlendirilmiştir. Oluşturulan model sınıflandırıcı performanslarını geliştirmek üzere meta sınıflayıcı olarak “logistik” sınıflayıcı kullanılmıştır.

Tablo 4.24’de ilk olarak random forest ve SMO sınıflandırıcılarının model 1’de stacking topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 94.6, duyarlılık değeri % 93.4, özgüllük değeri % 95.6, kesinlik değeri % 95.0 ve AUC değeri % 97.5 olarak bulunmuştur.

Tablo 4.24. Stacking Topluluk Öğrenme Yöntemine Göre Model 1’in Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	94.6
Duyarlılık	93.4
Özgüllük	95.6
Kesinlik	95.0
AUC	97.5

Tablo 4.25’de K-NN ve XGBoost sınıflandırıcılarının model 2’de stacking topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 75.6, duyarlılık değeri % 68.8, özgüllük değeri % 81.1, kesinlik değeri % 74.3 ve AUC değeri % 86.7 olarak bulunmuştur.

Tablo 4.25. Stacking Topluluk Öğrenme Yöntemine Göre Model 2’nin Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	75.6
Duyarlılık	68.8
Özgüllük	81.1
Kesinlik	74.3
AUC	86.7

Tablo 4.26’da K-NN ve random forest sınıflandırıcılarının model 3’de stacking topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 59.8, duyarlılık değeri % 54.8, özgüllük değeri % 63.8, kesinlik değeri % 54.7 ve AUC değeri % 75.7 olarak bulunmuştur.

Tablo 4.26. Stacking Topluluk Öğrenme Yöntemine Göre Model 3’nin Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	59.8
Duyarlılık	54.8
Özgüllük	63.8
Kesinlik	54.7
AUC	75.7

Tablo 4.27’de SMO ve XGBoost sınıflandırıcılarının model 4’de stacking topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 96.4, duyarlılık değeri % 95.9, özgüllük değeri % 96.7, kesinlik değeri % 96.5 ve AUC değeri % 96.2 olarak bulunmuştur.

Tablo 4.27. Stacking Topluluk Öğrenme Yöntemine Göre Model 4’ün Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	96.4
Duyarlılık	95.9
Özgüllük	96.7
Kesinlik	96.5
AUC	96.2

Tablo 4.28’de random forest, SMO ve K-NN sınıflandırıcılarının model 5’de stacking topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 59.8, duyarlılık değeri % 54.8, özgüllük değeri % 63.8, kesinlik değeri % 54.7 ve AUC değeri % 80.4 olarak bulunmuştur.

Tablo 4.28. Stacking Topluluk Öğrenme Yöntemine Göre Model 5'in Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	59.8
Duyarlılık	54.8
Özgüllük	63.8
Kesinlik	54.7
AUC	80.4

Tablo 4.29'da SMO, K-NN ve XGBoost sınıflandırıcılarının model 6'da stacking topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 75.6, duyarlılık değeri % 68.8, özgüllük değeri % 81.1, kesinlik değeri % 74.3 ve AUC değeri % 89.0 olarak bulunmuştur.

Tablo 4.29. Stacking Topluluk Öğrenme Yöntemine Göre Model 6'nın Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	75.6
Duyarlılık	68.8
Özgüllük	81.1
Kesinlik	74.3
AUC	89.0

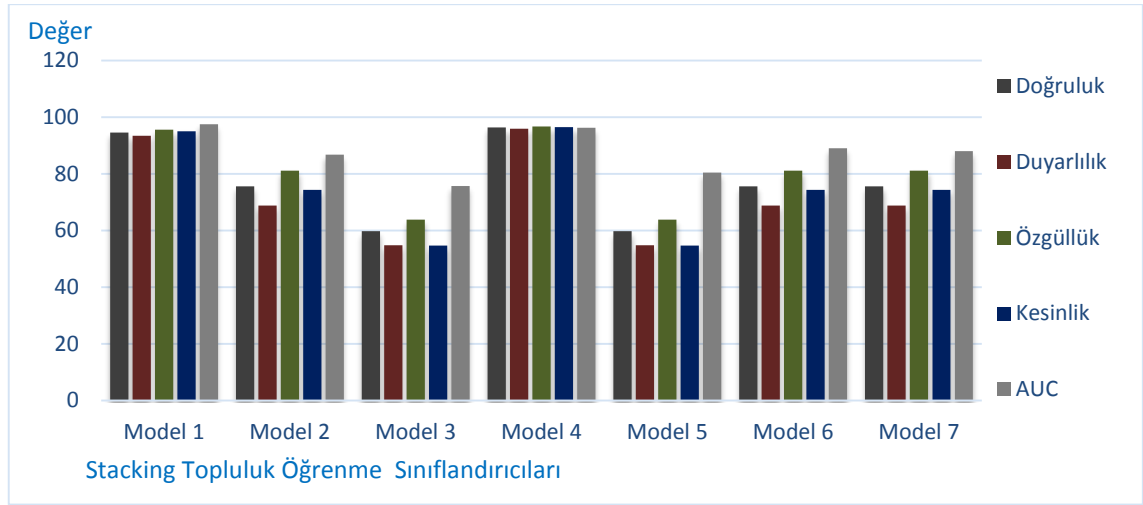
Tablo 4.30'de random forest, SMO, K-NN ve XGBoost sınıflandırıcılarının model 7'de stacking topluluk öğrenme yöntemi ile KHDAK hastalarında ölüm durumu performansı değerlendirilmiştir. Doğruluk değeri % 75.6, duyarlılık değeri % 68.8, özgüllük değeri % 81.1, kesinlik değeri % 74.3 ve AUC değeri % 88.0 olarak bulunmuştur.

Tablo 4.30. Stacking Topluluk Öğrenme Yöntemine Göre Model 7'nin Performans Metriği

Performans Metriği	Değer (%)
Doğruluk	75.6
Duyarlılık	68.8
Özgüllük	81.1
Kesinlik	74.3
AUC	88.0

Şekil 4.14'e göre stacking topluluk öğrenme yöntemi için oluşturulan model sınıflandırıcıların performans metriklerinin değerlendirildiğinde, doğruluk, duyarlılık,

özgüllük ve kesinlik değerlerinde en yüksek değere model 4 sahipken, ROC eğrisi altında kalan alan (AUC) değerinde ise model 1 en yüksek değerlere sahiptir.



Şekil 4.14. Stacking Topluluk Öğrenme Yöntemlerinde Sınıflandırıcıların Performans Metrikleri

4.7. Genel Değerlendirme

KHDAK hastalarında ölüm durumunun sınıflandırılmasında farklı sınıflayıcılar ve farklı topluluk öğrenme yöntemleri uygulanarak en iyi performansı sağlayan model belirlenmesi sağlanmıştır. Genel değerlendirme Tablo 4.31’de belirtilmiştir. Tabloda random forest, SMO (Ardışık Minimal Optimizasyon), K-NN (K-En Yakın Komşu) ve XGBoost (Aşırı Gradyan Artırma) sınıflayıcılarının temel sınıflayıcı performansları, bagging topluluk öğrenme yöntemindeki performansları ve boosting topluluk öğrenme yöntemindeki performansları değerlendirilmiştir. Ayrıca voting ve stacking topluluk öğrenme yöntemlerinde uygulanan Model 1 (random forest+SMO), Model 2 (XGBoost+K-NN), Model 3 (random forest+K-NN), Model 4 (XGBoost+SMO), Model 5 (SMO+K-NN+random forest), Model 6 (SMO+K-NN+XGBoost) ve Model 7 (SMO+K-NN+random forest+XGBoost)’nin farklı metriklerdeki performansları ifade edilmiştir.

Tablo 4.31. Genel Değerlendirme

		Doğruluk	Duyarlılık	Özgüllük	Kesinlik	AUC
Temel Sınıflayıcılar	Random					
	Forest	88.9	81.6	94.4	85.6	98.9
	SMO	91.5	89.5	93.4	91.9	96.9
	K-NN	59.8	54.8	63.8	54.7	84.6
	XGBoost	97.5	97.1	97.8	97.6	98.6
Bagging Topluluk Öğrenme	Random					
	Forest	98.2	97.1	98.9	98.9	99.4
	SMO	94.0	90.9	96.7	95.9	98.0
	K-NN	59.8	54.8	63.8	54.7	71.6
	XGBoost	97.5	97.1	97.8	97.6	99.4
Boosting Topluluk Öğrenme	Random					
	Forest	98.1	97.1	98.9	98.6	99.8
	SMO	92.2	89.6	94.6	93.6	97.4
	K-NN	59.8	54.8	63.8	54.7	84.6
	XGBoost	98.2	97.1	98.9	98.9	99.8
Voting Topluluk Öğrenme	Model 1	94.5	88.0	100.0	100.0	99.7
	Model 2	78.4	53.4	98.9	96.7	98.4
	Model 3	78.4	53.4	98.9	97.5	98.6
	Model 4	93.9	88.0	98.9	98.8	99.7
	Model 5	95.1	94.6	95.6	94.6	98.8
	Model 6	95.7	94.6	96.7	96.1	98.1
	Model 7	96.3	93.2	98.9	98.8	99.0
Stacking Topluluk Öğrenme	Model 1	94.6	93.4	95.6	95.0	97.5
	Model 2	75.6	68.8	81.1	74.3	86.7
	Model 3	59.8	54.8	63.8	54.7	75.7
	Model 4	96.4	95.9	96.7	96.5	96.2
	Model 5	59.8	54.8	63.8	54.7	80.4
	Model 6	75.6	68.8	81.1	74.3	89.0
	Model 7	75.6	68.8	81.1	74.3	88.0

Genel değerlendirmeye göre doğruluk, duyarlılık, kesinlik, özgüllük ve ROC eğrisi altında kalan alan metriklerinde en yüksek performansı boosting topluluk öğrenme yöntemi sağlamıştır. XGBoost ile en yüksek sınıflandırma performansını sağlayan boosting topluluk öğrenme yöntemi 0.982 doğruluk değeri, 0.971 duyarlılık değeri, 0.989 kesinlik değeri, 0.989 özgüllük değeri ve 0.998 ROC eğrisi altında kalan alan değerlerini elde etmiştir. XGBoost sınıflayıcısının sözde kodu (pseudo-code) şekil 4.15’de belirtilmiştir. Tüm sınıflayıcı ve topluluk yöntemlerinin sınıflandırma performansları Şekil 4.16’da ifade edilmiştir.

Input: $X \rightarrow$ Özellik matrisi ($N \times p$); Input: $B \rightarrow$ iterasyon sayısı

Input: $\eta \rightarrow$ öğrenme oranı; Input: $\Upsilon \rightarrow$ bir bölünme için gerekli minimum kayıp azaltma

Input: $\lambda \rightarrow$ ağırlık düzenleme parametresi; Input: $d \rightarrow$ maksimum ağaç derinliği

Input: $s \rightarrow$ sütunların alt örnek oranı

Başla;

1: $f_0(x_i) = \arg_{\Upsilon} \min \sum_{i=1}^N l(y_i, \Upsilon)$

2: for $b=1$ to B

- p ortak değişkenlerinin rastgele $\lfloor sp \rfloor$ alt örnekleme
- Maks. Derinlik d 'nin alt örnekleme ortak değişkenlerinden bir ağaç f_b 'si büyütmek için bir bölünme bulma algoritması kullanın
- $\sum_{i=1}^N l(y_i, \hat{y}_i^{(b-1)} + f_b(x_i)) + \sqrt{T_b} + \frac{1}{2} \lambda \|w_b\|^2$ (Notasyonel olarak, $\hat{y}_i^{(b)}$, b 'ninci yinelemede y_i 'nin tahminini temsil eder. T_b , b . ağaçtaki yaprak sayısını temsil eder ve w_b , b . ağaçındaki yaprakların ağırlıklarını temsil eder).
 - hesapla $l(y_i, \hat{y}_i^{(b)}) = y_i \hat{y}_i^{(b)} - \log(1 + \exp(\hat{y}_i^{(b)}))$ (Türevlenebilir dışbükey kayıp işlevi fonksiyonu).

3: $\hat{y}_i = \sum_{b=1}^B \eta f_b(x_i)$ (log-odds tahminleri)

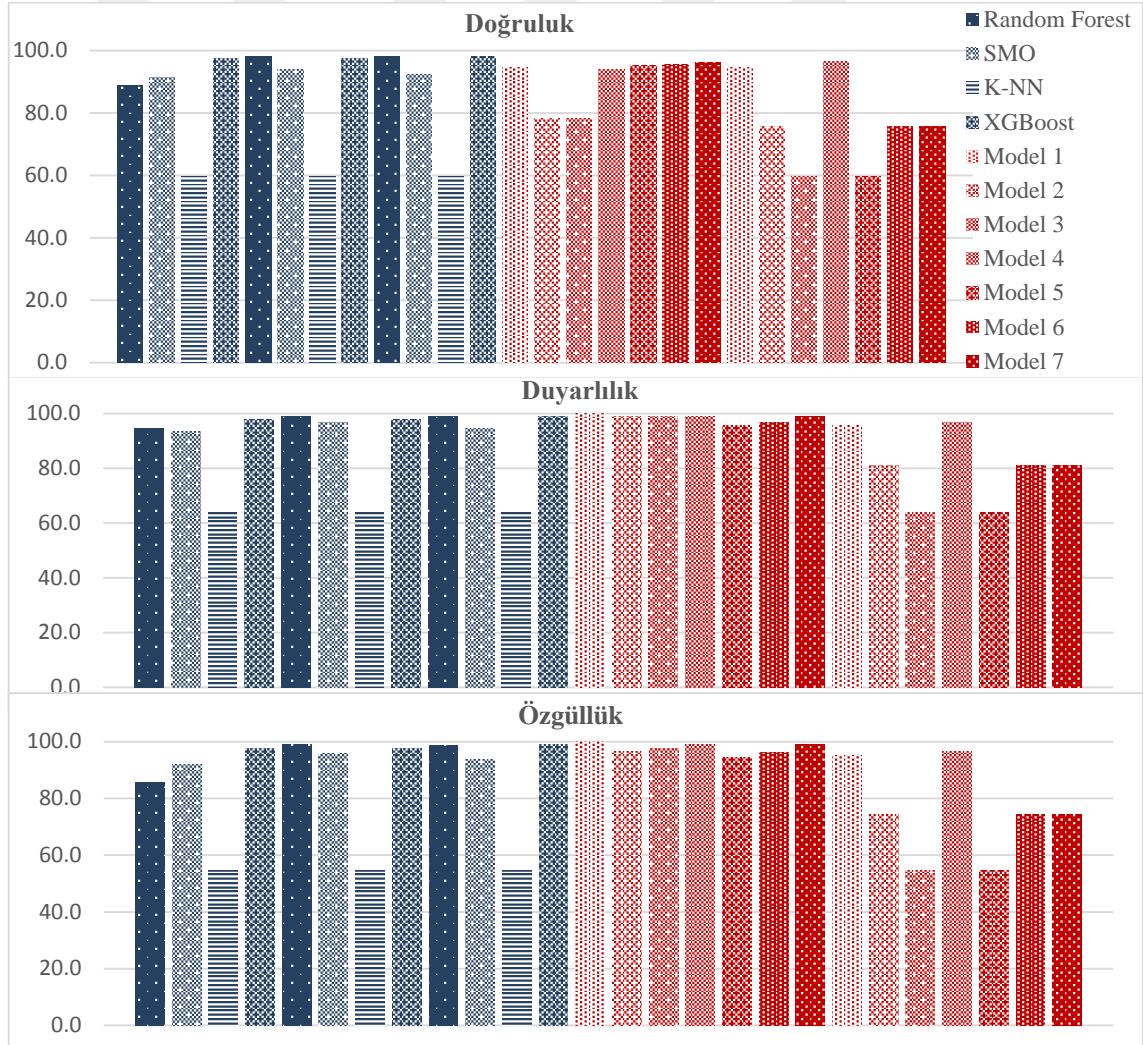
Son.

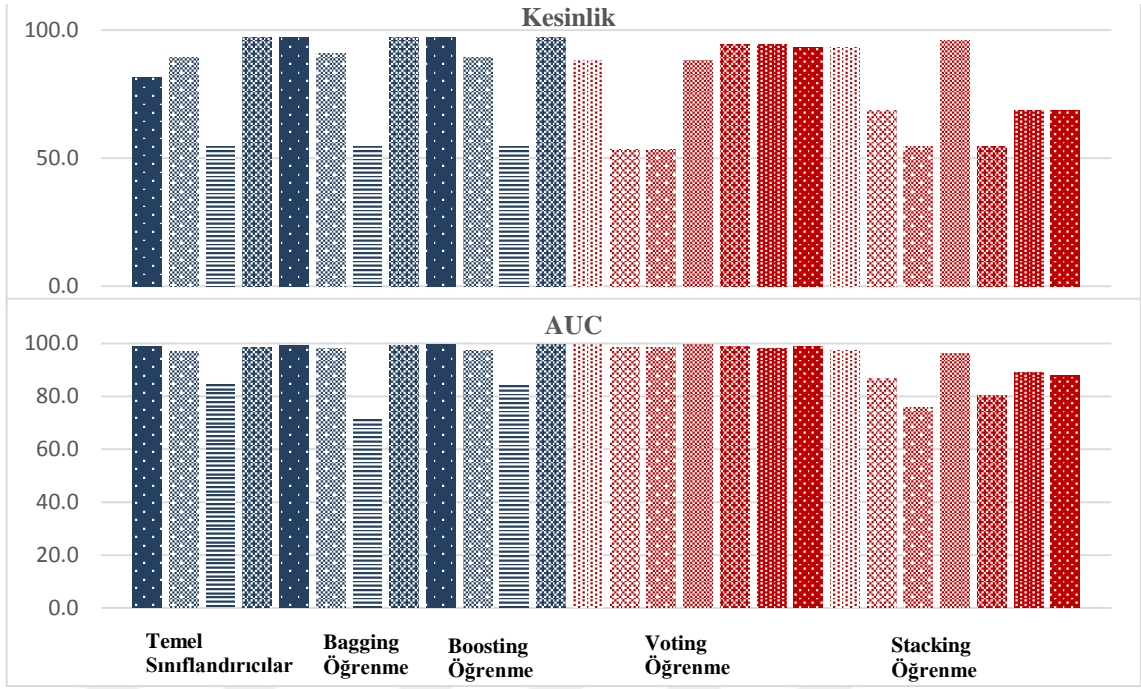
Şekil 4.15. XGBoost Sınıflayıcısına Ait Pseudo Kodu

Tablo 4.32'de KHDAK ölüm durumunun sınıflandırılmasında, bireysel sınıflandırıcılarda en iyi sonucu veren XGBoost sınıflandırıcısı ile topluluk öğrenme yönteminde en iyi sonucu veren boosting algoritmasındaki XGBoost sınıflandırıcısına ait değişkenlerin önem düzeyleri gösterilmiştir. XGBoost temel sınıflandırıcısında, GTV (% 29.5), CEA (% V10.4), WHOPS (% 10.0), yaş (% 8.5) ve OPN (%7.2) değişkenleri en yüksek önemi sağlarken cinsiyet (% 0.0), lymphnodes (% 2.5) ve SIL2R (% 2.9) en düşük önemi sağlayan değişkenler olmuştur. Boosting topluluk öğrenme algoritmasındaki XGBoost sınıflandırıcısında, GTV (% 11.2), CEA (%6.2), WHOPS (% 5.2), Total Dose 2 (% 5.2) değişkenleri en yüksek önemi sağlarken cinsiyet (% 0.0), FEV1s% (% 2.0) ve lymphnodes (% 2.1) en düşük önemi sağlayan değişkenler olmuştur.

Tablo 4.32. En iyi Modellere Ait Değişken Önemlilikleri

Değişken	XGBoost (Temel)		Boosting (XGBoost) (Topluluk)	
	Göreceli Önemlilik	Yüzde (%)	Göreceli Önemlilik	Yüzde (%)
GTV	29.5	29.4	11.2	11.2
CEA	10.4	10.4	6.2	6.2
WHOPS	10.0	9.9	5.2	5.3
Yaş	8.5	8.5	2.3	2.3
OPN	7.2	7.2	4.7	4.7
Total Dose(2)	6.6	6.6	5.2	5.1
FEV1s%	6.4	6.4	2.0	2
CA9	4.9	5.0	3.2	3.2
VEGF	4.6	4.5	4.1	4.1
IL8	3.6	3.6	2.6	2.6
TLR4	3.3	3.3	3.3	3.3
SIL2R	2.9	2.8	4.9	5.0
Lymphnodes	2.5	2.5	2.1	2.1
Cinsiyet	0.0	0.1	0.0	0.0





Şekil 4.16. Genel Performans Metrikleri Değerlendirmesi

5. TARTIŞMA

Akciğer kanseri en yaygın kanser türlerinden biridir ve akciğer hücrelerinin anormalleşmesi ile oluşarak kontrol dışı büyümeye başlar (60). Kanser hücreleri sayıları arttıkça bir tümör oluştururlar ve vücudun diğer bölümlerine yayılırlar (61). KHDAK, tüm akciğer kanseri vakalarının % 80'inden fazlasını oluşturur (62). Üç ana türü bulunan KHDAK'da, adenokarsinom genellikle akciğerin dışında görülen ve mukus salgılayan hücrelerin erken formlarında başlayan bir türüdür. Diğer bir türü olan skuamoz hücreli karsinom, akciğer kanserlerinin %25 ila 30'unu oluşturur ve akciğerlerde hava yollarının içini örten yassı hücreler olan skuamoz hücrelerin erken formlarında gelişir. Büyük hücreli karsinom, akciğer kanserlerinin %10 ila 15'ini oluşturur ve akciğerlerin herhangi bir bölümünde oluşabilir. Hızlı büyüme ve yayılma eğilimindedir ve bu tedavisini güç kılar. Büyük hücreli karsinomların bir alt türü olan büyük hücreli nöroendokrin karsinom, KHDAK'a benzer, hızlı büyüyen bir kanserdir (61-63). Erken evre KHDAK'li hastalar için temel tedavi cerrahidir. İleri vakalarda veya ameliyat olamayan hastalarda tedavi yöntemleri ışın tedavisi veya kemoterapidir (64). Akciğer kanseri hastalarının % 80'inden fazlası uygulama sırasında ameliyat edilemez. Tüm vakalar değerlendirildiğinde 5 yıllık sağkalım % 10'dan azdır. Kanser hastalarını mümkün olduğunca erken teşhis etmek yaygın olarak kabul edilen bir yaklaşımdır. Daha erken evrede teşhis edilen kanser hastalarında hayatta kalma ve iyileşme oranları daha yüksektir. Ancak Kanser hastalarında tanı ve tedavide gecikmeler yaygındır. Bazı çalışmalar teşhis ve tedavi gecikmelerinin tümör şiddetini ve prognozunu etkileyebileceğini gösterirken, bazı çalışma tümör evresi ile prognoz ve bu gecikmeler arasında önemli bir ilişki olmadığını göstermiştir (62, 65).

Son yıllarda, özellikle kanser hastalıklarının erken tanı ve tedavisinde makine öğrenimi ve yapay zeka yöntemlerinin kullanılması geniş bir uygulama alanı haline gelmiştir. Makine öğrenimi ve yapay zeka yöntemleri, özellikle kanser hastalıklarını tahmin etmede sıklıkla kullanılarak yüksek sınıflama performansı sağlamayı başarmıştır (66-69). Günümüzde birçok sağlık alanında kullanılan bu yöntemler, veri tabanlarında gizli kalmış örüntülerin otomatik olarak elde edilmesi ve veri analiz sürecinin

verimliliğine önem vermektedir (9).

Topluluk öğrenme yöntemi, bir optimal tahmin modelini üretmek için birkaç temel modeli birleştiren algoritmadır ve sınıflandırıcılar aracılığıyla eğitim veri kümesini eğitmek yerine birden çok sınıflandırıcıyı modelleyerek verileri eğitir (12, 15). Doğruluğu artırıp varyansı azaltmak için geliştirilen topluluk öğrenme algoritmaları, özellik çıkarımı, güven tahmini, hata düzeltme, dengesiz veriler ve durağan olmayan dağılımlarda sapmayı öğrenme gibi çeşitli makine öğrenme sistemlerinde başarılı sonuçlar vermektedir. Topluluk öğrenme, çoklu modelleri birleştirme yöntemidir ve tipik olarak modelin kestirim performansını geliştirerek verilecek karara olan güveni artırmaktadır (16). Tahmindeki hata değeri azaldığından, temel sınıflandırıcıya göre daha yüksek performans elde edilebilir.

Bu çalışmanın amacı, yapay zekâya dayalı topluluk öğrenme yöntemlerinin tıpta kullanılabilirliğini araştırılmak ve geliştirmektir. KHDAK ölüm durumunu sınıflandırmak için çeşitli klinik değişkenleri içeren veri seti kullanılarak, KHDAK değişkeni; yanıt/çıktı/hedef ve veri setinde yer alan ölçümler ile diğer faktörler açıklayıcı/tahminleyici/bağımsız değişkenler olmak üzere, topluluk öğrenme yöntemlerinin KHDAK ölüm durumu sınıflama performansı incelenecek ve en iyi model belirlenecektir. KHDAK hastalarında ölüm durumunun tahminine yönelik 4 temel sınıflandırıcı belirlenmiştir. Bu sınıflandırıcılar, SMO, K-NN, Random Forest ve XGBoost şeklindedir. Bu sınıflandırıcıların bireysel performanslarına bakılırken aynı zamanda topluluk öğrenme modeline dahil edilerek topluluk öğrenme performansları da incelenmiştir. Temel sınıflandırıcılarda en yüksek performansı XGBoost sınıflandırıcısı sağlamıştır. XGBoost sınıflandırıcısının doğruluk değeri % 97.5 iken ROC eğrisi altında kalan alan % 98.6 olarak elde edilmiştir. Topluluk öğrenme yönteminde ise en yüksek performansı boosting topluluk öğrenme yönteminde XGBoost sınıflandırıcısı sağlamıştır. Doğruluk değeri % 98.2 iken ROC eğrisi altında kalan alan % 99.8 olarak elde edilmiştir. Akciğer kanseri hastalıklarında sınıflama performansına yönelik kullanılan benzer makine ve yapay zeka öğrenme algoritmaları araştırılarak, bulgular değerlendirilmiştir. Aliferis ve arkadaşları dizi karşılaştırmalı genomik hibridizasyon kullanan akciğer kanseri hastalarının sınıflandırması için makine öğrenimi modelleri konulu çalışmalarında karar ağaçları, K-NN, lineer SVM, polinomal SVM, kernel SVM ve yapay sinir ağlarından yararlanmıştır. Performans metriklerine göre en yüksek başarıyı K-NN

sınıflayıcısı sağlamıştır (70). Nishio ve arkadaşları XGBoost ve Bayes optimizasyonu kullanarak akciğer nodülünün bilgisayar destekli teşhisi konulu çalışmalarında SVM ve XGBoost sınıflandırıcılarından faydalanmışlardır. Performans metriklerine göre en yüksek başarıyı XGBoost sınıflandırıcısı sağlamıştır (71). Wu ve arkadaşları erken T Evre KHDAK olan hastalarda lenf nodu metastazının makine öğrenimi algoritmaları ile preoperatif tahmini konulu çalışmalarında adabosting, yapay sinir ağları, karar ağaçları, lojistik regresyon, random forest, multinominal naive bayes (MNV) ve XGBoost sınıflayıcılarından yararlanmıştır. Performans metriklerine göre en yüksek başarıyı XGBoost sınıflandırıcısı sağlamıştır (72). Bireysel sınıflayıcı performanslarına ilişkin literatürdeki benzer çalışmalarda, XGBoost algoritması genelde en yüksek performans metriklerini sunarken, çalışmamızda da benzer şekilde bireysel sınıflandırıcılarından en yüksek başarıyı XGBoost sınıflayıcısı sağlamıştır. Boosting topluluk öğrenme algoritmasında XGBoost sınıflandırıcısı en yüksek performans metriklerine sahip olurken, bireysel sınıflandırıcıların performanslarından daha üstün bir başarı göstermiştir.

Topluluk öğrenme yöntemlerinde başarı iki temel kurgu üzerine dayanır. Bunlardan birincisi, bireysel sınıflandırıcıların performans başarısı ne kadar yüksek olursa topluluk öğrenme yöntemlerinin de başarısı o kadar yüksek olur. İkincisi ise bireysel sınıflandırıcıların sonuçlarının birbirinden farklı, ayrık olması (diversity) yine topluluk öğrenme başarısını artırmaktadır. Topluluk öğrenme yöntemlerinde genelde doğru sınıflandırıcılar ve birleştirme teknikleri kullanıldığında çok başarılı sonuçlar vermektedir. Safiyari ve arkadaşları topluluk öğrenme yöntemleri ile akciğer kanserinden sağkalım'ı tahmin etme konulu çalışmalarında, RIPPER (Repeated Incremental Pruning to Produce Error Reduction), DS (Decision Stamp), CART, C4.5, SMO, lojistik regresyon, NB ve random forest temel sınıflandırıcılarını bagging, dagging, adaboost, multiboosting ve random subspace gibi 5 adet topluluk öğrenme algoritmalarında kullanmışlardır. Temel sınıflandırıcıların performansını artırmada ve sağkalım tahmin sonuçlarında en iyi performans metriğini adaboost topluluk öğrenme algoritması sağlarken, en düşük performans metriğini random subspace algoritması sağlamıştır (35). Faisal ve arkadaşları, akciğer kanserinin erken evre tahmini için makine öğrenimi sınıflandırıcıları ve topluluklarının değerlendirilmesi konulu çalışmalarında, gradyan destekli ağaç (GBT), SVM, C4.5, karar ağaçları, çok katmanlı perceptron (MLP) ve NB dahil olmak üzere bir dizi sınıflandırıcı performansları ve ayrıca random forest (bagging) ve Majority Voting gibi tanınmış topluluklarla da karşılaştırılmıştır. Performans

değerlendirmelerine göre, GBT'nin diğer tüm bireysel ve topluluk sınıflandırıcılarından daha iyi performans gösterdiği görülmüştür (36). Topluluk öğrenme sınıflayıcı performanslarına ilişkin literatürdeki benzer çalışmalarda, boosting algoritması genelde en yüksek performans metriklerini sunarken, çalışmamızda da benzer şekilde topluluk sınıflandırıcılarından en yüksek başarıyı boosting algoritmasında XGBoost sınıflayıcısı sağlamıştır.

KHDAK vakalarında topluluk öğrenme yöntemine göre sınıflandırma ve tahmin araştırmalarının yeterli düzeyde olmadığı görülmektedir. Özellikle çok yaygın olarak görülen bu kanser türünde topluluk öğrenme yöntemlerinin kullanılması, veri setlerinin çok sayıda değişkeni içermesi ve çok sayıda örnek ile eğitim veri setindeki örüntünün öğrenebilmesi sayesinde yüksek sınıflandırma performansı sağlayacaktır. Çalışmanın temel amacı olan KHDAK hastalarındaki ölüm durumunun topluluk yöntemlerine göre sınıflandırılması başarılı bir şekilde tahmin edilmiştir. Topluluk yöntemlerinde kullanılan temel sınıflandırıcılardan SMO ve K-NN sınıflandırıcılarının XGBoost ve random forest sınıflandırıcılarına göre daha düşük sınıflandırma performansı sağlarken, birden fazla model oluşturularak karşılaştırma imkanı veren voting ve stacking gibi yüksek performans sağlayan algoritmaların beklenen sonucu vermediği görülmektedir. Bunun birinci sebebi, oluşturulan modellerin hepsinde düşük performans sağlayan temel sınıflandırıcıların yer alması ve ikinci sebep olarak bagging ve boosting topluluk öğrenme algoritmalarının random forest ve XGBoost temel sınıflandırıcılarının performansını çok başarılı bir şekilde artırmasından kaynaklanmaktadır. Bundan sonra yapılacak benzer çalışmalarda, sadece en yüksek bireysel başarıyı gösteren sınıflayıcılardan oluşturulacak modelin de voting ve stacking topluluk öğrenme algoritmalarına dahil edilmesi neticesinde sınıflandırma performansı karşılaştırılabilecektir. Bu çalışmada, KHDAK hastalarında ölüm durumunun sınıflandırılmasında güçlü bir topluluk öğrenme yöntemi oluşturulmuştur. Bu yöntem, boosting topluluk öğrenme algoritmasında XGBoost sınıflandırıcısı tarafından sağlanmıştır. Özellikle birden fazla topluluk öğrenme algoritmalarının aynı çalışmada kullanıldığı ve performans metriklerinin karşılaştırıldığı araştırmalar literatürde çok fazla bulunmamaktadır. Yapılan çalışmanın stacking gibi yeni topluluk öğrenme algoritmaları kullanılarak performans ölçümlerinin değerlendirilmesi ve aynı zamanda bireysel sınıflandırıcıların seçimi ve sınıflandırma performanslarının düşük olmasının topluluk öğrenme yöntemlerinin performansını olumsuz yönde etkilediğinin görülmesi açısından da önemli bir çalışmadır.

6. SONUÇ VE ÖNERİLER

Çalışmada, topluluk öğrenme yöntemleri kullanılarak KDHAH hastaları ölüm durumu tahmin edilmiştir. Modelde bagging topluluk öğrenme yöntemi, boosting topluluk öğrenme yöntemi, voting topluluk öğrenme yöntemi ve stacking topluluk öğrenme yönteminin yanısıra random forest, SMO, K-NN ve XGBoost gibi temel sınıflayıcıların performansları da değerlendirilmiştir. Topluluk yöntemi ve sınıflayıcıların performansları doğruluk, duyarlılık, kesinlik, özgüllük ve ROC eğrisi altında kalan alan metriklerinde değerlendirilmiştir.

En yüksek sınıflandırma performansını boosting topluluk öğrenme XGBoost sınıflandırıcısı sağlamıştır. Boosting topluluk öğrenme yöntemi KDHAH hastaları ölüm durumunu başarılı bir şekilde tahmin etmiştir. Boosting topluluk öğrenme yöntemi ile sınıflandırmada % 98.2 doğruluk, % 99.8 oranında da ROC eğrisi altında kalan alan değerleri elde edilmiştir. Topluluk öğrenme yöntemleri KDHAH hastaları ölüm durumunu tahmin etmede temel sınıflayıcılara göre daha yüksek performans elde etmektedir. Boosting topluluk öğrenme yöntemleri, bagging, stacking ve voting topluluk öğrenme yöntemlerinden daha yüksek sınıflandırma performansı sağlamıştır.

Temel sınıflandırıcılara göre topluluk öğrenme sınıflandırıcıları KDHAH ölüm durumunu sınıflandırmada daha iyi sonuçlar vermiştir. Boosting topluluk öğrenme yöntemi en iyi sınıflandırıcı performansı sağlarken, bagging topluluk öğrenme yöntemi de başarılı sonuçlar vermiştir. Özellikle bagging topluluk öğrenme algoritmasında random forest sınıflandırıcısı da yüksek performans metriği sağlamıştır.

Prevelansı yüksek kanser hastalıklarında, sınıflandırma problemleri için topluluk öğrenme yöntemlerinin kullanılması başarılı sonuçların elde edilmesi açısından önerilmektedir. Özellikle bireysel sınıflandırıcıların doğru bir şekilde seçilmesi ile stacking gibi yeni topluluk öğrenme yöntemlerinin daha başarılı olacağı önerilmektedir.

KAYNAKLAR

1. Günbatar H, Sertoğullarından B, Özbay B, Sünnetçioğlu A, Ekin S. Akciğer kanserli olguların değerlendirilmesi; 3 yıllık analiz. *Van Tıp Dergisi* 2012, 19(1): 13-2.
2. Damadoğlu E. Küçük Hücreli Dişi Akciğer Kanserinde Tani Ve Tedavi Gecikmeleri Ve Bu Gecikmelerin Patolojik Tümör Evresine Etkisi. Göğüs Hastalıkları ve Göğüs Cerrahisi Anabilim Dalı. Uzmanlık Tezi. İstanbul: SBÜ, 2007.
3. Türk Toraks Derneği, *Akciğer ve Plevra Maligniteleri Çalışma Grubu*, 13.Baskı, 2005: 13.
4. Türk Toraks Derneği, *Türkiye’de Temel Akciğer Sağlığı Sorunları ve Çözüm Önerileri*, 2010: 89-93.
5. Carvalho S, Troost EG, Bons J, Menheere P, Lambin P, Oberije C. Prognostic value of blood-biomarkers related to hypoxia, inflammation, immune response and tumour load in non-small cell lung cancer–A survival model with external validation. *Radiotherapy and Oncology* 2016, 119(3): 487-94.
6. Schreiber G, Douglas CM. Performance characteristics of different modalities for diagnosis of suspected lung cancer: summary of published evidence. *Chest* 2003, 123.1: 115-28.
7. Yoh WM. TNM classification for lung cancer. *Ann Thorac Cardiovasc Surg* 2003, 9(6):343-50.
8. Silahtaroglu G. *Veri madenciliği*, 3.Baskı. İstanbul, Papatya Yayınları 2008: 9-173
9. Karacan H, Yeşilbudak M. Kullanıcı Merkezli İnteraktif Veri Madenciliği: Bir Literatür Taraması. *Bilişim Teknolojileri Dergisi* 2010, 3(1):17-21.
10. Han J, Pei J, Kamber M. Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, 3th ed. Burnaby, Elsevier 2011, 5(4): 83-124.
11. Hsiao C-Y, Lo H-Y, Yin T-C, Lin S-D. Optimizing specificity under perfect sensitivity for medical data classification. In 2014 *International Conference on Data Science and Advanced Analytics (DSAA) IEEE* 2014,163-9.
12. Li K, Liu Z, Han Y. Study of selective ensemble learning methods based on support vector machine. *Physics Procedia* 2012, 33: 1518-25.

13. Bahl N. Optimizing Performance Measures in Classification Using Ensemble Learning Methods. Arizona State University, Master Thesis. Arizona, ASU, 2017.
14. Zhou Z-H. *Ensemble methods: foundations and algorithms*, 5th ed. Boca Raton, CRC press 2012:184
15. Zhang C, Ma Y. *Ensemble machine learning: methods and applications*, 5th ed. Boston, Springer 2012.
16. Polikar R. Ensemble learning. *Ensemble machine learning*, 10th ed. Boston, Springer 2012:1-34.
17. Lantz B. *Machine learning with R*, 1th ed. Birmingham, Packt publishing 2013.
18. Géron A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, 2th ed. Sebastopol , O'Reilly Media 2019:694
19. Alpaydin E. *Introduction to machine learning*, 4th ed. Massachusetts, MIT press 2020:665
20. Sümbüloğlu K, Akdağ B. *İyi klinik uygulamalar*, 1.Baskı. Denizli, Pamukkale Üniversitesi Yayınları 2010.
21. Rokach L. Pattern classification using ensemble methods, *World Scientific* 2010: vol.75.22-90.
22. Dasarathy BV, Sheela BV. A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE* 1979, 67(5): 708-13.
23. Wittner BS, Denker JS. Strategies for teaching layered networks classification tasks. *Neural information processing systems* 1988: 850-9.
24. Schapire RE. The strength of weak learnability. *Machine learning*, 1990, 5(2): 197-27.
25. Hansen LK, Salamon P. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence* 1990, 12(10): 993-1001. 1
26. Xu L, Krzyzak A, Suen CY. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE transactions on systems, man, and cybernetics* 1992, 22(3): 418-35.
27. Luo S-T, Cheng BW. Diagnosing breast masses in digital mammography using feature selection and ensemble methods. *Journal of medical systems* 2012, 36(2): 569-77.

28. Nagi S, Bhattacharyya DK. Classification of microarray cancer data using ensemble approach. *Network Modeling Analysis in Health Informatics and Bioinformatics* 2013, 2(3): 159-73.
29. Onan A. On the performance of ensemble learning for automated diagnosis of breast cancer. *Artificial Intelligence Perspectives and Applications* 2015: 119-29.
30. Tabibu S, Vinod P, Jawahar CV. Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning. *Scientific reports* 2019, 9(1): 1-9.
31. Wang Y, Wang D, Geng N, Wang Y, Yin Y, Jin Y. Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Applied Soft Computing* 2019, 77: 188-204
32. Cai Z, Xu D, Zhang Q, Zhang J, Ngai S-M, Shao J. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular BioSystems* 2015, 11(3): 791-800.
33. Bergquist SL, Brooks GA, Keating NL, Landrum MB, Rose S. Classifying lung cancer severity with ensemble machine learning in health care claims data. *Machine Learning for Healthcare Conference. PMLR* 2017:25-38.
34. Tan AC, Gilbert D. *Ensemble machine learning on gene expression data for cancer classification*, 1th ed. London, Brunel University 2003.
35. Safiyari A, Javidan R. Predicting lung cancer survivability using ensemble learning methods. *Intelligent Systems Conference (IntelliSys)* 2017: IEEE.
36. Faisal MI, Bashir S, Khan ZS, Khan FH. An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. *3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)* 2018: IEEE.
37. Breiman L. *Random forests*, 1th ed. Boston, Springer 2001.
38. Akman M, Genç Y, Ankarali H. Random forests yöntemi ve sağlık alanında bir uygulama. *Türkiye Klinikleri Biyoistatistik* 2011, 3(1): 36-48.
39. Wolpert DH. Stacked generalization. *Neural networks* 1992, 5(2): 241-59.
40. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 1999, 10(3): 61-74.
41. Nitze I, Schulthess U, Asche H. Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum

- likelihood for supervised crop type classification. *Proc. of the 4th GEOBIA 2012*, 35.
42. Cortes C, Vapnik V. Support-vector networks. *Machine learning* 1995, 20(3): 273-97.
 43. Ayhan S, Erdoğan Ş. Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi* 2014, 9(1): 175-201.
 44. Arslan A, Şen B. Detection of non-coding RNA's with optimized support vector machines. *23rd Signal Processing and Communications Applications Conference (SIU) IEEE*. 2015:1668-71.
 45. Yıldız O. Melanoma detection from dermoscopy images with deep learning methods: A comprehensive study. *Journal of the Faculty of Engineering and Architecture of Gazi University* 2019, 34(4): 2241-60.
 46. Alpar R. *Spor, Sağlık ve Eğitim Bilimlerinden Örneklerle Uygulamalı İstatistik Ve Geçerlilik-Güvenirlilik*, 4. Baskı. Ankara, Detay Yayıncılık 2016, 513-57.
 47. Santra A, Christy CJ. Genetic algorithm and confusion matrix for document clustering. *International Journal of Computer Science Issues (IJCSI)* 2012, 9(1):322.
 48. Kotu V, Deshpande B. *Predictive analytics and data mining: concepts and practice with rapidminer*, 1th ed. Waltham, Morgan Kaufmann 2014.
 49. Yu CH. Resampling methods: concepts, applications, and justification. *Practical Assessment, Research, and Evaluation* 2002, 8.1: 19.
 50. Beasley WH, Rodgers JL. Re-sampling methods. *The SAGE Handbook of Quantitative Methods in Psychology* 2009: 362-86.
 51. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research* 2010, 11: 2079-07.
 52. Yadav S, Shukla S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. *IEEE 6th International conference on advanced computing (IACC)*; IEEE.2016: 78-3.
 53. Temel GO, Erdoğan S, Ankaralı H. Sınıflama Modelinin Performansını Değerlendirmede Yeniden Örnekleme Yöntemlerinin Kullanımı. *Bilişim Teknolojileri Dergisi* 2012, 5(3): 1-8.

54. Arslan AK, Yaşar Ş, Çolak C, Yoloğlu S. WSSPAS: An Interactive Web Application for Sample Size and Power Analysis with R Using Shiny. *Türkiye Klinikleri Biyoistatistik* 2018, 10(3): 224-46.
55. Fayyad U. Knowledge discovery in databases: An overview. *Relational Data Mining*; Springer; 2001. p. 28-47.
56. YAŞAR Ş, ARSLAN A, Colak C, Yoloğlu SJMBSJoHS. A Developed Interactive Web Application for Statistical Analysis: Statistical Analysis Software. 6(2): 227-39.
57. Campbell M. *RStudio Projects*. Learn RStudio IDE: Boston, Springer; 2019: 39-48.
58. Hofmann M, Klinkenberg R. *RapidMiner: Data mining use cases and business analytics applications*, 3th ed. Boca Raton, CRC Press 2016.
59. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Statistics surveys* 2010, 4: 40-79.
60. Jemal A, Thomas A, Murray T, Thun M. Cancer statistics. *Ca-A Cancer Journal for Clinicians* 2002, 52(1): 23-47.
61. Cokkinides V, Albano J, Samuels A, Ward M, Thum JM. *American cancer society: Cancer facts and figures*. Atlanta: American Cancer Society 2005.
62. Capewell S. Patients presenting with lung cancer in South East Scotland. *Thorax* 1987, 42: 853-7.
63. Janssen-Heijnen M, Gatta G, Forman D, Capocaccia R, Coebergh J. Variation in survival of patients with lung cancer in Europe, *European Journal of Cancer* 1998, 34(14): 2191-6.
64. Robinson E, Mohilever J, Zidan J, Sapir D. Delay in diagnosis of cancer. Possible effects on the stage of disease and survival. *Cancer* 1984, 54(7): 1454-60.
65. Christensen E, Harvald T, Jendresen M, Aggestrup S, Petterson G. The impact of delayed diagnosis of lung cancer on the stage at the time of operation. *European journal of cardio-thoracic surgery* 1997, 12(6): 880-4.
66. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* 2015, 13: 8-17.
67. Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science* 2016, 83: 1064-9.

68. Liu H, Zhang R, Luan F, Yao X, Liu M, Hu Z, Fan BT. Diagnosing breast cancer based on support vector machines. *Journal of Chemical Information and Computer Sciences* 2003, 43(3): 900-7.
69. Maglogiannis I, Zafiropoulos E, Anagnostopoulos I. An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. *Applied intelligence*, 2009, 30(1): 24-36.
70. Aliferis CF, Hardin D, Massion PP. Machine learning models for lung cancer classification using array comparative genomic hybridization. *Proceedings of the AMIA Symposium 2002*: 7.
71. Nishio M, Nishizawa M, Sugiyama O, Kojima R, Yakami M, Kuroda T, Togashi K. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PloS one* 2018, 13(4): e0195875.
72. Wu Y, Liu J, Han C, Liu X, Chong Y, Wang Z, Gong L, Zhang J, Gao X, Guo C. Preoperative prediction of lymph node metastasis in patients with early-T-stage non-small cell lung cancer by machine learning algorithms. *Frontiers in Oncolog*, 2020, 10: 743.

EKLER

Ek-1. Özgeçmiş

A. Kişisel Bilgiler

Adı Soyadı	Mehmet KIVRAK
Doğum Tarihi	1983
Doğum Yeri	İskenderun
Yabancı Dil	İngilizce
Görev Yeri	TUIK Malatya Bölge Müdürlüğü
İletişim Bilgisi	_____

B. Eğitim Bilgileri

Derece	Bölüm	Üniversite	Yıl
Lisans	İstatistik	Dokuz Eylül Üniversitesi	2001-2006
Y.Lisans	Biyoistatistik ve Tıp Bilişimi	100.Yıl Üniversitesi	2008-2010
Doktora	Biyoistatistik ve Tıp Bilişimi	İnönü Üniversitesi	2016-2020

C. İş Tecrübesine İlişkin Bilgiler

Yıllar	Çalıştığı Kurum	Pozisyon
2007-2008	Özel Vazgal Rehabilitasyon Merkezi	Kurucu Müdür
2008-2009	Tarım Kredi Kooperatifi	Muhasebe Uzm. Yrd.
2009-Halen Devam Ediyor	TUIK Malatya Bölge Müdürlüğü	İstatistik Grup Müdürü

Ek-2. Etik Kurul Alınma Gerekçesi

SAĞLIK BİLİMLERİ ENSTİTİSÜ MÜDÜRLÜĞÜNE

13 Nisan 2013 tarih ve 28617 sayılı Resmi Gazetede yayımlanan “Klinik Araştırmalar Hakkında Yönetmenlik” ile birinci bölüm 2.maddesinin 1.fıkrası (Bu yönetmenlik, biyoyararlanım ve biyoşdeğerlik çalışmaları dahil, ruhsat veya izin alınmış olsa dahi insanlar üzerinde yapılacak olan ilaç, tıbbi ve biyolojik ürünler ile bitkisel ürünlerin klinik araştırmaları, klinik araştırma yerlerini ve bu araştırmaları gerçekleştirecek gerçek veya tüzel kişileri kapsar.) gereğince, doktora öğrencisi Mehmet KIVRAK’ın tezinin klinik bir çalışma olmaması ve deneysel bir amaçla kullanılmaması, kullanılan verinin açık kaynak web ortamından sağlanması dolayısıyla Etik Kurul kararı alınmamıştır.