

Esra TOKUR SONUVAR

YÜKSEK LİSANS TEZİ

2024-ANTALYA

T.C.
AKDENİZ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK VE TIBBİ BİLİŞİM ANABİLİM DALI

ULUSAL ELEKTRONİK SAĞLIK KAYITLARI
ÜZERİNDE MAKİNE ÖĞRENMESİ İLE KANSER RİSK
FAKTÖRLERİNİN İNCELENMESİ

Esra TOKUR SONUVAR

YÜKSEK LİSANS TEZİ

2024-ANTALYA

T.C.
AKDENİZ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK VE TIBBİ BİLİŞİM ANABİLİM DALI

ULUSAL ELEKTRONİK SAĞLIK KAYITLARI
ÜZERİNDE MAKİNE ÖĞRENMESİ İLE KANSER RİSK
FAKTÖRLERİNİN İNCELENMESİ

Esra TOKUR SONUVAR

YÜKSEK LİSANS TEZİ

DANIŞMAN
Doç.Dr. Kemal Hakan GÜLKESEN

“Kaynakça gösterilerek tezimden yararlanılabilir”

2024-ANTALYA

Saęlık Bilimleri Enstitüsü M¼d¼rl¼ę¼ne;

Bu alıřma j¼rimiz tarafından Biyoistatistik Ve Tıbbi Biliřim Anabilim Dalı Tıp Biliřimi Programında Y¼ksek Lisans tezi olarak kabul edilmiřtir. 08/07/2024

İmza

Tez Danıřmanı : Do. Dr. Kemal Hakan G¼LKESEN
Akdeniz ¼niversitesi

¼ye : Do. Dr. Neře ZAYİM
Akdeniz ¼niversitesi

¼ye : Dr. Öğr. Üyesi Yılmaz Kemal YÜCE
Alanya Alaaddin Keykubat ¼niversitesi

Bu tez, Enstit¼ Y¼netim Kurulunca belirlenen yukarıdaki j¼ri ¼yeleri tarafından uygun g¼r¼lm¼ř ve Enstit¼ Y¼netim Kurulu'nun/...../..... tarih ve/..... sayılı kararıyla kabul edilmiřtir.

Enstit¼ M¼d¼r¼

ETİK BEYAN

Bu tez çalışmasının kendi çalışmam olduğunu, tezin planlanmasından yazımına kadar bütün safhalarda etik dışı davranışımın olmadığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara kaynak gösterdiğimi ve bu kaynakları da kaynaklar listesine aldığımı beyan ederim.

Öğrenci

Esra TOKUR SONUVAR

İmza

Tez Danışmanı

Doç. Dr. Kemal Hakan GÜLKESEN

İmza

TEŞEKKÜR

Bu tez çalışmasını gerçekleştirme sürecinde destek ve yardımlarını esirgemeyen birçok kişiye teşekkür etmek isterim.

Öncelikle, danışmanım Doç. Dr. Kemal Hakan Gülkesen'e, yol gösterici rehberliği ve değerli katkıları için minnettarım. Çalışmalarımı sürekli olarak motive eden ve beni yönlendiren tavsiyeleri, bu tezin tamamlanmasında büyük rol oynamıştır.

Bölümümüzün kıymetli öğretim üyeleri Prof. Dr. Ahmet Yardımcı, Prof. Dr. Uğur Bilge, Doç. Dr. Öğr. Üyesi Neşe Zayım, Arş. Gör. Nevruz İlhanlı ve Arş. Gör. Hazal Taş'a teşekkür ederim. Ayrıca, Biyoistatistik ve Tıp Bilişimi Anabilim Dalı çalışanlarına da katkıları ve destekleri için minnettarım.

Sağlık Bakanlığı'ndan verilerin temin edilmesinde yardımcı olan Doç. Dr. Mustafa Mahir Ülgü ve İstatistik Uzmanı Abdullah Akünal'a özellikle teşekkür ederim.

Ayrıca Akdeniz Üniversitesi Hastanesi çalışma arkadaşlarıma, çalışmalarım sırasında gösterdikleri destek ve iş birliği için teşekkür ederim.

Kıymetli eşim Adem Sonuvar, sevgili ablam Doç. Dr. Makbule Tokur Kesgin, değerli ailem ve arkadaşlarım destekleri ve sabırlarıyla bu süreci benim için daha kolay ve anlamlı kıldılar. Onlara, yanımda oldukları ve bana inandıkları için minnettarım.

Son olarak, bu tez çalışmasında emeği geçen herkese içten teşekkürlerimi sunarım.

ÖZET

Amaç: Kanser, dünya genelinde önde gelen ölüm nedenlerinden biridir. Ulusal elektronik sağlık kayıtları (EHR), potansiyel kanser risk faktörlerini belirlemek için analiz edilebilecek zengin bir bilgi kaynağı sunar. Bu çalışmanın amacı, çeşitli makine öğrenme modelleri kullanarak kanser tahmininin doğruluğunu ve performansını değerlendirmektir. Ayrıca, çeşitli değişkenlerin kanser riski üzerindeki etkilerini incelemek ve bu bilgilerle klinik uygulamalara ve hasta yönetimine yönelik öneriler geliştirmek hedeflenmiştir.

Yöntem: Çalışmada kullanılan veri seti, 1 Ocak 2018 ile 31 Aralık 2022 tarihleri arasında kanser tanısı almış deney grubu, kanser tanısı almamış kontrol grubu 18 yaş üstü Türkiye Cumhuriyeti vatandaşları ve Türkiye’de oturma iznine sahip kişilerin e-Nabız sisteminden alınan verilerden oluşmaktadır. Veri setine standart ölçekleme işlemi uygulanmış ve makine öğrenme modelleri (lojistik regresyon [LR], SVM, XGBoost, karar ağaçları, rastgele ormanlar, yapay sinir ağları) uygulanmıştır. Modellerin performansı doğruluk, duyarlılık, kesinlik, F1 skoru, MCC, AUC-ROC ve precision-recall curve (PRC) metrikleri ile değerlendirilmiştir. Ayrıca, veri setindeki değişkenlerin kanser riski üzerindeki etkileri odds oranları, p-değerleri ve etki büyüklükleri kullanılarak analiz edilmiştir.

Bulgular: Yapılan analizlerde, XGBoost modeli 0,846 (0,841-0,850, %95 GA) AUC değeri ile en yüksek performansı göstermiştir. LR analizi ileri yaş, İstanbul bölgesinde ikamet, yüksek hemoglobin, düşük ALT ve bazı komorbiditelerin kanser riski ile ilişkili olduğunu göstermiştir.

Sonuç: Çalışmamız, kanser tahmininde XGBoost modelinin en yüksek performansı sergilediğini ortaya koymuştur. Bazı değişkenlerin kanser riski üzerindeki etkileri, klinik uygulamalar ve hasta yönetimi için kritik bilgiler sunmaktadır. Bu bulgular, makine öğrenimi modellerinin sağlık uygulamalarında kullanımını desteklemekte ve kanserin daha iyi anlaşılmasına katkıda bulunmaktadır.

Anahtar Kelimeler: büyük veri, makine öğrenmesi, yapay zeka, kanserin erken tanısı

ABSTRACT

Objective: Cancer is one of the leading causes of death worldwide. National electronic health records (EHR) provide a rich source of data that can be analyzed to identify potential cancer risk factors. The aim of this study is to evaluate the accuracy and performance of cancer prediction using various machine learning (ML) models. Additionally, the study aims to examine the effects of various variables on cancer risk and to develop recommendations for clinical applications and patient management based on this information.

Method: The data set used in the study consists of the data citizens of the Republic of Turkey and persons with a residence permit in Turkey, over the age of 18. The data was obtained from the e-Nabız system. The experimental group was diagnosed with cancer between 1 January 2018 and 31 December 2022, and the control group had no cancer diagnosis. The data underwent standard scaling, several ML models (logistic regression [LR], SVM, XGBoost, decision trees, random forests, artificial neural networks) were applied. Model performance was evaluated using accuracy, sensitivity, precision, F1 score, MCC, AUC-ROC, and precision-recall curve (PRC) metrics. Additionally, the effects of other variables on cancer risk were analyzed using odds ratios, p-values, and effect sizes.

Results: The analysis showed that the XGBoost model showed the highest performance with an AUC value of 0.846 (0.841-0.850, 95% CI). LR analysis showed that older age, residence in Istanbul region, high haemoglobin, low ALT and some comorbidities were associated with cancer risk.

Conclusion: Our study revealed that the XGBoost model performed the best in cancer prediction. The effects of certain variables on cancer risk provide critical information for clinical applications and patient management. These findings support the use of ML models in healthcare applications and contribute to a better understanding of cancer.

Key words: big data, machine learning, artificial intelligence, early detection of cancer

İÇİNDEKİLER

ÖZET	i
ABSTRACT	ii
İÇİNDEKİLER	iii
TABLolar DİZİNİ	iv
ŞEKİLLER DİZİNİ	v
SİMGELER ve KISALTMALAR	vii
1. GİRİŞ	1
2. GENEL BİLGİLER	3
2.1. Kanser	3
2.2. Makine Öğrenmesi	7
2.3. Model Değerlendirme Yöntemleri	19
2.4. Literatürde Makine Öğrenmesi ile Kanser Tahminlemesi Çalışmaları	21
3. GEREÇ ve YÖNTEM	31
3.1. Gereç	31
3.1.1. Veri Toplama ve Hazırlama Süreçleri	31
3.2. Yöntem	33
3.2.1. Veri Ön İşleme (Temizliği ve Filtreleme)	35
3.2.2. Veri Bölme (Eğitim ve Test)	42
3.2.3. Model Seçimi	42
3.2.4. Model Eğitimi (Hiperparametre Optimizasyonu)	42
3.2.5. Model Değerlendirme	43
3.2.6. Model Testi ve Validasyonu	43
4. BULGULAR	44
5. TARTIŞMA	57
6. KISITLILIKLAR	64
7. SONUÇ VE ÖNERİLER	66
KAYNAKLAR	70
EKLER	77
ÖZGEÇMİŞ	90

TABLÖLAR DİZİNİ

Tablo 2.1.	2020 Kanser İstatistikleri: 2020 Türkiye Genel Bakış (GLOBOCAN, 2020)	4
Tablo 4.1.	Veri Seti Kanser Türlerinin Sıklığı (e-nabız 2018-2022)	44
Tablo 4.2.	Kontrol ve Deney Gruplarının Komorbidite, Demografik ve Klinik Özelliklerine Göre Özet İstatistiksel Analizleri	45
Tablo 4.3.	Etki büyüklüğü 0.2'den küçük olup Lojistik Regresyonda anlamlı sonuç veren komorbiditeler	50
Tablo 4.4.	Kontrol ve Deney Gruplarının Bölgelere Göre İstatistiksel Analizleri	50
Tablo 4.5.	Modeller için Elde Edilen En İyi Hiperparametreler	54
Tablo 4.6.	Kanser Tahmini Performans Sonuçları	55

ŞEKİLLER DİZİNİ

Şekil 2.1.	Türkiye Geneline 2020 Yılında En Yaygın Görülen Kansere Türlerinin İnsidans ve Ölüm Oranları (GLOBOCAN, 2020)	3
Şekil 2.2.	DVM Grafiği	11
Şekil 2.3.	ANN Grafiği	16
Şekil 2.4.	LR Grafiği	18
Şekil 3.1.	Makine Öğrenimi Süreci	35
Şekil 3.2.	Veri Setindeki Değişkenler Arasındaki Korelasyon Matrisi	39
Şekil 3.3.	Veri Setindeki PDW Veri Dağılımı	40

SİMGELER ve KISALTMALAR

ALT	: Alanin Aminotransferaz
ANN	: Yapay Sinir Ağı
AST	: Aspartat Aminotransferaz
AUC	: ROC Eğrisi Altındaki Alan
CRP	: C-Reaktif Protein
EHR	: Elektronik Sağlık Kayıtları
EOS%	: Eozinofil Yüzdesi
EOS#	: Eozinofil Sayısı
GB	: Gradient Boosting
HCT	: Hematokrit
HGB	: Hemoglobin
LR	: Lojistik Regresyon
LYM%	: Lenfosit Yüzdesi
LYM#	: Lenfosit Sayısı
MCH	: Ortalama Korpüsküler Hemoglobin
MCHC	: Ortalama Korpüsküler Hemoglobin Konsantrasyonu
MCC	: Matthews Korelasyon Katsayısı
MCV	: Ortalama Korpüsküler Hacim
MÖ	: Makine Öğrenmesi
MPV	: Ortalama Trombosit Hacmi
NB	: Naive Bayes
PCT	: Plateletcrit
PDW	: Trombosit Dağılım Genişliği

PLT	: Trombosit Sayısı
RBC	: Kırmızı Kan Hücresi Sayısı
RF	: Rastgele Ormanlar
RDW	: Kırmızı Hücre Dağılım Genişliği
ROC	: Alıcı İşletim Karakteristiği
SVM	: Destek Vektör Makineleri
TSH	: Tiroit Uyarıcı Hormon
WBC	: Beyaz Kan Hücresi Sayısı
WHO	: Dünya Sağlık Örgütü
YZ	: Yapay Zekâ

1. GİRİŞ

Kanser, dünya çapında ölümlere neden olan en yaygın hastalıklardan biridir. Hücrelerin kontrolsüz büyümesi sonucu oluşan bu hastalığın, birçok farklı türü bulunmaktadır ve tedavi edilmediğinde ölümcül sonuçlar doğurabilmektedir (Siegel R. L., 2020).

Ulusal elektronik sağlık kayıtları (EHR), kanser risk faktörlerini belirlemek için önemli bir veri kaynağı olarak büyük bir potansiyele sahiptir (Ross, et al., 2014). Bu kayıtlar, hastaların demografik bilgileri, tıbbi geçmişi, yaşam tarzı alışkanlıkları ve kanser aile öyküsü gibi değerli bilgileri içerir. Erken ve doğru klinik teşhis, hastalıkların etkili tedavisi için kritik bir öneme sahiptir. Her hastalık, moleküler ve hücresel düzeyde farklılıklara yol açar veya bu farklılıklardan kaynaklanır. Bu değişikliklerin bir kısmı, kan tetkiklerindeki değişimler aracılığıyla da tespit edilebilmektedir (Gunçar, et al., 2018). Bu verilerin analizi, daha önce bilinmeyen veya yetersiz anlaşılan kanser risk faktörlerini ortaya çıkarabilmektedir ve bu da hastalar için erken teşhis, önleme ve tedavi stratejilerinin geliştirilmesine katkıda bulunabilmektedir. Microsoft Araştırma Laboratuvarları'nda direktör olarak görev yapan ve sağlık alanında yapay zekâ uygulamaları konusunda deneyimli olan Dr. Eric Horvitz, "Elektronik sağlık kayıtları, büyük miktarda değerli bilgi içeren devasa taş ocaklarına benziyor ve şu anda bu verileri keşfe başlıyoruz" demiştir (Patel, et al., 2020). Bu nedenle, elektronik sağlık kayıtlarından elde edilen verilerin analizi ve yapay zekâ uygulamaları, sağlık sektöründe daha etkili teşhis ve tedavi yöntemlerinin geliştirilmesinde büyük bir potansiyele sahiptir.

Yapay zekâ tekniklerinden biri olan makine öğrenimi (MÖ) teknolojilerinin kullanımı, büyük ve karmaşık veri kümelerini analiz ederek, geleneksel istatistiksel yöntemlerle gözden kaçabilecek ilişkileri ve örüntüleri tespit etmeye yardımcı olmaktadır. Bu sayede, ulusal elektronik sağlık kayıtlarının analizi, kanserle ilgili risk faktörlerini daha iyi anlamamızı sağlamakta ve hastaların yaşam kalitesini ve sağ kalım oranlarını artırmada önemli bir rol oynayabilir.

Bu nedenle, son yıllarda MÖ teknikleri kanser teşhisinde kullanılmaya başlamıştır (Mitsala, et al., 2021). MÖ, kanserli hücrelerin belirli özelliklerini analiz etmek için

kullanılmakta ve daha doğru bir kanser teşhisi konulmasına yardımcı olmaktadır. Bu teknikler ilerleyen teknoloji ile kanser taramasında da kullanılmakta ve kanserli hücrelerin erken aşamalarda tespit edilmesine yardımcı olmaktadır.

Çalışmada, e-Nabız Kişisel Sağlık Kaydı Sistemi'nden alınan ve büyük veri ortamına taşınan verilerle oluşturulan büyük veri ortamından yararlanılmıştır. Bu kaynak, sağlık verilerini güvenli ve kolay erişilebilir bir şekilde sunarak, kapsamlı analizler için önemli bir temel sağlar. Bu tezde, e-Nabız Kişisel Sağlık Kayıtları verilerinden cinsiyet, yaş, il, tam kan ve bazı serum değerleri kullanılarak MÖ yardımıyla kanser tahmin modelleri geliştirilmeye çalışılmıştır.

Bu tez, şu sorulara cevap arayacaktır:

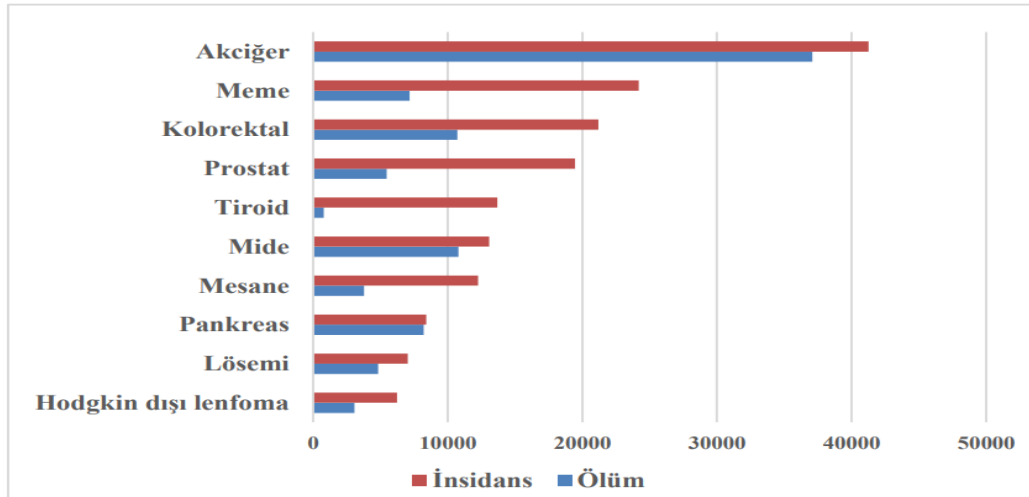
- Kanser risk faktörleri nelerdir?
- MÖ tekniklerinin kanser tahmininde performansları nedir?
- Cinsiyet, yaş, il, tam kan ve bazı serum değerleri kanser erken tahmin ve kanser tahmin faktörlerini belirlemede ne kadar etkilidir?

2. GENEL BİLGİLER

2.1. Kanser

Kanser, dünya çapında milyonlarca insanı etkileyen ve ölümlere neden olan karmaşık bir hastalıktır (Bray, et al., 2018). 2030 yılına kadar yıllık olarak tahmini 21,7 milyon yeni vaka ve 13 milyon kanser ölümü ile kanser, küresel bir halk sağlığı yüküdür (Ferlay, et al., 2019). Hücrelerin kontrolsüz bir şekilde büyümesi ve çoğalması sonucu oluşan kanser, vücudun herhangi bir bölgesinde meydana gelebilmektedir. Kanser türleri arasında akciğer, meme, kolorektal, prostat ve mide kanseri en yaygın olanlarıdır (Sung, et al., 2020).

Türkiye'de kanserle ilgili verilere göre, 2015 ile 2020 yılları arasında toplamda 581.636 kanser vakası kaydedilmiştir. 2020 yılında 233.834 kanser vakasının 126.335'inin ölümle neticelendiği görülmüştür. Önümüzdeki 20 yıllık süreçte, 2020'deki yaklaşık 234 bin olan kanser vakalarının sayısının 2040 yılında yaklaşık 393 bine yükselmesi ve ölümcül vaka sayısının ise yaklaşık 233 bine ulaşması beklenmektedir (GLOBOCAN, 2020). Türkiye'de yaygın olarak görülen kanser türleri ve bunlara ilişkin sayısal bilgiler, Şekil2.1. ve Tablo2.1. 'de sunulmuştur.



Şekil 2.1. Türkiye Geneline 2020 Yılında En Yaygın Görülen Kanser Türlerinin İnsidans ve Ölüm Oranları (GLOBOCAN, 2020)

Tablo2.1. 2020 Kanser İstatistikleri: 2020 Türkiye Genel Bakış (GLOBOCAN, 2020)

Cinsiyet	Kanser Vakaları	Nüfus Sayısı	75 Yaşına Değın Kanser Riski (%)	Kansere Bağlı Ölümmler	75 Yaşına Değın Kanserle İlgili Ölüm Riski (%)	5 Yıllık Kanser Görölme Oranı (2015-2020)
Erkek	132.816	41.636.125	29,2	78.949	18,2	300.071
Kadın	101.018	42.702.942	18,4	47.386	8,6	281.565
Toplam	233.834	84.339.067	23,3	126.335	13,0	581.636
		Erkek	Kadın		Toplam	
En Yaygın 5 Kanser Türü		Akciğır Prostat, Kolorektal Mesane Mide	Meme Tiroid Kolorektal Akciğır Rahim		Akciğır Meme Kolorektal Prostat Tiroid	

Kanser gelişiminin birçok nedeni vardır, bunlar genetik faktörler, yaşam tarzı faktörleri ve çevresel faktörler olarak sınıflandırılabilir (Anand et al., 2008). Genetik faktörler, aile öyküsü ve genetik mutasyonlar gibi, bireyin kanser geliştirme riskini etkilemektedir. Yaşam tarzı faktörleri arasında sigara içme, alkol tüketimi, fiziksel aktivite eksikliği ve sağlıksız beslenme bulunmaktadır. Çevresel faktörler ise, radyasyon ve kimyasal maddelere maruz kalma gibi, kanser riskini artırabilmektedir.

Kanserle ilgili risk faktörlerini belirlemek ve anlamak, erken teşhis ve önleme stratejileri geliştirme açısından önemlidir. Bu, hastaların yaşam kalitesini ve sağ kalım oranlarını artırabilmektedir (Colditz & Wei, 2012). Bu nedenle, kanser risk faktörlerini keşfetmek ve bu faktörlerle mücadele etmek için etkili yöntemler geliştirmek, kanser araştırmalarının temel amacıdır.

Dünya Sağlık Örgütü (WHO) verilerine göre 2030 yılına kadar kanserden kaynaklanan ölümlerin artması beklenmektedir (World Health Organization, 2022). Bu nedenle, WHO'nun önceliklerinden biri, hükümetler ve sağlık politikası yapıcılar için kanserle ilgili risk faktörlerini ve etkili önleme stratejilerini belirleyerek, dünya çapında kanserle mücadeleyi güçlendirmektir.

WHO, sigara içme ve alkol tüketimi gibi yaşam tarzı faktörlerinin kanser gelişimine katkıda bulunduğunu vurgulamaktadır (World Health Organization, 2022). Bu nedenle, WHO'nun öncelikli hedeflerinden biri, sigara ve alkol tüketimini azaltmak ve sağlıklı beslenme ve fiziksel aktiviteyi teşvik etmeye yönelik politikaları desteklemektir.

Kanser taramaları ve erken teşhis, WHO tarafından önemli bir kanser kontrol stratejisi olarak kabul edilir. Erken teşhis ve tedavi, kanserle mücadelede etkili bir şekilde yaşam kalitesini ve sağ kalım oranlarını artırabilir (World Health Organization, 2022). Bu nedenle, WHO, ülkelerin kanser tarama programlarını ve erken teşhis politikalarını desteklemekte ve geliştirmektedir.

Kanser türleri, büyüme hızı, yayılım şekli ve tedaviye yanıt gibi özellikler bakımından farklılık gösterir (Siegel, et al., 2023). Kanser, köken aldığı hücre çeşidine göre kategorize edilir. Beş ana kategori bulunmaktadır (Uluslararası Kanser Araştırma Ajansı, 2023):

Karsinomlar: Bu kanser türü, komşu dokuları ve organları işgal edebilir, lenf düğümlerine ve bedenin diğer kısımlarına yayılabilir. Bu gruptaki en yaygın kanser çeşitleri meme, prostat, akciğer ve kolon kanserleridir.

Sarkomlar: Kötü huylu kemik veya yumuşak doku tümörleridir. Sarkomların en sık rastlanan biçimleri leiomyosarkom, liposarkom ve osteosarkomdur.

Lenfoma ve miyelomlar: Lenfoma, lenfatik sistemin kanseri olup, tüm bedene yayıldığı için her yerde görülebilir. Miyelom ise, bağışıklık sisteminde rol alan antikorlar üreten plazma hücreleri adlı beyaz kan hücrelerinde başlar. Bu kanser, hücrenin antikor üretme yeteneğini etkileyebilmektedir.

Lösemi: Lösemi, beyaz kan hücrelerinin ve kan hücrelerini meydana getiren doku olan kemik iliğinin kanseridir. Birkaç farklı alt türü bulunmakta olup, en yaygın olanları lenfositik lösemi ve kronik lenfositik lösemidir.

Beyin ve omurilik kanserleri: Merkezi sinir sistemi kanserleri olarak adlandırılırlar. Bazıları iyi huylu iken, diğerleri büyüyebilmekte ve yayılabilmektedirler.

Kanser İstatistikleri 2023 raporuna göre, erkeklerde en yaygın üç kanser türü prostat, akciğer ve kolorektal kanser olup, erkeklerde teşhis edilen tüm kanser vakalarının yaklaşık %48'ini oluşturmaktadır. Öte yandan, kadınlarda en sık görülen üç kanser türü meme, akciğer ve kolorektal kanser olup, bu üç tür kadınlarda tespit edilen tüm kanser vakalarının %52'sinden biraz fazlasını kapsamaktadır (Siegel, et al., 2023). Bazı yaygın kanser türleri ve bu türlerle ilgili genel bilgiler aşağıdadır.

Meme kanseri: Kadınlarda en sık rastlanan kanser türüdür ve tüm kanser vakalarının yaklaşık %25'ini oluşturur (Sung et al., 2021). Risk faktörleri arasında yaş, aile öyküsü, genetik mutasyonlar ve hormonal faktörler bulunmaktadır.

Akciğer kanseri: Dünya genelinde en sık rastlanan kanser türü ve ölümlerin en büyük nedenidir (Sung et al., 2021). Sigara içme, radon gazına maruz kalma ve hava kirliliği, akciğer kanseri riskini artıran faktörlerdir.

Kolorektal kanser: Kalın bağırsak ve rektumda oluşan kanserdir ve dünya genelinde en yaygın üçüncü kanser türüdür (Sung et al., 2021). Yaş, aile öyküsü, diyabet ve yaşam tarzı faktörleri, kolorektal kanser riskini artırabilir.

Prostat kanseri: Erkeklerde en yaygın kanser türlerinden biridir ve yaşla birlikte risk artar (Sung et al., 2021). Aile öyküsü ve genetik faktörler, prostat kanseri riskini etkileyebilir.

Mide kanseri: Mide kanseri, dünya genelinde en yaygın kanser türlerinden biridir ve risk faktörleri arasında yaş, aile öyküsü, H. pylori enfeksiyonu ve tuzlu gıdaların tüketimi bulunmaktadır (Sung et al., 2021).

Karaciğer kanseri: Karaciğer kanseri, dünya genelinde önde gelen kanser ölümlerinden biridir ve risk faktörleri arasında kronik B veya C hepatiti, siroz ve aflotoksinlere maruz kalma bulunmaktadır (Sung et al., 2021).

2.2. Makine Öğrenmesi

Yapay zekâ (YZ), bilgisayarların insan zekâsını taklit etme amacıyla kullanılan bir araştırma disiplini. Makine öğrenimi ise, bilgisayar performansını geliştirmek adına matematiksel ve istatistiksel yöntemlerin kullanıldığı YZ'nin bir alt dalıdır (Shimizu & Nakayama, 2020). Makine öğrenimi, verilerden kuralları öğrenerek sorunları ele alır ve bu açıdan, bir doktorun deneyim ve gözlemlerine dayanarak hareket etmesine benzer. Algoritmalar, hasta düzeyinde gözlemlerle başlayarak, çok sayıda değişken arasından güvenilir şekilde sonuçları tahmin eden kombinasyonları bulmaya çalışır. Bu süreç, bir anlamda, geleneksel regresyon modellerine benzer: bağımsız değişkenler ve bunları birbirine bağlayan istatistiksel işlemler gibi. Ancak, makine öğreniminin gerçek gücü, büyük miktarda tahmin ediciyi bazen de gözlemlerden daha fazla tahmin edici ve bunları doğrusal olmayan ve son derece etkileşimli şekillerde entegre etme becerisidir. Bu yetenek, daha önce analiz etmeyi imkânsız kılan büyük boyutlu veya karmaşık veri türlerinin kullanılmasını mümkün kılar (Obermeyer & Emanuel, 2016).

Makine öğrenimi tekniklerinin temel hedefi, sınıflandırma, tahmin, kestirim gibi görevlerde kullanılabilen bir model oluşturmaktır. Öğrenme sürecinde en sık karşılaşılan görev sınıflandırmadır. Daha önce belirtildiği gibi, bu öğrenme işlemi, veri öğelerini önceden tanımlanmış birkaç sınıftan birine yerleştirir. Sınıflandırma modeli geliştirildiğinde, makine öğrenimi teknikleri ile eğitim ve genelleme hataları ortaya çıkar. İlki, eğitim verilerindeki yanlış sınıflandırma hatalarını ifade ederken, ikincisi, test verilerindeki beklenen hataları ifade eder (James, et al., 2013).

İyi bir sınıflandırma modeli, eğitim setine iyi uyum sağlamalı ve tüm örnekleri doğru bir şekilde sınıflandırmalıdır. Eğitim hata oranları düşse bile, bir modelin test hata oranları artmaya başladığında, modelde aşırı uyum (overfitting) olgusu meydana gelir. Bu durum, model karmaşıklığı ile ilgilidir. Model karmaşıklığı arttıkça, modelin eğitim hataları azaltılabilir. İdeal olarak, aşırı uyuma karşı hassas olmayan bir modelin en düşük genelleme hatasıyla en uygun karmaşıklığa sahip olması beklenir (Mitchell, 1997).

Öğrenme algoritmasının beklenen genelleme hatasını analiz etmek için resmi bir yöntem, önyargı-varyans ayrışmasıdır. Belirli bir öğrenme algoritmasının önyargı bileşeni, algoritmanın hata oranını ölçerken, varyans bileşeni, belirli boyuttaki tüm olası eğitim ve test setleri üzerindeki ikinci bir hata kaynağıdır. Bir sınıflandırma modelinin genel beklenen hatası, önyargı ve varyans toplamından oluşur (Geman, et al., 1997).

Bir veya daha fazla makine öğrenimi tekniği kullanarak bir sınıflandırma modeli elde edildikten sonra, sınıflandırıcının performansını değerlendirmek önemlidir. Her modelin performans analizi, doğruluk, duyarlılık (hassasiyet), seçicilik (özgüllük) ve ROC (Receiver Operating Characteristics) AUC (Area Under Curve, Eğri Altındaki Alan) ve benzeri ölçütler ile değerlendirilir. Duyarlılık, gerçek pozitiflerin doğru bir şekilde tespit edilme oranını ifade ederken, seçicilik ise gerçek negatiflerin doğru bir şekilde tespit edilme oranını gösterir. Ayrıca duyarlılık ve özgüllük, genellikle klinik tanı testlerinin performansını değerlendirmek ve bu testlerin ne kadar güvenilir ve tutarlı olduğunu belirlemek amacıyla kullanılır (Danjuma, 2015). Doğruluk, sınıflandırıcının tüm örnekler arasında doğru tahminler yapma oranını temsil eder ve eğrinin altındaki alan (AUC), sınıflandırıcının performansının toplam değerlendirmesi için kullanılan bir metriktir. Bu ölçütler, modelin genel başarısını ve uygulanabilirliğini değerlendirmede kullanılır (Fawcett, 2006).

Makine öğrenimi, son yıllarda klinik kanser araştırmalarında popüler hale gelmiş ve kanser tahmin performansında önemli başarılar elde edilmiştir (Huang, et al., 2020). Makine öğrenimi yöntemlerinin kanser progresyonu konusundaki bilgilerimizi artırma potansiyeli olsa da bu yöntemlerin günlük klinik uygulamada kullanılabilmesi için uygun doğrulama düzeyine ulaşması gerekmektedir (Konstantina , et al., 2015). Makine öğrenimi yöntemlerinin kanser araştırmalarında giderek daha fazla kullanılması göz

önünde bulundurulduğunda, kanser riskini, tahminini veya hasta sonuçlarını modellemek için en yaygın olarak kullanılan teknikleri burada ele alıyoruz.

Bir dizi YZ yöntemi bulunmaktadır, fakat tümü kanser tespitinde kullanılabilir değildir. Her bir yöntemin, spesifik sebeplerle uygulanabilir olup olmadığını belirleyen kendine özgü nitelikleri vardır. Yapay zekâ alanında öne çıkan bazı teknikler şunlardır: sinir ağları, genetik algoritmalar, bulanık mantık kümeleme, destek vektör makinesi (DVM) (SVM-Support Vector Machine), karar ağaçları (KA), rastgele orman (RO), Lojistik regresyon (LR-Logistic Regression), XGBoost (Extreme Gradient Boosting) ve yapay sinir ağları (ANN- Artificial Neural Networks). Bu yöntemler, sınıflandırma veya regresyon yaklaşımlarını belirler ve kanser tespiti için kapsamlı bir teknik ortaya koyabilmektedir (Patel, et al., 2020).

Kanser tespitinde en çok kullanılan makine öğrenmesi yöntemlerine aşağıda yer verilmiştir.

Destek Vektör Makineleri (DVM) (Support Vector Machine-SVM): Destek Vektör Makineleri (DVM), denetimli öğrenme algoritmaları kategorisinde yer alan, sınıflandırma ve regresyon analizi için kullanılan etkili bir yöntemdir. DVM, verilerin sınıflandırılması için karar düzeylerini kullanarak sınırlar belirler. Karar düzeyi, farklı sınıflardan gelen veri noktalarını birbirinden ayırır ve bu algoritma, veri kümesini sınıflara bölen hiperdüzlem veya hiperdüzlem kombinasyonlarını oluşturur. İki boyutlu bir alanda, iki sınıfı birbirinden ayıran hiperdüzlem bir çizgi olarak temsil edilir (Şekil 2.2.) (Geron, 2019).

DVM'nin temel bileşenleri ve çalışma prensipleri şunlardır:

Hiperdüzlem: Veri kümesini sınıflara bölen ve iki sınıfı birbirinden ayıran geometrik düzlemdir.

Marj: Karar sınırına en yakın veri noktaları arasındaki uzaklıktır. DVM, bu marjı maksimize etmeye çalışır.

Destek Vektörleri: Karar sınırına en yakın olan ve sınırların belirlenmesinde etkili olan veri noktalarıdır. Bu noktalar, hiperdüzlemin belirlenmesinde kritik rol oynar.

DVM'nin çalışma prensibi şu adımlardan oluşur:

Veri Kümesinin Hazırlanması: Giriş verileri sınıflandırma veya regresyon analizi için hazırlanır.

Hiperdüzlem Belirlenmesi: Veriyi sınıflara ayıran en uygun hiperdüzlem bulunur. Bu hiperdüzlem, sınıflar arasındaki marjı en büyük hale getirecek şekilde seçilir.

Marjın Maksimizasyonu: Marjın genişletilmesi için hiperdüzlem, destek vektörleri kullanılarak optimize edilir.

Sınıflandırma veya Regresyon: Yeni veri noktaları, belirlenen hiperdüzlem kullanılarak sınıflandırılır veya tahmin edilir.

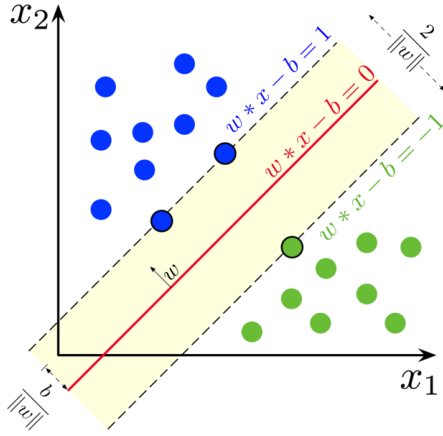
DVM'nin avantajları şunlardır:

Yüksek Doğruluk: Özellikle düşük örnek sayılı veya yüksek boyutlu veri kümelerinde etkili sonuçlar elde edilmesini sağlar.

Genelleştirme Yeteneği: Doğrusal ve doğrusal olmayan problemler için kullanılabilir ve genelleme yeteneği yüksektir.

Esneklik: Kernel fonksiyonları kullanarak, veriyi daha yüksek boyutlu uzaylara taşıyarak doğrusal olmayan karar sınırları oluşturabilir.

Destek Vektör Makineleri, düşük örnek sayılı veya yüksek boyutlu veri kümelerinde yüksek doğrulukla çalışabilmesi ve esnek yapısı ile sınıflandırma ve regresyon problemlerinde sıklıkla tercih edilen bir yöntemdir. (Geron, 2019).



Şekil 2.2. DVM Grafiği (Wikipedia)

Karar Ağaçları (KA) (DT-Decision Tree): Karar Ağaçları (KA), denetimli makine öğrenimi türlerinden biri olup, verilerin belirli parametrelere göre sürekli olarak ayrıldığı parametrik olmayan bir yöntemdir. Veri özniteliklerine dayalı basit karar kuralları öğrenerek, hedef değişkenin değerini tahmin eden bir model oluşturur. Karar ağacının en bilinen yöntemi ID3 (Iterative Dichotomiser 3) olsa da C4.5 (Successor of ID3) ve CART (Classification and Regression Tree) gibi teknikler de mevcuttur. Bu yöntem, doğrusal sınırlı problemlerle daha etkili çalışır ve yüksek doğruluk ile zamanın kritik olduğu durumlarda kullanılabilir (Murphy, 2021).

Karar ağaçlarının temel bileşenleri ve çalışma prensipleri şunlardır:

Giriş Katmanı: Giriş verilerini alır ve bu verileri karar ağacının diğer düğümlerine iletir.

Düğüm (Node): Veri kümesini belirli kurallara göre bölen ve sınıflandırma kararlarını veren temel birimdir.

Yaprak Düğüm (LeafNode): Karar ağacının son düğümü olup, sınıflandırma veya tahmin sonucunu temsil eder.

Karar ağaçları, veri kümesini böl-yönet yaklaşımıyla belirli kurallara göre küçük alt gruplara bölmeyi amaçlar. Ağacın oluşturulması sırasında, eğitim verilerinin hangi sırayla kullanılması gerektiğine karar vermek için entropi ölçümü kullanılır. Entropi ölçümü ne

kadar yüksekse, belirsizlik o kadar fazladır. Bu nedenle, karar ağacının kökünde entropi ölçümü en düşük olan alanlar tercih edilir. Karar düğümleri, veri kümesinde karar verme, sınıflandırma veya tahmin yapma amacıyla kullanılır. Entropi ölçümü ve bilgi kazancı, bu düğümlerin belirlenmesinde kritik rol oynar.

Karar ağaçlarının avantajları şunlardır:

Hız ve Anlaşılabilirlik: Karar ağaçları, hızlı ve kolay anlaşılabilirliği ile öne çıkar.

Yorumlanabilirlik: Karar ağaçları, sonuçların yorumlanmasını kolaylaştırır ve kullanıcıların modeli anlamasını sağlar.

Veri Tabanlarıyla Entegrasyon: Karar ağaçları, veri tabanları ile iyi entegre edilebilir ve büyük veri setleri üzerinde etkili bir şekilde çalışabilir.

Yüksek Doğruluk: Karar ağaçları, özellikle veri kümesi üzerinde çok fazla gürültü ve belirsizlik olmadığında yüksek doğruluk sunar.

Bu çalışmada, Karar Ağaçlarının seçilmesinin nedeni, hızlı ve etkili bir şekilde sınıflandırma doğruluğu sunabilmesidir. Ayrıca, karar ağaçları algoritmaları hekimler tarafından kolay anlaşılabilir olmaları nedeniyle tercih edilmiştir. Bu yöntem, karar ağaçlarının anlaşılabilir ve yorumlanabilir yapısı nedeniyle özellikle veri madenciliği ve makine öğrenimi uygulamalarında popülerdir (Murphy, 2021).

Rastgele Orman (RO) (Random Forest-RF)

Rastgele Orman, hem sınıflandırma hem de regresyon görevlerinde kullanılabilen güçlü ve çok yönlü bir denetimli öğrenme algoritmasıdır. Bu algoritma, eğitim sırasında birçok karar ağacı oluşturarak ve bu ağaçların sınıflarının modunu (sınıflandırma) veya ortalama tahminini (regresyon) çıkararak çalışır. RO'nın temel konsepti, eğitim verilerinin ve özelliklerin rastgele seçilen alt kümelerinden birçok karar ağacı oluşturmak ve bu ağaçların tahminlerini birleştirerek doğruluğu artırmak ve aşırı öğrenmeyi (overfitting) kontrol etmektir (Breiman, 2001).

Rastgele Orman algoritması aşağıdaki adımlardan oluşur:

Bootstrap Örnekleme: Orijinal verilerin rastgele alt kümeleri, tekrar kullanılarak oluşturulur. Her alt küme, bir bireysel karar ağacı oluşturmak için kullanılır.

Rastgele Özellik Seçimi: Bir ağaçtaki her düğümde, bölme için aday olarak rastgele bir özellik seçimi yapılır. Bu işlem, ağaçlar arasında çeşitlilik sağlar.

Ağaç İnşası: Her karar ağacı, budama yapılmadan maksimum düzeyde büyütülür.

Toplama: Tüm ağaçların tahminleri birleştirilerek nihai sonuç üretilir. Sınıflandırma görevleri için bu, çoğunluk oylama mekanizmasını içerirken, regresyon görevleri için tüm ağaç tahminlerinin ortalaması alınır.

Rastgele Orman'ların sunduğu avantajlar şunlardır:

Aşırı Öğrenmeye Karşı Dayanıklılık: Birden çok ağacın ortalaması alınarak model, özellikle bireysel karar ağaçlarına kıyasla aşırı öğrenme riskini azaltır.

Özellik Önemliliği: RO, veri kümesindeki farklı özelliklerin önemini tahmin etmek için kullanılabilir, bu da özellik seçimi ve veri özelliklerinin anlaşılmasında yardımcı olur.

Eksik Değerleri Ele Alma: RO, eksik değerleri etkili bir şekilde ele alabilir ve bu da modelin dayanıklılığını artırır.

Rastgele Orman, veri ve özellik seçiminde rastgelelik tanıtarak karar ağaçlarının performansını artıran bir topluluk öğrenme yöntemidir ve çeşitli tahmin modelleri görevleri için sağlam ve doğru bir araçtır (Ho, 1995).

XGBoost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting), makine öğrenimi yarışmalarında popüler hale gelen güçlü ve etkili bir denetimli öğrenme algoritmasıdır. Özellikle yüksek boyutlu veri setlerinde hızlı ve doğru tahminler yapabilme yeteneği ile bilinir. XGBoost, karar ağaçlarının bir araya getirilmesiyle oluşturulan bir topluluk yöntemidir ve bu ağaçlar,

sıralı olarak eğitilir. Her yeni ağaç, önceki ağaçların hatalarını düzeltmek için eklenir, bu da modelin genel doğruluğunu artırır (Chen & Guestrin, 2016).

XGBoost algoritmasının temel adımları şunlardır:

Başlangıç Modeli: İlk olarak, veriler üzerinde basit bir model (genellikle bir karar ağacı) eğitilir.

Artık Hesaplama: Mevcut modelin hatalarını (artıkları) hesaplar. Bu artıklar, yeni ağaçlar eğitmek için hedef olarak kullanılır.

Ağaç Eklenmesi: Her yeni ağaç, önceki modelin hatalarını düzeltmek için eğitilir ve mevcut modele eklenir.

Model Güncellemesi: Yeni ağacın katkısı, belirli bir öğrenme oranı ile ağırlıklandırılarak modele eklenir.

Durma Kriteri: Belirli bir sayıdaki ağaçtan sonra ya da model performansındaki iyileşme durduğunda eğitim süreci durdurulur.

XGBoost'un avantajları şunlardır:

Hız ve Verimlilik: Paralel hesaplama yetenekleri ve optimizasyon teknikleri sayesinde diğer gradient boosting algoritmalarına kıyasla daha hızlıdır.

Genelleştirme Yeteneği: Düzenleştirme (regularization) teknikleri kullanarak aşırı öğrenmeyi kontrol eder ve daha iyi genelleştirme sağlar.

Eksik Veri İşleme: Eksik verileri etkili bir şekilde işleyebilir, bu da modelin dayanıklılığını artırır.

Özellik Önemliliği: XGBoost, modelde kullanılan özelliklerin önemliliğini belirlemek için araçlar sunar, bu da özellik seçimi ve veri analizi açısından değerlidir.

XGBoost, güçlü ve esnek bir gradient boosting algoritması olarak, özellikle büyük ve karmaşık veri setleri üzerinde yüksek performans sergileyen bir modelleme aracıdır (Friedman, 2001).

Yapay Sinir Ağları (YSA)(ANN-Artificial Neural Networks): Yapay Sinir Ağları (YSA), insan beyninin çalışma şeklini taklit eden, öğrenme ve örüntü tanıma yeteneğine sahip güçlü bir makine öğrenimi modelidir (Haykin, 1999). YSA, birbirine bağlı düğümler veya nöronlardan oluşur ve bu nöronlar katmanlar halinde organize edilmiştir. Genellikle bir giriş katmanı, bir veya daha fazla gizli katman ve bir çıkış katmanı bulunur (Şekil 2.3.).

Giriş Katmanı: Dışarıdan gelen verileri alan ve bu verileri ağı diğer katmanlarına ileten katmandır.

Gizli Katmanlar: Daha karmaşık özellikler ve ilişkileri öğrenmeye yardımcı olan katmanlardır. Bir YSA birden fazla gizli katmana sahip olabilir ve bu katmanların sayısı ve büyüklüğü, ağı karmaşıklığını belirler.

Çıkış Katmanı: Gizli katmanlardan gelen işlemlerin sonucunu alır ve nihai tahmin veya sınıflandırmayı yapar.

YSA'nın çalışma prensibi şu adımlardan oluşur:

İleri Yayılım (Forward Propagation): Giriş verileri, giriş katmanından başlayarak, gizli katmanlar aracılığıyla çıkış katmanına kadar iletilir. Her nöron, aldığı girdileri belirli bir ağırlık ve aktivasyon fonksiyonu kullanarak işler ve bir çıktı üretir.

Hata Hesaplama: Çıkış katmanında üretilen tahminler, gerçek değerlerle karşılaştırılır ve hata hesaplanır.

Geri Yayılım (Backpropagation): Hesaplanan hata, ağı boyunca geriye doğru iletilir ve her nöronun ağırlıkları bu hatayı minimize edecek şekilde güncellenir. Bu işlem, ağı öğrenme sürecini oluşturur.

Ağırlık Güncelleme: Öğrenme oranına bağlı olarak ağırlıklar güncellenir ve modelin doğruluğu artırılır.

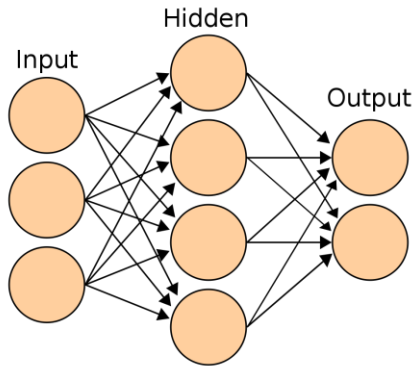
Yapay Sinir Ağlarının avantajları şunlardır:

Karmaşık İlişkileri Öğrenme: YSA, doğrusal olmayan ve karmaşık ilişkileri modelleyebilir, bu da onu geniş bir uygulama yelpazesi için uygun hale getirir.

Öğrenme Yeteneği: Büyük veri setlerinden öğrenebilir ve bu veri setlerindeki gizli kalıpları ortaya çıkarabilir.

Genelleştirme: Aşırı öğrenmeyi kontrol etmek için düzenleme teknikleri kullanarak yeni veriler üzerinde iyi genelleme yapabilir.

YSA'lar, biyolojik sinir ağlarının çalışma prensiplerinden esinlenerek geliştirilmiş, karmaşık ve doğrusal olmayan ilişkileri modelleyebilen makine öğrenme algoritmalarıdır (Rumelhart, 1986). Sınıflandırma, regresyon, örüntü tanıma ve önerme gibi problemleri çözmek için uygulanırlar. Modelin öğrenme süreci, gerçek değerlerle modelin tahminleri arasındaki hatayı azaltacak şekilde nöronların bağlantı ağırlıklarını ve biaslarını güncellemeye dayanır. Bu süreç, genellikle geri yayılım (backpropagation) algoritması ve gradyan inişi gibi optimizasyon yöntemleri kullanılarak gerçekleştirilir. TensorFlow 2, YSA'ların hızlı ve etkili bir şekilde uygulanmasına, eğitilmesine ve değerlendirilmesine olanak tanıyan popüler bir derin öğrenme kütüphanesidir (Sarang , 2020).



Şekil 2.3. ANN Grafiği (Wikipedia)

Lojistik Regresyon (LR): Lojistik regresyon (LR), denetimli sınıflandırma algoritmalarından biri olup, bir veya daha fazla bağımsız değişken içeren veri kümesini analiz ederek lojistik eğri üzerindeki verileri eşleştirir ve bir sonuca ulaşır (Şekil 2.4.). LR'nin çeşitli tipleri arasında ikili, sıralı ve multinominal bulunmaktadır. Genellikle ikili LR yöntemi yaygın olarak tercih edilir. Lojistik regresyonun hedefi, en az sayıda değişken kullanarak bağımsız değişkenler arasındaki ilişkiyi belirlemek ve bir model oluşturmaktır (Burkov, 2019).

LR, hem olasılık hesaplama hem de sınıflandırma amacıyla kullanılabilir. Bu yöntemin temelinde sigmoid fonksiyonu yer alır. Sigmoid fonksiyonu, negatif ve pozitif sonsuz değerler arasındaki girdileri alabilir, ancak çıktılar her zaman 0 ile 1 arasındadır. Fonksiyondaki xx değişkeni, modelde kullanılan tüm bağımsız değişkenleri temsil ederken, $f(x)$ ise bu değişkenlerin varlığının olasılığını gösterir (Haykin, 1999).

Lojistik regresyon, ilgili özniteliklerin olasılığını tahmin etmek için formül katsayılarını ve ilgililik seviyelerini üretir. Sigmoid fonksiyonunun minimum çıktı değeri 0, maksimum çıktı değeri ise 1'dir. Eğer xx değeri 0 ise, sigmoid fonksiyonunun çıktısı 0,5'tir. Bu nedenle, sınıflandırma problemleri için çıktı 0,5'ten büyükse sonuç 1, küçükse 0 olarak değerlendirilir.

LR'nin temel bileşenleri şunlardır:

Bağımsız Değişkenler: Modelin tahmin yapması için kullanılan girdilerdir.

Bağımlı Değişken: Tahmin edilmeye çalışılan hedef değişkendir.

Sigmoid Fonksiyonu: Lojistik regresyonun temelini oluşturan matematiksel fonksiyondur.

Lojistik regresyonun çalışma prensibi şu adımlardan oluşur:

Model Oluşturma: Bağımsız değişkenler kullanılarak bir lojistik model oluşturulur.

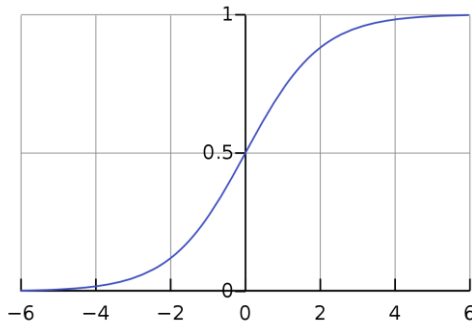
Olasılık Hesaplama: Sigmoid fonksiyonu kullanılarak her bir gözlem için olasılık hesaplanır.

Sınıflandırma: Hesaplanan olasılıklar, belirli bir eşik değeri (genellikle 0.5) kullanılarak sınıflara ayrılır.

Model Değerlendirme: Modelin doğruluğu, çeşitli performans metrikleri kullanılarak değerlendirilir.

Odds Oranı ve Güven Aralıkları: Lojistik regresyon analizinde, odds oranı (odds ratio) bağımsız değişkenlerin bağımlı değişken üzerindeki etkisini ölçer. Odds oranı, bağımsız değişkendeki bir birimlik artışın bağımlı değişken üzerindeki etkisini gösterir. Eğer odds oranı 1'den büyükse, bağımsız değişkenin artması bağımlı değişkenin olma olasılığını artırır; 1'den küçükse, bu olasılığı azaltır. Ayrıca, her bir odds oranı için güven aralığı (confidence interval) hesaplanır. Bu güven aralıkları, tahminin doğruluğunu ve istatistiksel önemliliğini belirlemeye yardımcı olur. Örneğin, güven aralığı 1'i içermiyorsa, odds oranı istatistiksel olarak anlamlı kabul edilir.

Lojistik regresyon, değişkenler arasındaki ilişkiyi belirleyebilmesi ve özniteliğin varlığının olasılığını tahmin edebilmesi özellikleriyle sınıflandırma problemlerinde sıkça tercih edilen bir yöntemdir.



Şekil 2.4. LR Grafiği (Wikipedia)

2.3. Model Değerlendirme Yöntemleri

Makine öğrenimi modellerinin performansını değerlendirmek, doğru ve güvenilir sonuçlar elde etmek için kritik bir adımdır. Model değerlendirme, bir modelin ne kadar iyi çalıştığını ve gelecekteki veri üzerinde nasıl performans göstereceğini belirlemeye yardımcı olur. Aşağıda, lojistik regresyon (LR), karar ağaçları (KA), destek vektör makineleri (DVM), yapay sinir ağları (YSA) ve diğer makine öğrenimi algoritmaları için yaygın olarak kullanılan model değerlendirme yöntemleri açıklanmaktadır.

Çapraz Doğrulama (Cross-Validation):

K-Fold Çapraz Doğrulama: Veri seti, k eşit parçaya bölünür ve her parça bir kez test seti olarak kullanılırken, kalan k-1 parça eğitim seti olarak kullanılır. Bu işlem k kez tekrarlanır ve sonuçlar ortalanarak model performansı değerlendirilir.

Leave-One-Out Çapraz Doğrulama (LOOCV): Veri集中的 her bir örnek sırasıyla test seti olarak kullanılırken, geri kalan veriler eğitim seti olarak kullanılır. Bu yöntem, özellikle küçük veri setleri için uygundur.

Karışıklık Matrisi (Confusion Matrix):

Sınıflandırma problemlerinde modelin doğru ve yanlış sınıflandırmalarını gösterir. True Positive (TP), True Negative (TN), False Positive (FP), ve False Negative (FN) değerlerini içerir.

Kesinlik (Precision): $TP / (TP + FP)$, modelin pozitif tahminlerinin doğruluğunu gösterir.

Duyarlılık (Recall): $TP / (TP + FN)$, modelin gerçek pozitifleri ne kadar iyi tespit ettiğini gösterir.

F1 Skoru: $2 * (Precision * Recall) / (Precision + Recall)$, modelin dengeli performansını ölçer.

Özgüllük / Seçicilik (Specificity) veya TNR (True Negative Rate): $TN / (TN + FP)$, modelin gerçek negatifleri ne kadar iyi tespit ettiğini gösterir.

ROC Eğrisi (Receiver Operating Characteristic Curve) ve AUC (Area Under Curve):

ROC eğrisi, modelin çeşitli eşik değerlerinde TPR ve FPR (False Positive Rate) değerlerini gösterir.

AUC, ROC eğrisinin altındaki alanı temsil eder ve modelin genel performansını değerlendirir. 1'e yakın değerler daha iyi performansı gösterir.

Doğruluk (Accuracy):

Modelin doğru sınıflandırdığı örneklerin oranıdır. $(TP + TN) / (TP + TN + FP + FN)$. Ancak, sınıflar arasındaki dengesizlik durumlarında yanıltıcı olabilir.

Kappa İstatistiği (Cohen's Kappa):

İki sınıflandırıcı arasındaki uyumu değerlendirir ve rastgele tahminlerin etkisini yok sayar. -1 ile 1 arasında bir değer alır, 1 mükemmel anlaşmayı gösterir.

Ortalamalar ve Sapmalar:

MSE (Mean Squared Error): Regresyon problemlerinde tahmin edilen ve gerçek değerler arasındaki farkların karelerinin ortalamasıdır.

RMSE (Root Mean Squared Error): MSE'nin kareköküdür ve hata oranını aynı birimde ifade eder.

MAE (Mean Absolute Error): Tahmin edilen ve gerçek değerler arasındaki farkların mutlak değerlerinin ortalamasıdır.

Model Karmaşıklığı ve Aşırı Öğrenme (Overfitting): BIC (Bayesian Information Criterion) ve AIC (Akaike Information Criterion): Modelin karmaşıklığını ve veri setine uyumunu değerlendirir. Daha düşük değerler genellikle daha iyi modeli gösterir.

Hata Matriksi (Error Matrix): Regresyon problemlerinde gerçek ve tahmin edilen değerlerin karşılaştırıldığı matristir. Hataların dağılımını ve modelin performansını görselleştirir.

Bu yöntemler, farklı makine öğrenimi algoritmalarının performansını değerlendirmek için kullanılır ve modellerin doğruluğunu, duyarlılığını, genelleştirme yeteneğini ve diğer kritik metriklerini belirlemeye yardımcı olur. Model değerlendirme, bir modelin gerçek dünyada ne kadar iyi performans göstereceğini anlamak için gereklidir ve bu nedenle makine öğrenimi sürecinin önemli bir parçasıdır.

2.4 Literatürde Makine Öğrenmesi ile Kanser Tahminlemesi Çalışmaları

Akademik literatürde, kanser tahmini ile ilgili çözümlere dair birçok çalışma bulunmaktadır. ML kullanılarak yapılan kanser tahminlemesinde laboratuvar verileri, hasta görüşme anketleri, radyolojik görüntüler gibi verilerin kullanılmaktadır. Birçok çalışmada, bilgisayarlı tomografi (CT- Computed Tomography), manyetik rezonans görüntüleme (MRI- Magnetic Resonance Imaging), X-ışını radyografisi ve pozitron emisyon tomografisi (PET- Positron Emission Tomography) gibi görüntüleme yöntemlerinden elde edilen veriler, temel olarak derin öğrenme mimarilerine dayanan otomatik teşhis modelleri oluşturmak amacıyla kullanılmıştır. Bu çalışmada, kanser tanısı ve saptanması için makine öğrenimi temelli modellerin geliştirilmesinde kullanılan görüntüleme, klinik, genomik ve diğer ilgili tıbbi verilere dayalı öncü çalışmaları sunmaktayız. Biz özellikle çalışmamızda da kullanacağımız rutin kan testleri ya da rutin alınan elektronik sağlık kaydı bilgilerini kullanarak yapılan çalışmaları burada ele almaya çalıştık.

2014 yılında yayınlanan retrospektif bir araştırma makalesinde kanser hastalarının hayatta kalma sürelerini tahmin etmek için makine öğrenimi yöntemleri kullanılmıştır. Bu

çalışmada, elektronik idari kayıtlar ve bir kanser kayıt sistemi kullanılarak, hastaların demografik, klinik ve tedavi bilgilerine dayalı olarak hayatta kalma sürelerinin tahmin edilmesi amaçlanmıştır. Kullanılan ML yöntemleri arasında lojistik regresyon (LR), destek vektör makineleri (DVM), rastgele orman (RO) ve Gradient Boosting (GB) gibi denetimli öğrenme algoritmaları yer almaktadır. Her yöntemin performansı, doğruluk, duyarlılık, özgüllük ve F1 skoru gibi çeşitli metriklerle değerlendirilmiştir. Çalışmanın sonuçları, makine öğrenimi yöntemlerinin kanser hastalarının hayatta kalma sürelerini etkili bir şekilde tahmin etmede başarılı olduğunu göstermiştir (Gupta, et al., 2014).

Joaquín ve arkadaşları tarafından 2016 yılında gerçekleştirilen bir çalışma, mevcut biyobelirteçler ile klinik ve demografik verilere dayalı bir kolon kanseri tahmin indeksi oluşturmayı ve doğrulamayı hedeflemiştir. Çok değişkenli lojistik regresyon kullanarak kan hemoglobin, kan serumu ve diğer özelliklere dayalı bir kolon kanseri tahmin modeli geliştirilmiştir; bu model COLONPREDICT olarak adlandırılmıştır. Araştırma sonuçları, COLONPREDICT'in kolorektal kanser (CRC) tespiti için oldukça doğru bir tahmin modeli olduğunu göstermiştir (Cubiella, et al., 2016).

Yarun ve arkadaşları 2016'da kolorektal kanser (CRC) riski altındaki bireyleri tespit etmek amacıyla kan sayımları, yaş ve cinsiyeti analiz eden iki uluslu retrospektif bir çalışma temelinde bir model geliştirip ve doğruladılar. Model İsrail veri setinin bir kısmı ile eğitilip, kalan kısmı ve İngiltere veri seti ile test edilip doğrulanmıştır. CRC tanısından 3-6 ay önce yapılan tüm tam kan tetkikleri (CBC-complete blood count) göz önünde bulundurularak elde edilen GBM (gradient boosting machine) ve RF yöntemlerinin birleşiminden elde edilen modelin ROC AUC değeri 0.81, odds oranı 40, seçicilik %94 olarak bulunmuştur (Kinar, et al., 2016).

Jacqueline ve arkadaşları 2017 yılında kolorektal kanser tahmini için 2,5 milyon hastanın elektronik sağlık kayıtlarından bir dizi makine öğrenimi aracıyla retrospektif bir analiz gerçekleştirdiler. Yaklaşık 3 milyon hasta ile yapılan bu çalışmada, rutin olarak toplanan birinci basamak verileri kullanılarak lojistik regresyon analizi yapılmıştır. Analiz sonucunda bir risk faktörü hesaplanmış ve %99,5 seçicilik sağlayan risk puanı değeri, önümüzdeki iki yıl boyunca %8,8'lik bir teşhis riski öngördüğünü buldular (Birks, et al., 2017).

Goshen ve arkadaşları 2017 yılında standart laboratuvar kayıtlarını kullanarak Maccabi Health Services (MHS) verilerinden kolon kanseri varlığı tahmininde bulundular. İsrail'in ikinci büyük elektronik sağlık kaydına sahip olan MHS verileri ile 2 milyon hasta ile yapılan tahminlemede 40 yaş üstü bireyler çalışmaya dahil edilmiştir. Hastaların 30-180 gün önce kaydedilen bir veya daha fazla rutin kan testi değerleri kullanılarak 2001-2011 arasındaki kanser tanısı tahminlemesi için bir makine öğrenimi tabanlı bir algoritma olan MeScore kullanılmıştır. Bu skorlama sonucunda elde edilen risk değerleri sonucunda CRC klinik tanıdan 2 yıl öncesine kadar tahmin edebilmiştir, ancak 2 yıldan uzun sürelerde ayırıcı yeteneği düşmüştür (Goshen, et al., 2017).

Hornbrook ve arkadaşları 2017 yılında cinsiyet, yaş ve tam kan sayımı verilerinden erken kolorektal kanser tespit eden bir makine öğrenimi modeli geliştirdiler. KPNW isimli bir Amerikan sigorta şirketi elektronik tıbbi kayıt sisteminde kayıtlı yaklaşık 17 bin hasta ile yapılan çalışmada, kolorektal kanser olasılığının daha yüksek olduğunu gösteren ve kolonoskopi için sevk edilmeye yönlendiren kan sonuçları olan hastaları tanımlanmaya çalışılmıştır. Bu CRC belirleme modelinin temel hedefi, demografik niteliklere ve detaylı sağlık geçmişi bilgilerine sahip bir yetişkin nüfusu içinde, geçmiş laboratuvar test sonuçlarına dayalı olarak CRC riski yüksek olan bireyleri saptamaktır. Tam kan testlerini içeren sağlık verilerinin algoritma temelli değerlendirmesinde, yalnızca HGB'ye göre, tam kan testlerinden 6 ay sonrası ve 6-12 ay arasında CRC vakalarını belirlemede daha fazla hassasiyete sahip olduğu görüldü. Karar ağaçları ve çapraz doğrulama tekniklerinin birleşiminden oluşan ColonFlag algoritması temelli inceleme sonucunda CRC tanısından önceki 6-12 aya göre ilk 12 ay içerisinde tespit edilen CRC vakalarını belirlemekte ve 40-89 yaş arası genel nüfusa kıyasla 50-75 yaş aralığındaki nüfusta CRC vakalarını saptamada daha fazla hassasiyet göstermiştir. KPNW elektronik sağlık kayıtlarına uygulanan ColonFlag için ROC AUC değeri 0,81 olarak bulunmuştur (Hornbrook, et al., 2017).

2018 yılında Robert ve arkadaşları, tam kan sayımı verilerine dayalı bir makine öğrenme algoritması kullanarak kolonoskopi taramalarındaki bulguları tahmin etmeye çalışmışlardır. Yaş, cinsiyet, kırmızı kan hücresi değişkenleri, enflamatuar hücreler ve trombositlere dayanarak, bir ML algoritması olan ColonFlag ile 0 ila 100 arasında bir puan

üretip yüksek riskli hastalar ile normal hastalar arasında modelin ayırım yapıp yapmadığı değerlendirilmiştir. Bu çalışmada İsrail, İngiltere ve Amerika elektronik sağlık kayıtları kullanılarak yaklaşık 20 bin hasta verisi kullanılmıştır. Model doğrulama çalışmasında, kolorektal kanserin saptanması için ROC AUC değeri 0,82 olarak bulunmuştur (Hilsden, et al., 2018).

Stark ve arkadaşları (2019), makine öğrenimi tekniklerini kullanarak gelecekteki kanser riski için öngörücü modeller geliştirdiler. Araştırmalarında, popülasyon, kavram ve bağlam gibi farklı alanlara odaklanarak kapsamlı bir şekilde konuyu ele alan tanınmış bir kapsam inceleme yaklaşımı kullanılmıştır. Meme kanseri riskini tahmin etmek için kullanılan Gail Modeli ve onun bir uygulaması olan Meme Kanseri Risk Tahmin Aracı (BCRAT), yaşa özgü meme kanseri tehlike oranları ve kişisel öykü verilerini kullanır. BCRAT, Cox regresyonları, lojistik regresyonlar ve Gail Modeli gibi basit istatistiksel modellerden yararlanır. Bu çalışma, beş yıllık meme kanseri riskinin BCRAT'tan daha iyi tahmin edilip edilemeyeceğini incelemektedir. Modeller, Prostat, Akciğer, Kolorektal ve Over Kanseri Tarama Denemesi (PLCO) veri seti üzerinde eğitilmiş ve değerlendirilmiştir. Veri seti, 50-78 yaş arası 78,215 kadından oluşmaktadır. Çalışmada altı tip sınıflandırıcı kullanılmıştır: lojistik regresyon, Gauss naif Bayes, karar ağacı, doğrusal diskriminant analizi, destek vektör makinesi ve ileri beslemeli yapay sinir ağı. Lojistik regresyon, doğrusal diskriminant analizi ve sinir ağı modelleri, BCRAT'tan daha güçlü sonuçlar vermiştir (Stark, et al., 2019).

2019 yılında Hart ve arkadaşları akciğer kanseri risk tahmini için yapay sinir ağı modeli geliştirdiler. 1997-2015 Ulusal Sağlık Görüşme Anketi (NHIS) yetişkin verilerini kullanarak: cinsiyet, yaş, VKİ, diyabet, sigara içme durumu, amfizem, astım, ırk, Hispanik etnik kökeni, hipertansiyon, kalp hastalıkları, güçlü egzersiz alışkanlıkları ve inme öyküsü özellikleri ile 648 kanser vakası tespit ettiler. Modelin AUC değeri 0,86 olarak bulunmuştur. Bu çalışma, kişisel sağlık bilgilerine dayalı bir yapay sinir ağı (ANN) kullanımının akciğer kanseri tespiti için yüksek seçicilik ve mütevazı duyarlılık sağladığını ve risk sınıflandırması için uygun maliyetli ve invaziv olmayan bir klinik araç sunduğunu göstermiştir (Hart, et al., 2019).

Taninaga ve arkadaşları, 2019 yılında kapsamlı tıbbi kontrol verilerini kullanarak gelecekteki mide kanseri riskinin tahmin edilmesi için bir makine öğrenimi algoritması geliştirdiler. 2006 ile 2017 yılları arasında Japonya'daki tek bir tesiste gerçekleştirilen birden fazla endoskopi uygulamasına katılan 25.942 kişiden boylamsal ve kapsamlı tıbbi kontrol verileri toplandı. Katılımcılar, 122 aylık süre zarfında mide kanseri tespit edilirse vaka grubu ($y = 1$) olarak, tespit edilmediyse kontrol grubu ($y = 0$) olarak sınıflandırıldı. Sonuçlar, XGBoost algoritmasının lojistik regresyondan daha iyi performans sergilediğini ve en yüksek alanın eğri değerinin (ROC AUC) 0,899 olduğunu gösterdi. Bu sonuçlar, makine öğrenimi algoritmalarının mide kanseri riskinin tahmininde kullanılmasının potansiyel değerini ortaya koymaktadır (Taninaga, et al., 2019).

Cai ve arkadaşları, 2019 yılında Çin'deki yüksek riskli bireyler arasında mide kanseri riskini belirlemek için bir tahmin kuralı geliştirmeyi ve doğrulamayı amaçlayan ülke genelinde çok merkezli bir çalışma gerçekleştirdiler. Çalışmaya, mide kanseri tarama gastroskopisi için hastaneleri ziyaret eden 40-80 yaş arası 14.929 katılımcı dahil edildi. Yeni mide kanseri risk tahmin kuralında, yaş, cinsiyet, PG I/II oranı, G-17 düzeyi, H. Pylori enfeksiyonu, salamura yiyecekler ve kızarmış yiyecekler gibi yedi değişkeni içeren ve 0 ile 25 arasında puanlama sistemi kullanıldı. Tahmin kuralının eğri altındaki alan (ROC AUC) değeri 0,76 olarak bulunmuş ve kalibrasyon için ($p < 0,001$) iyi bir ayırt edicilik gösterdi. Katılımcılar önce bir anket doldurdular, daha sonra PG, G-17 ve anti-H. pylori IgG antikor konsantrasyonlarını ölçmek için açlık kan örnekleri verdiler ve en son gastroskopi işlemine tabi tutuldular. Yeni bir lojistik regresyon yöntemi kullanarak, mevcut özellikleri dikkate alarak tahminler yapıldı ve daha önce belirlenmiş biyobelirteçlere dayalı çalışmalardan da elde edilen sonuçlar kendi modellerine uyguladılar. Geliştirilen lojistik regresyon modelinin ROC AUC değeri, daha önce yapılan çalışmalara göre daha yüksek olan 0,75 olarak bulundu. Bu çalışma sonucunda geliştirilen ve doğrulanan tahmin kuralının, Çin'deki yüksek mide kanseri riski taşıyan bireyleri belirlemekte iyi bir performans sergilediği görülmüştür (Cai , et al., 2019).

Zhu ve arkadaşları 2020 yılında, makine öğreniminin noninvaziv özelliklere dayalı bir çalışmayla mide kanseri teşhisi koymak için Zhejiang İl Halk Hastanesi'nde retrospektif olarak 709 hastayı inceleyerek bir çalışma yaptılar. Çalışmada, yaş, cinsiyet, kan hücre

sayısı, karaciğer fonksiyonu, böbrek fonksiyonu, kan lipitleri, tümör belirteçleri ve patolojik sonuçlar gibi değişkenler analiz edilmiştir. Bu kapsamda, mide kanseri hastalarının klinik verileri, laboratuvar sonuçları ve radyolojik görüntülemeler gibi noninvaziv özellikler değerlendirilmiştir. Çok değişkenli analiz sonucunda, cinsiyet, yaş, nötrofil-lenfosit oranı, hemoglobin, albümin, karsinoembriyonik antijen (CEA), karbonhidrat antijeni 125 (CA125) ve karbonhidrat antijeni 199 (CA199) gibi bağımsız değişkenlerin mide kanseri ile ilişkili olduğu belirlenmiştir. Gradyan artırma karar ağacı (GBDT) yöntemi kullanılarak oluşturulan modelin AUC değeri 0,91 olarak bulunmuştur. Sonuç olarak, çalışma, mide kanseri teşhisinde makine öğrenme algoritmalarının başarılı bir şekilde uygulanabileceğini göstermektedir. Noninvaziv özelliklerin kullanılması, hastaların kanser durumlarının doğru ve hızlı bir şekilde belirlenmesine katkıda bulunarak, daha etkili tedavi stratejileri geliştirmeye yönelik önemli bir adım olarak değerlendirilebilir (SL, et al., 2020).

Lu ve arkadaşları (2020) tarafından gerçekleştirilen retrospektif bir çalışma, makine öğrenimi kullanarak over kanseri tahminlemesini incelemiştir. Çin'deki Soochow Üniversitesi'nden toplanan 349 hastanın verileriyle yapılan bu çalışma, 2011-2017 yılları arasındaki verileri eğitim ve doğrulama verisi olarak kullanmış; 2017-2018 yıllarındaki veriler ise test verisi olarak değerlendirilmiştir. Çalışmada kullanılan özellikler, demografik bilgiler, kan rutin test sonuçları, genel kimya ve tümör belirteçlerini içeren değişkenlerden oluşmaktadır. Bu özelliklerle over kanseri tahminlemesi gerçekleştirilmiştir. Toplamda 49 değişkeni içeren bir veri seti oluşturulmuş olup, bu veri setinde histoloji, demografik bilgiler, kan rutin test sonuçları, genel kimya ve tümör belirteçleri bulunmaktadır. Yapılan analizlerde, Karar Ağacı (CART), Lojistik Regresyon ve ROMA (Risk Ovarian Malignancy Algorithm) adlı bir risk değerlendirme algoritması kullanılmıştır. Sonuçlar, Karar Ağacı algoritmasıyla (CART) elde edilen modelin, ROC AUC 0,88 ile en iyi tahminlemeyi gerçekleştirdiğini göstermiştir (Lu, et al., 2020).

Eunsaem ve arkadaşları 2021 yılında gerçekleştirdikleri bir araştırmada, ülke genelindeki sağlık hizmeti verilerini kullanarak kanser tahmin modelleri geliştirmeye odaklanmıştır. Bu çalışma, makine öğrenimi yöntemlerini kullanarak lojistik regresyon, rastgele orman, Hafif Degrade Yükseltme Makinesi (LGBM-Light Gradient Boosting Machine) ve çok

katmanlı algılayıcı (MLP- Multilayer Perceptron) gibi dört yaygın modelle eğitim sağlamıştır. Ayrıca, tek sınıflı bir gömme sınıflandırıcı (OCEC- Organizational CoEvolutionary algorithm for Classification) adlı derin anomali tespiti temelli bir model oluşturulmuştur. Araştırma, yedi yaygın kanser türü (karaciğer, mide, kolorektal, akciğer, pankreas, meme ve rahim ağzı kanseri) için gelecekteki insidansı tahmin etmeyi amaçlayan makine öğrenimi modelleri geliştirmiştir. Sonuçlar, en iyi performansı gösteren modellerin LGBM ve OCEC olduğunu ortaya koymuştur. LGBM yöntemi ile kanser tahminlerindeki en önemli özellikler: cinsiyet, yaş, alkol tüketimi ve sıklığı, toplam kolesterol, hemoglobin, glikoz, vücut kitle endeksi (VKİ) olmuştur. Bu iki model, diğerlerine kıyasla daha yüksek ROC AUC değerlerine sahiptir. Özellikle OCEC, ROC AUC puanı açısından en iyi sonucu elde etmiş ve normal derin öğrenme yöntemi olan MLP'den daha iyi performans sergilemiştir. Araştırmanın en etkili model yapısı, tek sınıf anomali içgörülerıyla yoğun boyut küçültme yöntemi kullanılarak elde edilmiştir (Lee , et al., 2021).

2021 yılında Li ve arkadaşları geleneksel laboratuvar test verileri kullanarak makine öğrenimi modelleriyle kolorektal kanser tespitinde bulunmaya çalıştılar. 2500 hasta ile yapılan çalışmada LR, RF, K-NN, SVM ve NB makine öğrenimi yöntemleri kullanılmıştır. CRC'yi tanımlamak için güçlü, invaziv olmayan ve uygun maliyetli bir yöntem olarak 0,90 ROC AUC ile en iyi performansı karsinoembriyonik antijen (CEA), hemoglobin (HGB), lipoprotein (a) (Lp (a)) ve yüksek yoğunluklu lipoprotein (HDL) dayalı lojistik regresyon modeli, bulunmuştur (Li, et al., 2021).

İmmünoterapi alan kanser hastalarında mortaliteyi tahmin etmek için 2022 yılında Yougen ve arkadaşları C-reaktif protein, nötrofil-lenfosit oranı albümin (ALB), hemoglobin, beyaz kan hücresi sayısı gibi birçok özelliği kullanarak makine öğrenimi tabanlı bir risk sınıflandırma modeli oluşturdular. Üç kanser türünde (küçük hücreli olmayan akciğer kanseri, mesane geçiş hücreli karsinom ve renal hücreli karsinom) verileri dahil ettiler. Tahmin modellerinde GBM, RF, kement düzenlileştirme ile LR, SVM ve K-NN yöntemleri kullanılarak kanser hastalarında mortalite tahmini ve risk sınıflaması yapılmış ve en iyi yöntemin RF olduğu görülmüştür (Wu, et al., 2023).

2022 yılında Sanam ve arkadaşları hasta verilerinden denetimli makine öğrenmesi yöntemlerini (Rastgele Orman, Degrade Artırma, Destek Vektör Makinesi, Yapay Sinir Ağı ve Çok Katmanlı Algi) kullanarak meme kanseri tahmini için yeni bir makine öğrenimi tabanlı model geliştirdiler. Meme kanseri tahmini için en sık kullanılan veri kümeleri SEER Meme Kanseri (SEERBCD), Coimbra Meme Kanseri (CBC) veri kümesi, Wisconsin (Prognostik) Meme Kanseri (WPBC) veri kümesi, Wisconsin (Diagnostik) Meme Kanseri veri kümesi, Wisconsin Orijinal Meme Kanseri (WOBC) veri kümesi ve Meme Dokusu Veri Kümesi (BTD) kullanılmıştır. CBC veri kümesi rutin kan analizlerini de içermektedir. Çok Katmanlı Algi Modeli, incelenen diğer tüm modellerden daha iyi performans göstermiştir (%99,12 doğrulukla) (Aamir, et al., 2022).

Olga ve arkadaşları 2022 yılında hem tam kan değerleri (CBC) hem de moleküler verileri veya CBC verilerini tek başına kullanarak Miyelodisplastik sendrom varlığını tahmin etmek için RF makine öğrenimi de dahil olmak üzere çeşitli algoritmalar kullanarak parametre tabanlı modeller geliştirdiler. Yaklaşık 1000 kişilik bir örneklem ile yapılan çalışmada sadece CBC değerleri ile modelin ROC AUC değeri %86 gösterirken demografik verilerle ile ROC AUC değeri %93'e yükselmiştir (Pozdnyakova, et al., 2022).

Guo ve arkadaşları 2022 yılında tiroid tümörlü hastalar için geliştirmekte olan bir yardımcı tanı yöntemi sağlamak amacıyla, iyi huylu ve kötü huylu tiroid tümörlerinin tanısal bir modelini oluşturmak amaçlı bir çalışma yaptılar. Bu amaçla, dört farklı makine öğrenme algoritması kullanılmıştır: aşırı gradyan artırma (XGBoost), rastgele orman (RF), hafif gradyan artırma makinesi (LightGBM) ve tahmine dayalı modeller oluşturmak için kullanılan uyarlanabilir güçlendirme (AdaBoost). Çalışmada, 2.042 hasta ve 12 değişken dahil edilmiştir. Hastalar, Chongqing Genel Hastanesinden (Chongqing, Çin) Temmuz 2020'den Eylül 2021'e kadar seçilmiştir. Çalışmada kullanılan değişkenler arasında cinsiyet, yaş, BRAFV600E geni, lenfosit sayısı, nötrofil sayısı, nötrofil/lenfosit oranı (NLR-Neutrophil/Lymphocyte Ratio), trombosit/lenfosit oranı (PLR-Platelet/Lymphocyte Ratio), kırmızı kan hücresi dağılım genişliği (RDW- Red Cell Distribution Width), trombosit sayısı (PLT-Platelet Count), kırmızı kan hücresi dağılım genişliği-varyasyon katsayısı (RDW-CV), alkalın fosfataz (ALP) ve paratiroid hormonu

(PTH) bulunmaktadır. Tek deęişkenli ve LASSO analizi ile belirlenen istatistiksel olarak anlamlı göstergeler şunlardır: cinsiyet, yaş, Lenf#, PLR, RDW ve BRAFV600E. RF, XGBoost, LightGBM ve AdaBoost algoritmalarının elde ettięi ROC AUC deęerleri sırasıyla 0,874; 0,868; 0,861; 0,837 olarak bulunmuştur. Sonuç olarak, bu çalışma, iyi huylu ve kötü huylu tiroid tümörlerini tahmin etmek için güçlü ve umut verici bir araç olabilecek yeni biyobelirteçleri içeren bir model önermektedir (Guo, et al., 2022).

Zhang ve arkadaşları 2022'de, derin öğrenme kullanarak 33 mevcut kan indeksini temel alan mide kanseri tanısı için bir araç geliştirdiler. Toplamda 2951 örnekten elde edilen verilere dayanarak, yaş, cinsiyet ve 58 rutin kan biyokimyasal indeks, tanı modeli oluşturmak için kapsamlı indeksler olarak ele alındı. H2O çerçevesi kullanarak, rutin kan verilerini ve kan biyokimyasal verilerini işleyen yapay sinir ağından yola çıkarak, mide kanseri tanısı için 33 önemli indeksi içeren bir tanımlama modeli ve kullanıcı dostu bir web test terminali kurdular. Test seti için modelin ROC AUC deęeri 0,9152 olarak bulunmuştur. Geleneksel tanı yöntemlerine kıyasla, bu çalışmada oluşturulan tanı modeli daha yüksek doğruluk, noninvazivlik ve uygun maliyetli tespit avantajlarına sahip olduęu sonucuna varılmıştır (Zhang, et al., 2022).

Read ve arkadaşları 2023 yılında Boylamsal Elektronik Sağlık Kaydı Verileri Kullanılarak Gastrointestinal (GI) Sistem Kanserlerinin Tahmini için retrospektif bir kohortta, elektronik sağlık kaydında rutin olarak bulunan verileri kullanarak lüminal gastrointestinal sistem kanserlerinin (özofagus, mide, ince bağırsak, kolorektal, anal) öngörülmesi için modeller geliştirdiler. Kohort, 1025 gastrointestinal sistem kanseri olan 148.158 kişiyi içeriyordu. Bu çalışmada lojistik regresyon ve rastgele orman makine öğrenimi yöntemlerini kullanarak laboratuvar çalışmaları ve hasta özelliklerini kullanarak lümen GI yolu kanserlerini tahmin etmek için modeller geliştirmek amaçlanmıştır. Her bir tahmin teknięi için, altı aylık, bir yıllık, üç yıllık ve beş yıllık lüminal gastrointestinal sistem kanserlerinin teşhisi için bir tahmin modeli geliştirilmiştir. Lojistik Regresyon (Tek Zaman Noktası): 0,703 (5 yıl), Lojistik Regresyon (boylamsal): 0,735 (3 yıl), Rastgele Orman (boylamsal): 0,750 (3 yıl). Bu puanlara göre, üç yıllık zaman noktasında Rastgele Orman (boylamsal) modelinin en yüksek ROC AUC puanına sahip olduęu görülüyor:

0,750. Bu sonuç, Rastgele Orman (boylamsal) modelinin bu karşılaştırmada en iyi performans gösteren model olduğunu düşündürmektedir (Read, et al., 2023).

Sebastien ve ekibi, 2021 yılında, 55-69 yaş arası evre küçük hücreli dışı akciğer kanseri hastalarında (toplam 298 hasta) immün kontrol noktası inhibisyonuna yanıtın belirleyicilerini araştırmışlardır. Çalışmada, tedavi öncesi alınabilen kan örneklerindeki çeşitli basit belirteçlerin yanıtla ilişkisini değerlendirmişlerdir. İlk olarak, klasik istatistiksel yöntemler kullanarak analiz yapmışlar ve daha sonra, bireysel tahminler için bir makine öğrenimi algoritması geliştirmişlerdir. Bu yöntemle %69 doğruluk oranı elde etmişlerdir. Hemoglobin düzeyleri ve performans durumu, yanıtın en güçlü belirleyicileri olarak bulunmuştur. Ayrıca, nötrofil-lenfosit oranının da sonuçlarla ilişkili olduğu tespit edilmiştir. Sekiz farklı makine öğrenimi modelinin karşılaştırıldığı bu çalışmada, en iyi performans gösteren modelin, Rastgele orman modeli olup ROC AUC değeri 0,74 olarak bulunmuştur. Klinik ve biyolojik verilerin birleştirilmesi, hastalık kontrol oranını bireysel düzeyde doğru bir şekilde tahmin etmeye olanak sağlamıştır (Benzekry, et al., 2021).

3. GEREÇ ve YÖNTEM

3.1. Gereç

3.1.1. Veri Toplama ve Hazırlama Süreçleri

Bu çalışma, evrenin tamamı üzerinde yapılmıştır. Çalışma evreni, 1 Ocak 2018 ile 31 Aralık 2022 tarihleri arasında kanser tanısı almış 18 yaş üstü Türkiye Cumhuriyeti vatandaşları ile Türkiye’de oturma iznine sahip kişilerin e-Nabız sisteminden alınan verilerden oluşmaktadır. Veri setinde yer alan bireylerin yaşları 18 ile 123 arasında değişmektedir.

2015 yılında e-Nabız sistemi açılmış olup, 2016 yılında veri gönderim kalitesi ve hastanelerin entegrasyon sayısı daha sağlıklı bir seviyeye ulaşmıştır. Çalışmamıza alınacak hastaların seçiminde kanser ATC kodu ve/veya ICD tanı kodu bulunması kriterleri kullanılmıştır.

Deney Grubu

Bu gruptaki hastalar, 2018 ve 2022 yılları arasında ilk defa kanser tanısı almış bireylerdir. Bu hastaların patoloji raporlarında malign hastalıklara ait ICD-O-3 (International Classification of Disease for Oncology, version 3) kodu bulunmaktadır.

Kontrol Grubu

Kontrol grubu, daha önce hiç kanser teşhisi almamış (tüm zamanlarda) hastalardan oluşmaktadır. Bu gruba, 2016-2017 yıllarında kan tetkik verileri olan kişiler alınmıştır.

Veri:

Çalışmamızda, başlangıç aşamasında hasta ve kontrollerin 2016-2017 yıllarındaki tetkiklerine ait idhash, yas, dogum_yili, cinsiyet, islem_kodu_n, ALT, AST, CRP, EOS#, EOS%, Eritrosit, Glukoz, HCT, HGB, Keton, Kreatinin, LYM#, LYM%, MCH, MCHC, MCV, MPV, PCT, PDW, PLT, Potasyum, RBC, RDW, TSH, WBC, ure, durum (kanser

değil-0 / kanser-1) verileri gelmiştir. Bu verilerde yer yer eksikler ve hatalar bulunmaktadır. Bundan dolayı verilerin tümü analizlerde kullanılamamıştır. Verilerle ilgili olarak ilerleyen bölümlerde detaylı bilgi verilecektir.

Veri kaynağı olarak e-Nabız sistemi kullanılmıştır. Veriler, T.C. Sağlık Bakanlığı tarafından kimliksizleştirilmiş (de-identifiye edilmiş) olarak tarafımıza verilmiştir. Bu süreçte, AB General Data Protection Regulation (GDPR) 2016/679 içeriğinde tanımlanan “pseudonymisation” prosedürü uygulanmıştır. Hastanın kim olduğuna tıbbi nedenlerle ulaşmak gerektiğinde kullanılmak üzere, “hash function” ile bir geri tanımlama kodu oluşturularak hastanın diğer verileri ile ilişkilendirilmiştir. Daha sonra, hastaların kimliğinin belirlenmesine yol açabilecek tüm bilgiler (ad, adres, telefon ve faks numaraları, cihaz tanımlayıcıları ve seri numaraları, e-posta adresleri, sosyal güvenlik numaraları, vatandaşlık numarası, tıbbi kayıt numaraları, parmak ve ses kaydı dahil biyometrik tanımlayıcılar, geri tanımlama kodu hariç, her türlü benzersiz tanımlayıcı numara, özellik veya kod) silinmiştir.

Türkiye Cumhuriyeti Sağlık Bakanlığı'nda (TCSB) tüm veriler, raporlama ve analiz amacıyla Hadoop tabanlı bir büyük veri platformu olan Cloudera (Cloudera CDH, sürüm 6.3.2) sistemine sürekli olarak aktarılmaktadır. Veriler, TurboBoard (sürüm 2020.07, E-Kalite Ltd., Ankara) aracılığıyla yönetilmektedir. Büyük veri platformunda, canlı kaynaklardan gelen verilerin işlenmesi için Kudu tablo tipi kullanılırken, analitik sorgulamalar ve rapor verileri için Parquet depolama formatı tercih edilmektedir. Veri kaydı ve sorgulamalar, Apache Impala (sürüm 3.2.0) sorgulama motoru aracılığıyla gerçekleştirilmektedir.

Orijinal verinin oluşturulması:

Hazırlanan verilerden oluşturulan raporlar, TurboBoard (v2020.07, E-Kalite Ltd., Ankara) platformunda geliştirilmiştir. Bu platform, Python ve JavaScript teknolojileri kullanılarak geliştirilen bir iş zekâsı platformudur. TurboBoard, farklı veritabanı kaynaklarına (Impala, MS-SQL, PostgreSQL, MySQL, Oracle, Vertica) bağlanma ve bu farklı kaynaklardan alınan görsellerin tek bir ekranda birleştirilerek gösterilmesini sağlar.

Verilerin temizlenmesi ve analizlere uygun tablo formatlarının hazırlanması için Impala SQL kullanılmıştır. Gösterimde yapılacak filtreleme ve/veya ayıklama işlemleri ise Turboard'daki filtreleme özellikleri ile gerçekleştirilmiştir. Turboard sistemi, hastaların T.C. kimlik numaralarını bir hash fonksiyonu ile kendine özgü bir id numarasına çevirerek hastaların kimliğini gizli tutmaktadır.

Verilerin analizi ve düzenlenmesi, tahmin edici modellerin geliştirilmesi ve belirlenen algoritmaların uygulanması amacıyla Python programlama dili (sürüm 3.10.9) kullanılmıştır. Python, makine öğrenimi projelerinde yaygın olarak tercih edilen bir yazılım aracıdır. Çalışmanın gerçekleştirilmesi için, kullanıcı dostu arayüzü ve geniş kütüphane desteği ile bilinen Anaconda platformu ve Jupyter Lab tercih edilmiştir. Anaconda, scikit-learn, matplotlib, numpy, pandas ve seaborn gibi gerekli kütüphanelerin kolayca yönetilmesini sağlayan bir paket yönetim sistemi sunar. Jupyter Lab ise, bu kütüphanelerin entegrasyonunu destekleyen etkileşimli bir geliştirme ortamıdır. Bu özellikleri sayesinde, algoritmaların çalıştırılması ve verilerin görselleştirilmesi gibi ihtiyaç duyulan tüm işlemler kolaylıkla gerçekleştirilmiştir.

Bunun yanı sıra, istatistiksel analizler için IBM SPSS Statistics 22, jamovi 2.4.4 ve Microsoft 365 Excel yazılımları kullanılmıştır. Bu araçlar, verilerin detaylı analizi, istatistiksel işlemler ve görselleştirmeler için tercih edilmiştir.

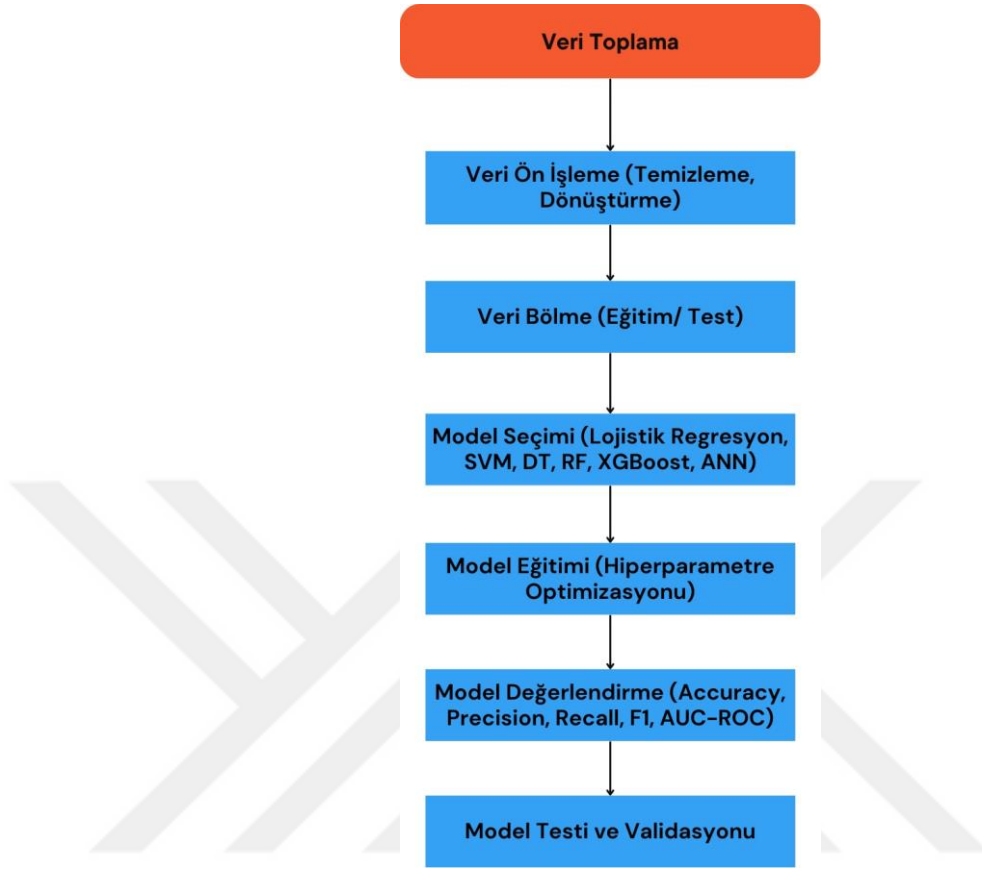
3.2. Yöntem

Bu çalışmada, kanser sınıflandırması için makine öğrenimi modeli geliştirme sürecinde belirli adımlar izlenmiştir (Şekil 3.1.). Öncelikle, 1 Ocak 2018 ile 31 Aralık 2022 tarihleri arasında kanser tanısı almış 18 yaş üstü bireylerin e-Nabız sisteminden alınan verileri kullanılmıştır. Toplanan veriler, eksik veya hatalı değerlerin düzeltilmesi ve gerekli dönüşümlerin uygulanmasıyla ön işleme tabi tutulmuştur. Ardından, veriler eğitim ve test setlerine bölünerek, lojistik regresyon, destek vektör makineleri (SVM), XGBoost ve yapay sinir ağları (ANN) gibi farklı makine öğrenmesi algoritmaları kullanılarak modeller oluşturulmuştur. Bu modeller, hiperparametre optimizasyon teknikleri ile eğitilmiş ve performansları Doğruluk, Duyarlılık, Kesinlik, F1 skoru, MCC ve AUC-ROC gibi metriklerle değerlendirilmiştir. Son olarak, modeller test seti üzerinde doğrulanarak

genelleme yetenekleri test edilmiştir. Bu adımlar izlenerek, kanser sınıflandırmasında kullanılabilecek etkili ve güvenilir bir makine öğrenimi modeli geliştirilmiştir.

Model geliştirme sürecinde aşağıdaki Python kütüphaneleri kullanılmıştır:

- Scikit-learn (Pedregosa et al., 2011)
- TensorFlow (Abadi et al., 2016)
- Matplotlib (Hunter, 2007)
- Seaborn (Waskom, 2021)
- Pandas (McKinney, 2012)
- NumPy (McKinney, 2012)
- Plotly (Plotly Technologies Inc., 2015)
- Streamlit (Streamlit Inc. (2021))



Şekil 3.1. Makine Öğrenimi Süreci

3.2.1. Veri Ön İşleme (Temizliği ve Filtreleme)

Bu çalışmada, büyük veri platformundaki milyonlarca satırı bulunan tablolarda yer alan kayıtlar incelenmiştir. Belirlenen özel kriterlere uyan kayıtların seçilmesi sürecinde bazı zorluklarla karşılaşmıştır. Özellikle, gerçekçi olmayan değerler ve analiz sürecinde doğru sonuçların elde edilmesi için veri temizliği çalışmaları yapılmıştır. Veri temizleme süreci, sorgulamalar sırasında yüksek performansın sağlanması ve güvenilir sonuçlar elde etmek amacıyla gerçekleştirilmiştir.

Makine öğrenmesi modellerini eğitmeden önce veri setindeki değişkenler ölçeklendirilmiştir. Bu işlem, değişkenlerin değerlerini ortak bir ölçeğe getirerek, modelin daha etkin ve doğru çalışmasını sağlamak için uygulanmıştır. Özellikle, veri standartlaştırma tekniği kullanılarak, değişkenlerin ortalama değeri sıfır ve varyansı bir olacak şekilde Standard Scaler kullanılarak dönüştürülmüştür. Bu ölçeklendirme işlemi,

algoritmaların eğitim sürecini hızlandırır ve daha kararlı hale getirir, böylece modelin genel performansını iyileştirir (Cao, 2016).

1. **Veri Formatı Dönüşümü:** Veri seti, Sağlık Bakanlığı'ndan alınan 17 değişken ve 16.854.639 satır veriden oluşmaktadır. İlk olarak, veriler arasındaki ayırıcı olarak kullanılan 'I' karakteri virgülle değiştirilerek CSV formatına dönüştürülmüştür.
2. **Sütun Yapılandırma:** Kan tetkikleri, tek sütunda birden fazla idhash ile geldiği için, bu veri setine önce kan tetkiklerinin her birinin ayrı bir sütunda yer alacağı şekilde pivot dönüşümü uygulanmıştır. Her bir kan tetkik değişkeni ayrı bir sütun olacak şekilde düzenlenmiş ve ilgili değerler bu sütunlara yerleştirilmiştir. Bu işlemler sonucunda veri seti, 43 değişken ve 6.439.825 satırdan oluşan bir veri setine dönüşmüştür.
3. **Benzersiz Kayıt Düzenlemesi:** Çoklu idhash ve yıl bilgileri olan veriler, idhash ve yıl değerleri birlikte benzersiz olacak şekilde düzenlenerek, sadece dolu kan değerleri dikkate alınmıştır.
4. **Veri Birleştirme:** 2016 yılı ilk tanı alınan kan değerleri baz alınarak, boş veriler 2017 yılındaki dolu değerlerle doldurulmuştur. Bu şekilde tek satırda tek hasta yani benzersiz idhash değerlerinin oluşturulması sağlanmıştır.

26 Kan tetkik, üç demografik ve bir bağımlı değişken ile veri seti ön işlemeye devam edilmiştir.

ALT, AST, CRP, EOS#, EOS%, Eritrosit, Glukoz, HCT, HGB, Keton, Kreatinin, LYM#, LYM%, MCH, MCHC, MCV, MPV, PCT, PDW, PLT, Potasyum, RBC, RDW, TSH, WBC, ure, yas, dogum_yili, cinsiyet, durum.

5. Değer Sınıflandırma ve Dönüşümü:

- Eritrosit, keton ve glukoz değerleri, uygun ve hatalı değerler olarak sınıflandırılmıştır.

- Diğer kan değerleri, verinin yapısı ve dağılımı dikkatlice incelenmiş ve aykırı değerler silinerek, uygunsuz değerler (object değerler floata dönüştürülerek) düzenlenerek analize hazır hale getirilmiştir.

6. Türetilen Veriler:

Boş olan PLT değerleri için $PLT = PCT * 10000 / MPV$ formülü kullanılarak 17,206 yeni veri türetilmiştir.

- Boş olan MPV değerleri için $MPV = PCT * 10000 / PLT$ formülü kullanılarak 51,599 yeni veri türetilmiştir.

- Boş olan EOS% değerleri için $EOS\% = EOS\# / WBC * 100$ formülü kullanılarak 42,698 yeni veri türetilmiştir.

- Boş olan LYM% değerleri için $LYM\% = LYM\# / WBC * 100$ formülü kullanılarak 68,246 yeni veri türetilmiştir.

- Boş olan RBC değerleri için $RBC = HCT * 100 / MCV$ formülü kullanılarak 50,439 yeni veri türetilmiştir.

- Boş olan MCV değerleri için $MCV = HCT * 100 / RBC$ formülü kullanılarak 36,810 yeni veri türetilmiştir.

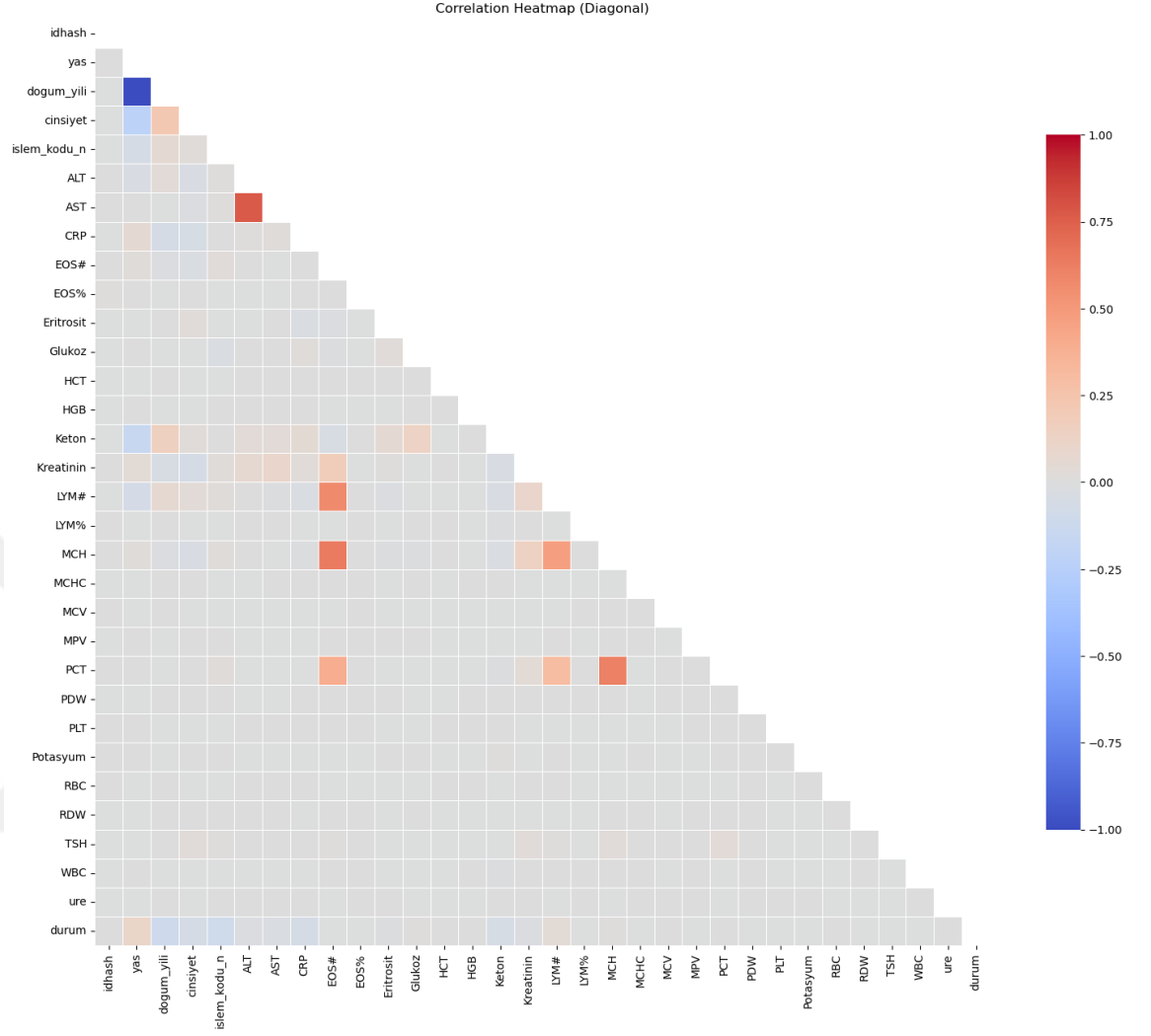
- Hemoglobin (HGB) Türetme:

$HGB = (MCH * RBC / 100)$ formülü kullanılarak 21,737 yeni HGB verisi türetildi.

$HGB = (MCHC * HCT / 100)$ formülü kullanılarak 13,848 yeni HGB verisi türetildi. Toplamda 35,584 yeni HGB verisi türetildi.

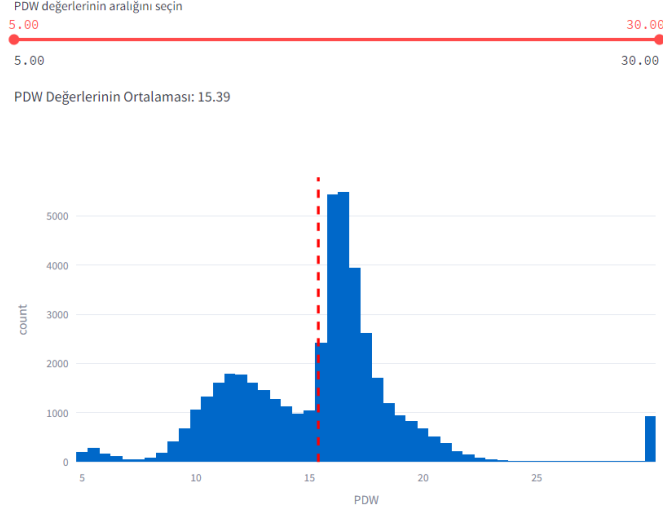
7. Temizlik ve Filtreleme:

- Veri setindeki değişkenlerden boş değerleri %30 üstü olanlar veri setinden çıkarılmıştır (Potasyum, RDW, Eritrosit, TSH, MCH) (Dong, 2013).
- Glukoz, Keton, Kreatinin, ure veri kalitesi (kan ve idrar sonuçlarının karışması) ve üre, LYM%, LYM#, EOS#, PCT, CRP uygun olmayan veri dağılımı nedeniyle veri setinden çıkarılmıştır.
- RBC, HCT, MCHC, WBC, MCV ve HGB değerleri arasında multikolinearite olduğu için bu çalışmada sadece HGB kullanılmıştır. Literatürde, HGB hesaplamasında diğer kan parametrelerinin kullanıldığına dair pek çok çalışma bulunmaktadır. Örneğin, Sarma (1990), HGB düzeylerinin diğer parametreler kullanılarak doğru bir şekilde tahmin edilebileceğini göstermiştir. Bu hesaplama yöntemi, laboratuvar analizlerinde gereksiz tekrarı önleyerek veri setlerinin daha kompakt ve yönetilebilir hale getirilmesine katkı sağlamaktadır. Bu çalışmaya dayanarak, veri setimizde ilgili kan parametrelerinin çıkarılması, modelin performansını olumsuz etkilememekte, aksine veri yönetimini kolaylaştırmaktadır. Bu nedenle, bu tetkiklerin veri setinden çıkarılması uygun bulunmuştur (James G. W., 2013).
- Yüksek korelasyon nedeniyle modelde çoklu bağlantı probleminin önüne geçmek için AST ve ALT arasında bir seçim yapılması gerekmiştir (Şekil 3.2). Literatürde, yüksek korelasyonlu değişkenlerin modele dahil edilmesinin çoklu bağlantı problemlerine yol açtığı ve modelin tahmin gücünü olumsuz etkilediği belirtilmektedir (Dormann, 2013). Bu durumu önlemek için, veri setinde veri kalitesi ve dağılımı daha iyi olan ALT modele dahil edilerek, AST çıkarılmıştır.



Şekil 3.2 Veri Setindeki Değişkenler Arasındaki Korelasyon Matrisi

- PDW değerlerindeki ölçüm hataları, modelin eğitim sürecinde yanlış desenleri öğrenmesine neden olur ve bu da modellerin gerçek dünya verileri üzerinde doğru bir genelleme yapmasını engeller (Şekil 3.3). Yanlış girilmiş ölçüm birimleri içeren veriler, modelin test aşamasında beklenen doğruluk düzeyini sunamamasına yol açar; bu verilerin veri setinden kaldırılması, modelin tahmin doğruluğunu önemli ölçüde iyileştirir. Bu durumu düzeltmek için, ölçüm hatalarına maruz kalmış veri noktalarının elimine edilmesi, modellerin güvenilirliğini ve performansını artırmak için zorunludur (Paullada, et al., 2021). Bundan dolayı PDW değişkeni veri setinden çıkarılmıştır.



Şekil 3.3. Veri Setindeki PDW Veri Dağılımı

- Yaş: Çalışmaya alınacak kanser ve kanser olmayan hasta kayıtlarının yaşları, ilk tetkik kayıt yılından hastanın doğum yılı çıkarılarak elde edilmiştir ve yaş değişkeni oluşturulup yeni bir veri oluşturulmuştur. Örneğin 2017 yılında kayıtlı bir hastanın yaşı hesaplanırken doğum yılı 1987 ise $2017-1987 = 30$ yaşında olarak hesaplanıp tablolara bu şekilde kaydedilmiştir.
- İl veri setinde, il bilgisi olmayan 83 kişi veri setinden çıkarılmış ve kan değerleri veri seti ile il veri seti birleştirilmiştir.

İl: Veri setimizde bulunan hasta il bilgileri, Türkiye İstatistik Kurumu'nun nüfus istatistiklerinde yer alan illerin bölge bilgi kodlarına (<https://nip.tuik.gov.tr/Home/GelenekselNufusSayimlari>) göre etiketlenmiş (1-13) ve veri setine bölge kodları ile dahil edilmiştir.

ICD-O-3

Patoloji raporlarında yer alan ICD-O-3 kodları, 31 başlık altında sınıflandırılmıştır. Bu süreçte Canadian Cancer Registry Classification

(<https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=182622&adm=0&dis=0>) ve National Cancer Institute Rare Cancer Classification (<https://seer.cancer.gov/seerstat/variables/seer/raresiterecode>) dikkate alınmıştır.

8. ICD Kodları ile Çalışma:

1936 adet ICD10 kodundan, 1000'den fazla kaydı olan 707 koda lojistik regresyon yapıldı. p değeri 0.01'den büyük olan 457 kod kullanılmaya uygun bulundu. Ayrıca ICD10 kodlarından, doluluk oranı %0.5'ten az olanların analiz üzerindeki etkisi çok önemsizdir (Cowling, 2020). Bundan dolayı %0.5'ten az olan kodlar veri setinden çıkarıldı. 137 ICD10 kodu ile analizlere devam edildi.

- Orijinal veride ICD10 kodları sütununda hastanın 2016-2017 yıllarında toplam kaç kez bu tanıyı aldığı verisi bulunmaktadır. ICD10 kodları için yapılan lojistik regresyon analizlerinde hem normal hem de binary modeller uygulanmıştır. Özellikle, ziyaret sayılarıyla yapılan lojistik regresyon analizlerinde elde edilen Cox-Snell R^2 değeri 0,262 ve Nagelkerke R^2 değeri 0,382 olarak hesaplanmıştır. Diğer yandan, veriler 0-1 (binary) formatına dönüştürüldüğünde Cox-Snell R^2 değeri 0,263 ve Nagelkerke R^2 değeri 0,383 olarak ölçülmüştür. Bu sonuçlar, ziyaret sayılarını analizde kullanmanın anlamlı olmadığını göstermektedir. Dolayısıyla ICD kodları için binary formatta devam edilmesine karar verilmiştir.

Klasik analizler

Tüm değişkenler için kanser olup olmama durumu arasındaki farklar değerlendirilmiştir. Bu analizler, kanser tanısı alan ve almayan gruplar arasındaki istatistiksel farkları belirlemek amacıyla yapılmıştır. Önce grupların Q-Q grafiği ile normal dağılım gösterip göstermediği incelenmiştir. Normal dağılım gösteren sayısal değişkenler için bağımsız örneklem t-testi kullanılmış, normal dağılım göstermeyen değişkenler için ise Mann-Whitney U testi uygulanmıştır. Kategorik veriler ise ki-kare testi ile değerlendirilmiştir.

Her hipotez testi için etki büyüklüğü hesaplanmış, t-testi dışındakiler de için Cohen's d değerine dönüştürülmüştür (https://www.psychometrica.de/effect_size).

Kanser olup olmama durumunu sınıflandırmak için bağımsız değişkenlerle lojistik regresyon analizi yapılmış ve odds oranları ile güven aralıkları değerlendirilmiştir. Lojistik regresyon modelimizi değerlendirmeden önce, bağımsız değişkenler arasındaki potansiyel çoklu doğrusal bağlantı sorunlarını tespit etmek için Varyans Şişme Faktörü (VIF) analizi gerçekleştirilmiştir. Yapılan analiz sonucunda, modelimizde yer alan değişkenlerin VIF değerlerinin hepsi 1.54'ün altında bulunmuştur.

Makina öğrenmesi modelleri, sekiz temel değişken (yas, cinsiyet, il_bolge, ALT, EOS%, HGB, MPV, PLT) ve 137 ICD10 (Dermatofitoz (B35), Kandidiyaz (B37), Vitamin B12 eksikliği anemisi (D51) vb.) kod değişkeni kullanılarak eğitime ve test edilmeye başlanmıştır. Çalışmada toplam 413812 kişinin verisi kullanılmıştır.

3.2.2. Veri Bölme (Eğitimi ve Test)

Bu çalışmada, elde edilen veriler eğitim ve test setlerine bölünmüştür. Eğitim seti, modelin öğrenmesi için kullanılırken, test seti modelin performansını değerlendirmek amacıyla kullanılmıştır. Verilerin %80'i eğitim için, %20'si ise test için ayrılmıştır.

3.2.1.3 Model Seçimi

Çalışmada çeşitli makine öğrenimi algoritmaları kullanılmıştır. Bu algoritmalar şunlardır: Lojistik Regresyon, Destek Vektör Makineleri (SVM), Karar Ağaçları (DT) Rastgele Ormanlar (RF), XGBoost, Yapay Sinir Ağları (ANN).

3.2.4. Model Eğitimi (Hiperparametre Optimizasyonu)

Seçilen modellerin her biri için hiperparametre optimizasyonu yapılmıştır. Hiperparametreler, modelin performansını doğrudan etkileyen ve önceden belirlenmesi gereken parametrelerdir. Bu optimizasyon işlemi, Grid Search ve Random Search yöntemleri kullanılarak gerçekleştirilmiştir. Böylece, her model için en uygun hiperparametre değerleri belirlenmiştir. Ayrıca, model validasyonu çapraz doğrulama

(cross-validation) yöntemleri ile yapılmıştır. Bu yöntem, modelin performansının tutarlı olup olmadığını ve overfitting'in önlenip önlenmediğini kontrol etmek için kullanılmıştır.

3.2.5. Model Değerlendirme

Model performansı çeşitli metriklerle değerlendirilmiştir: Doğruluk, Duyarlılık, Kesinlik, F1 Score, MCC, AUC-ROC.

3.2.6. Model Testi ve Validasyonu

Eğitim süreci tamamlandıktan sonra, modeller test veri seti üzerinde değerlendirilmiştir. Model test işlemi, modelin genel performansını ve yeni verilere karşı ne kadar iyi genelleme yapabildiğini ölçmek için kullanılmıştır.

4. BULGULAR

e-Nabız'dan alınan ve ön işleme yapılmış 2018-2022 yıllarına ait 413812 satır veri setine göre, analizde kullanılacak kan değerlerine sahip hastaların Türkiye'de görülen 31 farklı kanser türüne ilişkin dağılımı Tablo 4.1.'de detaylandırılmıştır. Meme kanseri %12.28 ile örneklemdaki en yaygın kanser türüdür, ardından %12.16 ile epitelial deri kanseri gelmektedir. Akciğer kanseri, %9.94 ile üçüncü en yaygın kanser türüdür. Kolon ve rektum kanserleri (%7.21) ile mide kanseri (%6.92), gastrointestinal kanserler arasında önemli bir yer tutmaktadır. Tiroid kanseri (%6.65) ve prostat kanseri (%6.62) oranları yakın olup, bu kanser türleri için erken tanı yöntemlerinin önemi büyüktür. Mezotelyoma, kemik ve eklem kanserleri, gırtlak kanseri ve özofagus gibi daha az yaygın kanser türleri, toplam kanser vakalarının küçük bir yüzdesini oluşturmaktadır. Endometrium, penis kanseri ve timus bezi kanseri gibi nadir kanser türleri, %0.01 gibi çok düşük prevalans oranlarıyla dikkat çekmektedir.

Tablo 4.1. Veri Seti Kanser Türlerinin Dağılımı (e-nabız 2018-2022)

Kanser Türü	%n
Meme Kanseri	13866 (%12.28)
Epitelial deri	13739 (%12.16)
Akciğer Kanseri	11222 (%9.94)
Kolon ve rektum	8142 (%7.21)
Mide	7813 (%6.92)
Tiroid	7511 (%6.65)
Prostat	7475 (%6.62)
Diğer iyi tanımlanmamış	6061 (%5.37)
Myelom	5491 (%4.86)
Mesane	5090 (%4.51)
Korpus uteri	4705 (%4.17)
Lenfoma	4024 (%3.56)
Karaciğer	2703 (%2.39)
Böbrek Kanseri	2452 (%2.17)
Beyin	1984 (%1.76)
Serviks uteri	1512 (%1.34)
Ağız Kanseri	1490 (%1.32)
Pankreas	1484 (%1.31)
Yumuşak doku	1358 (%1.20)
Gırtlak Kanseri	1039 (%0.92)
Mezotelyoma	1031 (%0.91)

Kemik ve eklem	846 (%0.75)
Over	629 (%0.56)
Özofagus	536 (%0.47)
Testis	284 (%0.25)
Diğer non-epitelial	208 (%0.18)
Göz	143 (%0.13)
Hodgkin	40 (%0.04)
Timus bezi kanseri	39 (%0.03)
Endometrium	13 (%0.01)
Penis Kanseri	10 (%0.01)
Toplam	112940 (%100)

Tablo 4.2.'de, hastalıkların deney ve kontrol gruplarındaki dağılımı ve ilgili istatistiksel analizlerin sonuçları özetlenmiştir. Lojistik regresyon analizinde $p < 0.001$ değerine sahip olmayan değişkenler ve kıkare testi sonuçlarından elde edilen etki büyüklüğü (effect size) 0.2'den küçük olan özellikler çıkarılarak bu tablo elde edilmiştir. Asıl büyük tablo ek (Ekler 2) olarak sunulmuştur. Bu seçim, istatistiksel olarak anlamlı ve klinik olarak önemli farkları vurgulamak amacıyla yapılmıştır.

Tablo 4.2. Kontrol ve Deney Gruplarının Komorbidite, Demografik ve Klinik Özelliklerine Göre Özet İstatistiksel Analizleri. EB: Etki Büyüklüğü, GA: Güven Aralığı. Etki büyüklükleri Cohen d'sine dönüştürülmüştür. *Ayrıntılı bölge istatistikleri Tablo 4.3'te gösterilmektedir.

						Lojistik Regresyon		
Özellik	Hastalık	Kontrol (%n)	Deney (%n)	EB	p	p	Odds Oranı	%95 GA
Demir eksikliği anemisi	Var	40897 (%26.9)	9510 (%16.6)	0.22	<0.001	<0.001	0.879	0.853 - 0.907
Anemi, diğer	Var	60538(%39.8)	13917 (%24.3)	0.29	<0.001	<0.001	0.739	0.719 - 0.759
Kronik iskemik kalp hastalığı	Var	66718 (%43.9)	18078 (%31.5)	0.23	<0.001	<0.001	0.915	0.886 - 0.944
Atrial fibrilasyon ve flutter	Var	23082 (%15.2)	3459 (%6)	0.25	<0.001	<0.001	0.623	0.594 - 0.653
Kalp yetmezliği	Var	33847 (%22.2)	4372 (%7.6)	0.34	<0.001	<0.001	0.583	0.558 - 0.608

Pnömoni, tanımlanmamış organizmalar	Var	46232 (%30.4)	9353 (%16.3)	0.29	<0.001	<0.001	0.666	0.635 - 0.699
Kronik obstrüktif akciğer hastalığı diğer	Var	40467 (%26.6)	9865 (%17.2)	0.2	<0.001	<0.001	0.987	0.961 - 1.014
Solunum yolu yetmezliği, başka yerde sınıflanmamış	Var	10055 (%6.6)	653 (%1.1)	0.22	<0.001	<0.001	0.95	0.920 - 0.981
Akut böbrek yetmezliği	Var	13514 (%8.9)	745 (%1.3)	0.27	<0.001	<0.001	0.363	0.334 - 0.395
Kronik böbrek yetmezliği	Var	25273 (%16.6)	2333 (%4.1)	0.34	<0.001	<0.001	0.325	0.309 - 0.343
Böbrek yetmezlikleri, tanımlanmamış	Var	10843 (%7.1)	935 (%1.6)	0.21	<0.001	<0.001	0.563	0.520 - 0.609
Sistit	Var	43891 (%28.8)	10327 (%18)	0.22	<0.001	<0.001	0.843	0.818 - 0.868
Üriner sistemin diğer bozuklukları	Var	74040 (%48.7)	17496 (%30.5)	0.33	<0.001	<0.001	0.738	0.719 - 0.758
Kalp atım anormallikleri	Var	30869 (%20.3)	6815 (%11.9)	0.2	<0.001	<0.001	0.88	0.850 - 0.911
Solunum anormallikleri	Var	42545 (%28)	7219 (%12.6)	0.33	<0.001	<0.001	0.684	0.661 - 0.708
Boğaz ve göğüs ağrı	Var	77480 (%50.9)	18680 (%32.6)	0.33	<0.001	<0.001	0.738	0.718 - 0.757
Abdominal ve pelvik ağrı	Var	87738 (%57.7)	19399 (%33.8)	0.44	<0.001	<0.001	0.64	0.624 - 0.656
Bulantı ve kusma	Var	51547 (%33.9)	9456 (%16.5)	0.35	<0.001	<0.001	0.718	0.697 - 0.740
Sebebi bilinmeyen ateş	Var	24852 (%16.3)	3565 (%6.2)	0.27	<0.001	<0.001	0.613	0.588 - 0.639
Ağrı, başka yerde sınıflanmamış	Var	67219 (%44.2)	19091 (%33.3)	0.2	<0.001	<0.001	0.833	0.813 - 0.854
Yakınma veya bilinen teşhisi olmayan kişilerin genel muayene ve incelemesi	Var	129222 (%84.9)	40118 (%70)	0.34	<0.001	<0.001	0.651	0.633 - 0.669
Gebelik durumu	Var	17623 (%11.6)	1366 (%2.4)	0.29	<0.001	<0.001	0.316	0.293 - 0.341
Normal gebeliğin gözlemi	Var	11297 (%7.4)	803 (%1.4)	0.23	<0.001	<0.001	0.589	0.537 - 0.647
Kalp ve damar implant ve greftleri	Var	14240 (%9.4)	1894 (%3.3)	0.2	<0.001	<0.001	0.392	0.37 - 0.416

	Kontrol		Deney				Lojistik Regresyon		
Özellik	(%n)		(%n)		EB	p	p	Odds Oranı	%95 GA
Bölge*	152139 (%72.63)		57325 (%27.37)		0.23	<0.001	-	-	-
Cinsiyet	152140 (%72.63)		57339 (%27.37)		0.12	<0.001	0.075	0.971	0.939-1.003
Kadın	95635 (%62.9)		32069 (%55.9)						
Erkek	56504 (%37.1)		25256 (%44.1)						
	(%n)	Median (%25-75)	(%n)	Median (%25-75)	EB	p	p	Odds Oranı	%95 GA
Yaş	152139 (%72.63)	59 (42-71)	57325 (%27.37)	62 (52-70)	0.16	<0.001	<0.001	1.031	1.029-1.032
ALT	152139 (%72.63)	17 (12-25)	57325 (%27.37)	18 (13-24.8)	0.09	<0.001	<0.001	0.998	0.997-0.998
EOS	152139 (%72.63)	2.25 (1.5-3.4)	57325 (%27.37)	2.25 (1.5-3.4)	0.00	0.718	0.778	0.999	0.992-1.006
PLT	152139 (%72.63)	247 (201-300)	57325 (%27.37)	247 (201-299)	0.00	0.793	0.869	0.869	1.009-1.021
	(%n)	Mean±Std	(%n)	Mean±Std	EB		p	Odds Oranı	%95 GA
HGB	152139 (%72.63)	12.99 ± 2.00	57325 (%27.37)	13.14 ± 1.96	0.07	<0.001	<0.001	1.015	1.009-1.021
MPV	152139 (%72.63)	9.20 ± 1.49	57325 (%27.37)	9.19 ± 1.49	0.00	0.642	0.733	0.999	0.991-1.006

Anemi, Diğer

Anemi ile ilişkili diğer durumlar, kontrol grubunda %39.8 oranında bulunurken, deney grubunda %24.3 oranında saptanmıştır. Etki büyüklüğü 0.29 olan bu fark, istatistiksel olarak anlamlıdır ($p < 0.001$). Lojistik regresyon analizi sonuçları da bu durumu desteklemekte olup, p değeri <0.001 ve odds oranı 0.739 (95% GA: 0.719 - 0.759) olarak bulunmuştur. Bu, anemi ile ilişkili diğer durumların kanser riski üzerinde önemli bir etkisi olduğunu göstermektedir. Anemi gibi yaygın ve ciddi sağlık durumlarının kanser riski ile ilişkilendirilmesi, özellikle kansızlık gibi kronik durumların kanserin gelişiminde rol oynayabileceğini düşündürmektedir.

Kronik İskemik Kalp Hastalığı

Kronik iskemik kalp hastalığı kontrol grubunda %43.9, deney grubunda ise %31.5 oranında bulunmuştur. Etki büyüklüğü 0.23 olup, bu fark istatistiksel olarak anlamlıdır ($p < 0.001$). Lojistik regresyon analizinde p değeri <0.001 ve odds oranı 0.915 (95% GA: 0.886 - 0.944) olarak bulunmuştur. Bu, kronik iskemik kalp hastalığının kanser riski ile ilişkili olduğunu göstermektedir. Kalp hastalıklarının genel sağlık üzerindeki olumsuz etkileri, kanser gelişimini tetikleyen sistemik inflamasyon ve oksidatif stres gibi mekanizmalarla ilişkilendirilebilir.

Üriner Sistemin Diğer Bozuklukları

Üriner sistem bozuklukları kontrol grubunda %48.7, deney grubunda %30.5 oranında saptanmıştır. Etki büyüklüğü 0.33 olan bu fark istatistiksel olarak anlamlıdır ($p < 0.001$). Lojistik regresyon analizi sonuçları, p değerinin <0.001 ve odds oranının 0.738 (95% GA: 0.719 - 0.758) olduğunu göstermektedir. Bu, üriner sistem bozukluklarının kanser riski üzerinde önemli bir etkisi olduğunu göstermektedir. Üriner sistem bozukluklarının kanserle ilişkili olabileceği, bu sistemdeki kronik inflamasyon ve hücrel hasarın kanser gelişimine katkıda bulunabileceğini düşündürmektedir.

Boğaz ve Göğüste Ağrı

Boğaz ve göğüste ağrı kontrol grubunda %50.9, deney grubunda %32.6 oranında bulunmuştur. Etki büyüklüğü 0.33 olup, bu fark istatistiksel olarak anlamlıdır ($p < 0.001$). Lojistik regresyon analizinde p değeri <0.001 ve odds oranı 0.738 (95% GA: 0.718 - 0.757) olarak bulunmuştur. Bu sonuçlar, boğaz ve göğüste ağrının kanser riski ile ilişkili olduğunu göstermektedir. Kronik ağrıların kanser riski ile ilişkili olması, bu ağrıların altında yatan inflamatuvar süreçlerin kanser gelişimini destekleyebileceğini düşündürmektedir.

Abdominal ve Pelvik Ağrı

Abdominal ve pelvik ağrı kontrol grubunda %57.7, deney grubunda %33.8 oranında saptanmıştır. Etki büyüklüğü 0.44 olan bu fark istatistiksel olarak anlamlıdır ($p < 0.001$). Lojistik regresyon analizi, p değerinin <0.001 ve odds oranının 0.64 (95% GA: 0.624 - 0.656) olduğunu göstermektedir. Bu, abdominal ve pelvik ağrının kanser riski üzerinde önemli bir etkisi olduğunu ortaya koymaktadır. Pelvik ve abdominal ağrıların kanserle ilişkili olabileceği, bu bölgelerdeki kronik inflamasyon ve hücrel anormalliklerin kanser gelişimine yol açabileceğini düşündürmektedir.

Ağrı, Başka Yerde Sınıflanmamış

Başka yerde sınıflanmamış ağrı, kontrol grubunda %44.2, deney grubunda %33.3 oranında bulunmuştur. Etki büyüklüğü 0.2 olup, bu fark istatistiksel olarak anlamlıdır (p

< 0.001). Lojistik regresyon analizinde p değeri <0.001 ve odds oranı 0.833 (95% GA: 0.813 - 0.854) olarak bulunmuştur. Bu, başka yerde sınıflanmamış ağrının kanser riski ile ilişkili olduğunu göstermektedir. Kronik ağrıların kanser riski ile ilişkili olması, bu ağrıların altında yatan inflamatuvar süreçlerin kanser gelişimini destekleyebileceğini düşündürmektedir.

Yakınma veya Bilinen Teşhisi Olmayan Kişilerin Genel Muayene ve İncelemesi

Genel muayene ve inceleme, kontrol grubunda %84.9, deney grubunda %70 oranında saptanmıştır. Etki büyüklüğü 0.34 olan bu fark istatistiksel olarak anlamlıdır ($p < 0.001$). Lojistik regresyon analizi sonuçları, p değerinin <0.001 ve odds oranının 0.651 (95% GA: 0.633 - 0.669) olduğunu göstermektedir. Bu, genel muayene ve incelemenin kanser riski üzerinde önemli bir etkisi olduğunu ortaya koymaktadır. Genel muayene ve inceleme gibi sağlık hizmetlerinin kanser riski ile ilişkili olması, bu hizmetlerin erken teşhis ve önleyici tedavi uygulamaları açısından önemini vurgulamaktadır.

Bu bulgular, belirli hastalıkların ve semptomların kanser riski üzerindeki etkilerini anlamada önemli ipuçları sunmaktadır. Özellikle etki büyüklüğü yüksek olan hastalıklar, kanser riski ile ters orantılı bir ilişki göstermektedir.

Her hastalık için, tanı oranları ve istatistiksel önemlilik kanser risk faktörleri hakkında fikir verirken, bu bilgilerin klinik uygulamalar ve hasta yönetimi için düşünülmesi gereken faktörler olduğu unutulmamalıdır. Bu analizler, kanserle ilişkili risk faktörlerini daha iyi anlamamıza ve bu bilgileri kullanarak hastalığın önlenmesi ve yönetiminde yenilikçi stratejiler geliştirmemize yardımcı olacaktır.

Tablo 4.3.'de, kanser tanısı alan bireylerde anlamlı sonuçlar veren ancak etki büyüklüğü 0.2'den küçük olan komorbiditeler incelenmiştir. Dişler ve destekleyici yapıların diğer hastalıkları, dış kulağın diğer bozuklukları, refraksiyon ve akomodasyon bozuklukları, omuz lezyonları ve prostat hiperplazisi gibi komorbiditeler kanser tanısı alan bireylerde daha yüksek oranlarda bulunmuştur. Özellikle refraksiyon ve akomodasyon bozuklukları ile prostat hiperplazisi, kontrol grubuna göre daha yüksek oranlarda görülmüş ve bu durum kanser riskini anlamlı derecede artırmaktadır. Lojistik regresyon sonuçlarına göre,

bu komorbiditelerin odds oranları birden büyük olup, bu hastalıkların kanser riskini artırdığı gözlemlenmiştir. Ancak, etki büyüklüklerinin 0.2'den küçük olması, bu ilişkilerin görece düşük derecede olduğunu göstermektedir. Bu nedenle, bu komorbiditelerin kanser riskinin belirlenmesindeki bağımsız etkisi daha düşük olabilir ve başka faktörlerle birlikte değerlendirilmelidir. Bu bulgular, çeşitli komorbiditelerin kanser riski ile ilişkili olduğunu ortaya koysa da, daha fazla araştırma ve daha büyük veri setleri ile bu ilişkilerin daha detaylı incelenmesi gerekmektedir.

Tablo 4.3. Etki büyüklüğü 0.2'den küçük olup Lojistik Regresyonda anlamlı sonuç veren komorbiditeler

Özellik		Kontrol (%n)	Deney (%n)	EB	p	Lojistik Regresyon		
						p	Odds Oranı	%95 GA
Dışler ve destekleyici yapıların diğer hastalıkları	Var	14340 (%9.4)	6601 (%11.5)	0.06	<0.001	<0.001	1.152	1.108 - 1.199
Dış kulağın diğer bozuklukları	Var	15085 (%9.9)	6249 (%10.9)	0.03	<0.001	0.004	1.058	1.018 - 1.100
Refraksiyon ve akomodasyon bozuklukları	Var	48335 (%31.8)	21187 (%37)	0.10	<0.001	<0.001	1.185	1.155 - 1.216
Omuz lezyonları	Var	10152 (%6.7)	4062 (%7.1)	0.01	0.001	0.001	1.078	1.030 - 1.128
Prostat hiperplazisi	Var	24204 (%15.9)	10860 (%18.9)	0.07	<0.001	<0.001	1.397	1.344 - 1.452

Bölgesel Farklılıklar

Lojistik regresyonda İstanbul bölgesi referans alınmıştır. Buna göre Batı Anadolu ile İstanbul arasında fark yokken ($p=0,238$), Batı Marmara, Ege, Akdeniz, Orta Anadolu, Batı Karadeniz, Doğu Karadeniz, Kuzeydoğu Anadolu, Ortadoğu Anadolu ve Güneydoğu Anadolu bölgelerinde kanser riski İstanbul bölgesine göre daha düşüktür ($p<0.001$) (Tablo 4.4.).

Tablo 4.4. Kontrol ve Deney Gruplarının Bölgelere Göre İstatistiksel Analizleri

Bölgeler	Kontrol (%n)	Deney (%n)	p	Lojistik Regresyon	
				Odds Oranı	%95 GA
İstanbul	20262 (%13.3)	10331 (%18)			
Batı Marmara	13489 (%8.9)	4930 (%8.6)	<0.001	0.406	0.387-0.427
Ege	14956 (%9.8)	7190 (%12.5)	<0.001	0.668	0.638-0.700
Doğu Marmara	9744 (%6.4)	5157 (%9)	<0.001	0.970	0.921-1.021
Batı Anadolu	28892 (%19)	10436 (%18.2)	0.238	0.543	0.521-0.566
Akdeniz	15101 (%9.9)	5323 (%9.3)	<0.001	0.636	0.606-0.668
Orta Anadolu	7133 (%4.7)	1682 (%2.9)	<0.001	0.361	0.337-0.386
Batı Karadeniz	13378 (%8.8)	4471 (%7.8)	<0.001	0.597	0.568-0.628
Doğu Karadeniz	6578 (%4.3)	2356 (%4.1)	<0.001	0.654	0.613-0.697
Kuzeydoğu Anadolu	4655 (%3.1)	1123 (%2)	<0.001	0.515	0.475-0.559
Ortadoğu Anadolu	7510 (%4.9)	2389 (%4.2)	<0.001	0.664	0.623-0.707
Güneydoğu Anadolu	10441 (%6.9)	1937 (%3.4)	<0.001	0.439	0.412-0.469

Cinsiyet

Cinsiyet dağılımına bakıldığında, kanser tanısı almış bireyler ile kontrol grubundaki bireyler arasında anlamlı bir fark bulunmaktadır ($p < 0.001$): Kadınlar kontrol grubunda %62.9, deney grubunda %55.9 iken; erkekler kontrol grubunda %37.1, deney grubunda %44.1 oranındadır.

Lojistik regresyon analizine göre cinsiyetin kanser riski üzerindeki etkisi istatistiksel olarak anlamlı değildir ($p = 0.075$, odds oranı 0.971; %95 GA: 0.939-1.003). Kikare testinin anlamlı sonuç vermesine rağmen lojistik regresyonun anlamlı sonuç vermemesi, cinsiyetin kanser riski üzerindeki etkisinin küçük olduğunu ve daha basit testlerde anlamlı olarak görülebileceğini göstermektedir. Ayrıca, lojistik regresyon modelinde başka değişkenler de etkili olduğundan, cinsiyetin bağımsız etkisi daha düşük olabilir ve bu nedenle p-değeri anlamlılık sınırını aşamamış olabilir.

Yaş

Yaşın kanser riski üzerinde anlamlı bir etkisi vardır ($p < 0.001$). Bu sonuç, yaş değişkeninin kontrol ve deney gruplarında farklı dağılımlar gösterdiğini ve yaşın kanser tanısı ile ilişkili olabileceğini göstermektedir.

Lojistik regresyon analizi, yaş değişkeninin kanser riski üzerindeki etkisini daha detaylı bir şekilde ortaya koymaktadır: $p < 0.001$ yaş değişkeninin kanser riski üzerinde anlamlı bir etkisi olduğunu göstermektedir. Odds Oranı (1.031, %95 GA: 1.029 - 1.032, yaş arttıkça kanser tanısı alma olasılığının arttığını göstermektedir. Her bir birim yaş artışı için, kanser tanısı alma olasılığı %3.1 oranında artmaktadır. Bu artış küçük gibi görünse de, yaşı artışıyla birlikte kanser riskinin istikrarlı bir şekilde arttığını göstermektedir.

ALT (Alanine Aminotransferase)

Kikare testi sonuçlarına göre, ALT seviyelerinin kanser riski üzerinde anlamlı bir etkisi vardır ($p < 0.001$). ALT değişkeni kontrol ve deney gruplarında farklı dağılımlar göstermektedir ve ALT seviyelerinin kanser tanısı ile ilişkili olabileceğini göstermektedir. Lojistik regresyon analizi, ALT seviyelerinin kanser riski üzerinde anlamlı ($p < 0.001$) bir etkisi olduğunu göstermektedir (Odds Oranı: 0.998, %95 GA: 0.997 - 0.998). ALT seviyelerinin kanser tanısı alma olasılığını az da olsa azalttığını göstermektedir. ALT seviyelerindeki her bir birim artış için, kanser tanısı alma olasılığı yaklaşık %0.2 oranında azalmaktadır. Hem kikare testi hem de lojistik regresyon analizi, ALT seviyelerinin kanser riski üzerinde anlamlı bir etkiye sahip olduğunu göstermektedir. Özellikle, lojistik regresyon analizi, ALT seviyelerinin kanser tanısı alma olasılığını az da olsa azalttığını açıkça ortaya koymaktadır.

EOS (Eosinophil)

Kontrol grubunda medyan EOS değeri 2.25 (25-75 persentil: 1.5-3.4) iken, deney grubunda da medyan EOS değeri 2.25 (25-75 persentil: 1.5-3.4) olarak bulunmuştur. EOS seviyeleri arasında belirgin bir farklılık bulunmamakta olup, odds oranı 0.999 (%95 GA: 0.992-1.006) ve $p=0.718$ 'dir. Bu sonuçlar EOS seviyelerinin kanser tanısı ile anlamlı bir ilişki göstermediğini belirtmektedir.

PLT (Platelet)

Kontrol grubunda medyan PLT değeri 247 (25-75 persentil: 201-300) iken, deney grubunda medyan PLT değeri 247 (25-75 persentil: 201-299) olarak bulunmuştur. PLT seviyeleri arasında belirgin bir farklılık bulunmamakla birlikte, odds oranı 1.015 (%95

GA: 1.009-1.021) ve $p=0.793$ 'dir. Bu sonuç, PLT seviyelerinin kanser tanısı ile anlamlı bir ilişki göstermediğini belirtmektedir.

HGB (Hemoglobin)

Kontrol grubunda ortalama HGB değeri 12.99 ± 2.00 iken, deney grubunda ortalama HGB değeri 13.14 ± 1.96 olarak bulunmuştur. HGB seviyeleri arasında anlamlı bir farklılık bulunmaktadır. Odds oranı 1.015 (%95 GA: 1.009-1.021) ve $p<0.001$ 'dir. Bu sonuç, HGB seviyelerinin kanser tanısı ile anlamlı bir ilişki gösterdiğini belirtmektedir.

MPV (Mean Platelet Volume)

Kontrol grubunda ortalama MPV değeri 9.20 ± 1.49 iken, deney grubunda ortalama MPV değeri 9.19 ± 1.49 olarak bulunmuştur. MPV seviyeleri arasında belirgin bir farklılık bulunmamakla birlikte, odds oranı 0.999 (%95 GA: 0.991-1.006) ve $p=0.642$ 'dir. Bu sonuç, MPV seviyelerinin kanser tanısı ile anlamlı bir ilişki göstermediğini belirtmektedir.

Makine Öğrenmesi Hiperparametreleri

Kanser tahmini için çeşitli makine öğrenme modellerinin performansları değerlendirilmiş ve en iyi hiperparametreler belirlenmiştir. Tablo 4.5.'te gösterildiği gibi, her model için en iyi hiperparametreler grid search yöntemi kullanılarak seçilmiştir.

- Lojistik Regresyon modeli için en iyi hiperparametreler Penalty: 12, C: 0.004832930238571752, Solver: liblinear olarak belirlenmiştir. Bu parametreler, modelin doğruluğunu ve genelleme yeteneğini artırmıştır.
- SVM (Destek Vektör Makineleri) modeli için en iyi hiperparametre C: 0.01 olarak belirlenmiştir. Bu parametre, modelin karar sınırlarını optimize ederek doğruluğu artırmıştır.
- XGBoost modeli, Learning Rate: 0.1, Max Depth: 5, n_estimators: 200 hiperparametreleri ile en iyi performansı göstermiştir. Bu hiperparametreler, modelin öğrenme hızını ve derinliğini dengeleyerek overfitting riskini azaltmıştır.

- Karar Ağaçları (DT) modeli için en iyi hiperparametreler Criterion: entropy, Max Depth: 10, Min Samples Split: 10, Min Samples Leaf: 6 olarak belirlenmiştir. Bu parametreler, ağacın dallanma kriterlerini ve maksimum derinliğini belirleyerek modelin performansını optimize etmiştir.
- Rastgele Ormanlar (RF) modeli, Max Depth: 40, Max Features: sqrt, Min Samples Leaf: 2, Min Samples Split: 9, n_estimators: 400 hiperparametreleri ile en iyi performansı göstermiştir. Bu hiperparametreler, ormanın derinliğini ve dallanma kriterlerini dengeleyerek modelin genelleme yeteneğini artırmıştır.
- Yapay Sinir Ağları (ANN) modeli için en iyi hiperparametreler activation: relu, solver: adam olarak belirlenmiştir. Bu parametreler, modelin öğrenme sürecini ve aktivasyon fonksiyonunu optimize ederek performansını artırmıştır.

Tablo 4.5. Modeller için Elde Edilen En İyi Hiperparametreler

Model	Hiperparametreler
Lojistik Regresyon	Penalty: l2, C: 0.004832930238571752, Solver: liblinear
SVM	C: 0.01
XGBoost	Learning Rate: 0.1, Max Depth: 5, n_estimators: 200
Karar Ağaçları (DT)	Criterion: entropy, Max Depth: 10, Min Samples Split: 10, Min Samples Leaf: 6
Rastgele Ormanlar (RF)	Max Depth: 40, Max Features: sqrt, Min Samples Leaf: 2, Min Samples Split: 9, n_estimators: 400
Yapay Sinir Ağları (ANN)	activation: relu, solver: adam

Modellerin Performans Ölçütleri

Tablo 4.6.'da modellerin genel performanslarını ve metrikleri özetlenmektedir. Ayrıca ROC ve Precision Recall eğri grafikleri Ek 3 de bulunmaktadır.

Lojistik Regresyon modeli, %80 doğruluk ve %83 AUC değeri ile yüksek performans göstermiştir. Duyarlılık değeri %51, F1 skoru %58, ve kesinlik değeri %68 olarak kaydedilmiştir. Precision-recall curve (PRC) değeri ise %66'dır.

SVM (Destek Vektör Makineleri) modeli, %80 doğruluk ve %83 AUC değeri ile Lojistik Regresyon modeline benzer performans sergilemiştir. Duyarlılık değeri %49, F1 skoru %57, ve kesinlik değeri %69'dur. PRC değeri %66 olarak belirlenmiştir.

XGBoost modeli, %81 doğruluk ve %85 AUC değeri ile en yüksek performansı göstermiştir. Duyarlılık değeri %55, F1 skoru %61, ve kesinlik değeri %69'dur. PRC değeri ise %70 ile diğer modellerden daha yüksektir.

Karar Ağaçları (DT) modeli, %77 doğruluk ve %79 AUC değeri ile diğer modellere kıyasla daha düşük performans sergilemiştir. Duyarlılık değeri %40, F1 skoru %49, ve kesinlik değeri %62 olarak kaydedilmiştir. PRC değeri %57'dir.

Rastgele Ormanlar (RF) modeli, %79 doğruluk ve %82 AUC değeri ile performans göstermiştir. Duyarlılık değeri %36, F1 skoru %48, ve kesinlik değeri %75'tir. PRC değeri %65 olarak belirlenmiştir.

Yapay Sinir Ağları (ANN) modeli, %79 doğruluk ve %81 AUC değeri ile orta seviyede performans göstermiştir. Duyarlılık değeri %51, F1 skoru %57, ve kesinlik değeri %65'dir. PRC değeri ise %61 olarak kaydedilmiştir.

Bu sonuçlar, dengesiz veri seti ile yapılan çalışmada, XGBoost modelinin en yüksek performansı sergilediğini göstermektedir. Özellikle yüksek AUC-ROC ve PRC değerleri, bunun yanında duyarlılık, F1 skor ve MCC diğer modellere göre daha yüksek olması, bu modeli tercih sebebi yapmaktadır. Diğer modeller belirli metriklerde iyi performans sergilese de, genel olarak XGBoost modelinin üstünlüğü dikkate değerdir. Bu bulgular, kanser tahmini gibi uygulamalarda XGBoost modelinin tercih edilmesini önermektedir.

Tablo 4.6. Kanser Tahmini Performans Sonuçları

Modeller	Doğruluk	Duyarlılık	F1	AUC (ROC) (%95 GA)	Kesinlik	Precision Recall Curve	MCC
LR	0.80	0.51	0.58	0.829 (0.824-0.833)	0.68	0.658	0.47
SVM	0.80	0.49	0.57	0.827 (0.823-0.832)	0.69	0.656	0.46
XGBOOST	0.81	0.55	0.61	0.846 (0.841-0.850)	0.69	0.704	0.49
DT	0.77	0.40	0.49	0.786 (0.781-0.791)	0.62	0.567	0.36
RF	0.79	0.36	0.48	0.816 (0.811-0.820)	0.75	0.683	0.36
ANN	0.79	0.51	0.57	0.813 (0.804-0.813)	0.65	0.614	0.44

***En yüksek performansı gösteren XGBOOST modeli matematiksel kesme noktasında duyarlılık 0.55, seçicilik 0.91 olarak bulunmuştur. Değişik kesme noktalarındaki duyarlılık ve seçiciliklerin gösterildiği ayrıntılı tablo ve grafik EK 4'tedir.**

5. TARTIŞMA

Çalışmamızda, komorbiditelerin, kan tetkiklerinin ve hastanın demografik bilgilerinin sonraki yıllardaki kanser riski üzerindeki etkileri üzerine önemli bilgiler elde edilmiştir. Bu riski hesaplamak için Lojistik Regresyon, Destek Vektör Makineleri, XGBoost, Karar Ağaçları, Rastgele Orman, Yapay Sinir Ağları makine öğrenmesi modelleri uygulanmış, en başarılı modelin XGBOOST olduğu görülmüştür.

Komorbiditelerin kanser riski ile anlamlı ilişki gösterenlerinin tamamına yakınının kanser olmayanlarda (kontrol grubunda) daha yüksek olduğu görüldü. Bu durumun çalışmanın tasarımı ile ilgili olduğunu düşünüyoruz. Kontrol grubuna, elimizde yeterli olması amacı ile çok tetkik sonucu bulunan hastalar seçilmişti. Kontrol grubunda çok görülen komorbiditelerin genellikle kronik hastalıklar olduğu görülmektedir. Bu hastaların hekim ziyaretlerinin sık olması ve e-nabız'daki veri hacminin önemli bir kısmının oluşturmalarının doğal olduğunu düşünmekteyiz. Ancak, bu durum bu bulgunun değerini azaltmamaktadır. Bu çalışmanın nihai amacı, rutin elektronik sağlık kaydından gelecekteki kanser olgularını tahmin ederek erken tanı koymaktır. Kanser hastalarında hastalık ortaya çıkmadan önceki dönemde komorbiditelerin genel olarak az olması da bizim amacımıza katkı sağlayan bir bilgidir.

Öte yandan, bazı komorbiditelerin kanser hastalarında daha az görüldüğüne dair başka çalışmalar da vardır. Örneğin demir eksikliği anemisi ve diğer anemi türleri, kanser tanısı alan bireylerde anlamlı derecede düşük oranlarda görülmüştür. Lojistik regresyon analizinde demir eksikliği anemisi için odds oranı 0.879 ve diğer anemi türleri için 0.739 olarak bulunmuştur. Ludwig ve arkadaşlarının (2004) çalışmasında da benzer şekilde anemi ile kanser riski arasında negatif bir ilişki bulunmuştur. Anemi durumunun, vücudun savunma mekanizmalarını harekete geçirdiği ve kanser hücrelerinin büyümesini engelleyebileceği önerilmiştir (Ludwig et al., 2004). Kronik iskemik kalp hastalığı, kanser tanısı alan bireylerde anlamlı derecede düşük oranlarda görülmüştür (odds oranı 0.915). Bu bulgu, kalp hastalığı olan bireylerin kanser riskinin azaldığını göstermektedir. Ancak, bu ilişki karmaşık olabilir ve daha fazla araştırma gerektirmektedir. Tu ve arkadaşlarının (2018) çalışmasında da kronik hastalıkların kanser riskini nasıl etkilediği

tartışılmıştır (Tu et al., 2018). Üriner sistemin diğer bozuklukları, kanser tanısı alan bireylerde düşük oranlarda bulunmuştur (odds oranı 0.738). Bu durum, üriner sistem bozukluklarının kanser riskini azalttığını göstermektedir. Ancak, bu bulgunun klinik anlamı tam olarak anlaşılamamıştır ve daha fazla araştırma gereklidir. Boğaz ve göğüste ağrı, kanser tanısı alan bireylerde anlamlı derecede düşük oranlarda bulunmuştur (odds oranı 0.738). Bu, bu tür ağrıların kanser riskini azaltıcı bir etkisi olabileceğini göstermektedir. Bu bulgu, Multhoff ve arkadaşlarının (2012) çalışmasında benzer şekilde rapor edilmiştir (Multhoff et al., 2012). Abdominal ve pelvik ağrı, kanser tanısı alan bireylerde düşük oranlarda bulunmuştur (odds oranı 0.64). Bu, bu tür ağrıların kanser riskini azaltabileceğini düşündürmektedir. Bu bulgu, gastrointestinal sistemle ilgili inflamatuvar durumların kanser riskini etkileyebileceğini belirten çalışmaları desteklemektedir (McConnell et al., 2009). Diğer belirtilmemiş ağrı türleri de kanser tanısı alan bireylerde düşük oranlarda görülmüştür (odds oranı 0.833). Bu bulgu, ağrının kanser riskini azaltıcı bir etkisi olabileceğini göstermektedir. Ancak, bu ilişkinin altında yatan mekanizmalar tam olarak anlaşılmamıştır ve daha fazla araştırma gereklidir. Yakınma veya bilinen teşhisi olmayan kişilerin genel muayene ve incelemesi, kanser tanısı alan bireylerde düşük oranlarda bulunmuştur (odds oranı 0.651). Bu, genel sağlık taramalarının kanser riskini azaltmada etkili olabileceğini göstermektedir. Ancak, bu bulgunun klinik önemi ve uygulamaları hakkında daha fazla araştırma yapılması gerekmektedir.

Çalışmada, lojistik regresyon analizi kullanılarak Türkiye'nin çeşitli bölgelerinde kanser riski değerlendirilmiştir. İstanbul bölgesi referans alınarak yapılan analizler, İstanbul ile karşılaştırıldığında diğer bölgelerde kanser riskinin daha düşük olduğunu ortaya koymuştur. Batı Anadolu bölgesi ile İstanbul arasında anlamlı bir fark bulunmamakla birlikte ($p = 0.238$), Batı Marmara, Ege, Doğu Marmara, Akdeniz, Orta Anadolu, Batı Karadeniz, Doğu Karadeniz, Kuzeydoğu Anadolu, Ortadoğu Anadolu ve Güneydoğu Anadolu bölgelerinde kanser riskinin istatistiksel olarak anlamlı derecede düşük olduğu tespit edilmiştir ($p < 0.001$). Bu bulgu, bölgesel farklılıkların kanser riski üzerindeki etkisini vurgulamaktadır. Literatürde, çevresel faktörlerin, yaşam tarzı alışkanlıklarının ve genetik yatkınlıkların kanser insidansında önemli rol oynadığı belirtilmiştir (Vineis et al.,

2004). Örneğin, İstanbul gibi büyük metropol alanlarında hava kirliliği, stres ve yaşam tarzı faktörlerinin kanser riskini artırabileceği düşünülebilir. Diğer bölgelerdeki daha düşük risk, bu faktörlerin daha az yoğun olmasıyla açıklanabilir. Ayrıca, sağlık hizmetlerine erişim ve kanser tarama programlarının yaygınlığı da bu farklılıklarda etkili olabilir. Bu sonuçlar, sağlık politikalarının ve kanser tarama programlarının bölgesel düzeyde özelleştirilmesi gerektiğini önermektedir. Özellikle yüksek risk taşıyan bölgelerde daha yoğun tarama ve önleme programlarının uygulanması, kanser insidansını azaltmada etkili olabilir. Ayrıca, düşük riskli bölgelerde de mevcut risk faktörlerinin izlenmesi ve gerekli müdahalelerin yapılması önemlidir. Bu çalışma, Türkiye'deki bölgesel kanser risklerini anlamada önemli bir adım olup, gelecekteki araştırmalara ve sağlık politikalarına rehberlik edebilir.

Cinsiyetin kanser riski üzerindeki etkisi çalışmamızda belirgin değildir. Kikare testi sonuçlarına göre, cinsiyetin kanser riski üzerindeki etkisi anlamlı bulunmuş ($p < 0.001$), ancak lojistik regresyon analizinde bu etki anlamlı bulunmamıştır ($p = 0.075$, odds oranı 0.971, %95 GA: 0.939-1.003). Kontrol grubunun %62.9'i, Çalışma grubunun %55.9'u kadındır. Ki-kare testinin büyük gruplarda aşırı anlamlı olma eğilimde olduğu unutulmamalıdır. Bu durum, cinsiyetin kanser riski üzerindeki etkisinin küçük olabileceğini ve daha karmaşık modellerde bu etkinin kaybolabileceğini göstermektedir. Ayrıca, bu sonuç cinsiyetin diğer örtük değişkenlerle etkileşime girebileceğini ve bu etkileşimlerin kanser riskini doğrudan etkileyebileceğini göstermektedir (Ferlay et al., 2015). Ayrıca yaş, kanser riskinde önemli bir faktör olarak bulunmuştur ($p < 0.001$, Odds ratio: 1.031, %95 GA: 1.029-1.032). Yaş arttıkça kanser riskinin arttığı gözlemlenmiştir. Bu bulgu, literatürdeki yaşın kanser riskini artırdığına dair genel görüşü desteklemektedir (Ferlay et al., 2015).

EOS seviyeleri kanser tanısı alan ve almayan gruplar arasında anlamlı bir farklılık göstermemiştir ($p = 0.718$, odds oranı 0.999, %95 GA: 0.992-1.006). Bu bulgu, eosinophil seviyelerinin kanser riski ile ilişkili olmadığını göstermektedir. Literatürde eosinophil seviyelerinin bazı kanser türleri ile ilişkili olabileceği belirtilse de, bu çalışma genel kanser riskinde EOS seviyelerinin belirleyici bir faktör olmadığını ortaya koymaktadır (Varricchi et al., 2017). PLT seviyeleri, kanser tanısı alan ve almayan gruplar arasında

belirgin bir farklılık göstermemiştir ($p = 0.793$, odds oranı 1.015, %95 GA: 1.009-1.021). Bu, PLT seviyelerinin kanser riski ile anlamlı bir ilişki göstermediğini belirtmektedir. Ancak, bazı çalışmalarda yüksek PLT seviyelerinin belirli kanser türlerinde prognostik bir gösterge olabileceği belirtilmiştir. Bu nedenle, PLT seviyelerinin kanser riskinde genel bir belirleyici faktör olmadığı, ancak belirli kanser türlerinde önemli olabileceği düşünülmelidir (Song et al., 2019). HGB seviyeleri arasında anlamlı bir farklılık bulunmuş olup, kanser tanısı alan bireylerde HGB seviyeleri daha yüksektir ($p < 0.001$, odds oranı 1.015, %95 GA: 1.009-1.021). Bu bulgu, yüksek hemoglobin seviyelerinin kanser riski ile ilişkili olabileceğini göstermektedir. Literatürde de hemoglobin seviyelerinin belirli kanser türlerinde prognostik bir gösterge olabileceği belirtilmiştir (Lind et al., 2002). MPV seviyeleri, kanser tanısı alan ve almayan gruplar arasında belirgin bir farklılık göstermemiştir ($p = 0.642$, odds oranı 0.999, %95 GA: 0.991-1.006). Bu durum, MPV seviyelerinin kanser riski ile anlamlı bir ilişki göstermediğini göstermektedir. Literatürde MPV'nin belirli kanser türlerinde prognostik bir gösterge olabileceği belirtilmiş olsa da, genel kanser riskinde önemli bir faktör olmadığı anlaşılmaktadır (Detopoulou et al., 2023).

Bu bulgular, çalışmamızda değerlendirilen çeşitli hematolojik değişkenlerin kanser riski ile ilişkisini ortaya koymaktadır. Genel olarak, bazı değişkenler kanser riski ile anlamlı ilişkiler gösterirken, diğerleri anlamlı bir ilişki göstermemiştir. Bu sonuçlar, kanser riskini belirlemede kullanılan parametrelerin önemini vurgulamakta ve bu alanda daha fazla araştırma yapılması gerektiğini göstermektedir. Literatürdeki çalışmalarla uyumlu olarak, kanser tanısında ve risk değerlendirmesinde çoklu faktörlerin dikkate alınmasının önemi bir kez daha ortaya konmaktadır.

Çalışmamızda elde edilen bulguların literatürdeki benzer çalışmalarla uyumlu olduğunu ve bazı durumlarda yeni perspektifler sunduğunu gözlemledik. Daha fazla araştırma, bu ilişkilerin daha iyi anlaşılmasına ve kanser riski yönetiminde yenilikçi stratejiler geliştirilmesine yardımcı olacaktır.

Aynı zamanda bu çalışmada, kanser tahmini için çeşitli makine öğrenme modellerinin performansları incelenmiştir. Sonuçlar, özellikle XGBoost modelinin diğer modellere

kıyasla daha yüksek performans gösterdiğini ortaya koymuştur. XGBoost modeli, doğruluk, MCC AUC-ROC ve PRC metriklerinde en yüksek değerleri elde etmiştir. Bu bulgu, literatürdeki benzer çalışmalarla uyumlu olup, XGBoost'un sağlık verileri üzerinde etkili bir tahmin aracı olduğunu desteklemektedir. Örneğin, Chen ve Guestrin'in (2016) çalışması, XGBoost'un hız ve performans açısından diğer gradient boosting algoritmalarına göre üstün olduğunu göstermiştir.

Lojistik regresyon, SVM, karar ağaçları, rastgele ormanlar ve yapay sinir ağları (ANN) gibi diğer modeller de yüksek doğruluk oranları sunmakla birlikte, XGBoost'un özellikle duyarlılık, F1 skor ve MCC gibi kritik metriklerde üstünlük sağlaması, bu modeli kanser tahmini için ideal bir seçenek haline getirmektedir. Bu sonuçlar, Hastie, Tibshirani ve Friedman'ın (2009) makine öğrenimi yöntemlerinin tıbbi veri analizlerinde önemli katkılar sağlayabileceğine dair bulgularını desteklemektedir. Bununla birlikte, yapay sinir ağları (ANN) gibi derin öğrenme modelleri de sağlık verilerinde yüksek doğruluk oranları ile tanınmaktadır. Özellikle ANN'nin karmaşık verileri modellemedeki yeteneği, genellikle daha derin ve geniş veri setleri üzerinde üstün performans göstermektedir. Ancak, bu çalışmada ANN'nin diğer modellere göre daha düşük performans göstermesi, veri setinin dengesizliği ve modelin hiperparametre optimizasyonundaki sınırlamalarla açıklanabilir.

Çalışmamızda kullanılan ANN modeli, %79 doğruluk ve %78 AUC değeri ile orta seviyede performans göstermiştir. Duyarlılık değeri %42, F1 skoru %57, precision değeri %67 ve MCC %44 olarak kaydedilmiştir. PRC değeri ise %56 olarak belirlenmiştir. Bu sonuçlar, ANN'nin bu veri setinde XGBoost kadar etkili olmadığını göstermektedir. Ancak, ANN'nin potansiyeli göz ardı edilmemelidir. LeCun, Bengio ve Hinton (2015) tarafından yapılan çalışmalar, ANN'nin büyük ve dengeli veri setlerinde üstün performans sergilediğini göstermektedir.

Diğer taraftan, lojistik regresyon modeli, %80 doğruluk ve %83 AUC değeri ile yüksek performans göstermiştir. Duyarlılık değeri %51, F1 skoru %58 ve precision değeri %68 olarak kaydedilmiştir. Precision-recall curve (PRC) değeri ise %66'dır. Lojistik regresyon, özellikle yorumlanabilirliği ve sadeliği nedeniyle tıbbi verilerde yaygın olarak

kullanılmaktadır. Ancak, bu modelin karmaşık veri yapılarında sınırlamaları olabileceği de göz önünde bulundurulmalıdır (Hosmer, Lemeshow ve Sturdivant, 2013).

Destek vektör makineleri (SVM) modeli, %80 doğruluk ve %83 AUC değeri ile lojistik regresyon modeline benzer performans sergilemiştir. Duyarlılık değeri %49, F1 skoru %57 ve precision değeri %69'dur. PRC değeri %66 olarak belirlenmiştir. SVM'nin yüksek boyutlu veri setlerinde etkili olduğu bilinmektedir. Ancak, kernel seçimi ve hiperparametre optimizasyonu, model performansını önemli ölçüde etkileyebilir (Cortes ve Vapnik, 1995).

Karar ağaçları (DT) modeli, %77 doğruluk ve %79 AUC değeri ile diğer modellere kıyasla daha düşük performans sergilemiştir. Duyarlılık değeri %40, F1 skoru %49 ve precision değeri %62 olarak kaydedilmiştir. PRC değeri %57'dir. Karar ağaçları, modelin yorumlanabilirliğini artırırken, overfitting riskini de beraberinde getirebilir. Bu modelin performansı, özellikle büyük ve karmaşık veri setlerinde sınırlı kalabilir (Breiman, 2001).

Rastgele ormanlar (RF) modeli, %79 doğruluk ve %82 AUC değeri ile performans göstermiştir. Duyarlılık değeri %36, F1 skoru %48 ve precision değeri %75'tir. PRC değeri %65 olarak belirlenmiştir. RF modeli, overfitting riskini azaltarak genelleme yeteneğini artırabilir, ancak modelin karmaşıklığı ve hesaplama maliyeti, uygulamada zorluklar yaratabilir (Breiman, 2001).

Bu çalışmanın bulguları, dengesiz veri seti ile yapılan analizlerde XGBoost modelinin üstün performansını göstermektedir. Özellikle yüksek MCC (Matthews Correlation Coefficient) değeri, XGBoost modelinin dengesiz veri setlerinde bile doğru ve güvenilir tahminler yapabilme yeteneğini vurgulamaktadır. MCC, özellikle sınıflar arasında dengesizlik olduğunda daha güvenilir bir performans ölçütü sunar ve bu çalışmada MCC değerlerinin değerlendirilmesi, modellerin dengesiz veri setlerinde ne kadar etkili olduğunu gösterir. Bu nedenle, MCC gibi metriklerin dikkate alınması, model performansının kapsamlı bir şekilde değerlendirilmesini sağlar (Chicco ve Jurman, 2020). Ancak, diğer modellerin de belirli metriklerde iyi performans sergilemesi, farklı uygulama alanlarında bu modellerin de dikkate alınması gerektiğini ortaya koymaktadır.

Gelecekteki alıřmalar, daha byk ve dengeli veri setleri ile yapılmalı ve farklı makine ğrenme modelleri arasındaki karşılařtırmalar daha derinlemesine incelenmelidir.

Sonuç olarak, demografik deėiřkenler, komorbidite analizleri ve makine ğrenme modelleri kanserle ilgili risk faktrlerinin daha iyi anlařılmasına aynı zamanda bu bilgiler doėrultusunda klinik uygulamaların iyileřtirilmesine olanak tanımaktadır.



6. KISITLILIKLAR

Bu çalışma, çeşitli komorbidite ve kanser riski arasındaki ilişkiye dair değerli bilgiler sunmasına rağmen, dikkate alınması gereken bazı sınırlılıklara sahiptir:

1. Veri Kaynağı ve Kalitesi:

- Bu çalışmada kullanılan veriler, tek bir ulusal veri tabanından (e-Nabız) çıkarılmıştır. Kapsamlı olmasına rağmen, bazı tutarsızlıklar veya eksik veriler sonuçları etkileyebilir. Ayrıca, farklı sağlık hizmeti sağlayıcıları arasındaki veri girişi uygulamalarındaki değişiklikler yanlılığa yol açabilir.

2. Genellenebilirlik:

- Bu çalışmanın bulguları Türk nüfusuna özgüdür ve diğer nüfuslara veya sağlık ortamlarına genellenemez. Sağlık sistemleri, kültürel uygulamalar ve genetik yapılar arasındaki farklılıklar, bu sonuçların diğer bağlamlarda uygulanabilirliğini etkileyebilir.

3. Kanser Sınıflandırması:

- Çalışma, kanser sınıflandırması için ICD-O-3 kodlarına dayanmıştır, bu da kanser teşhislerinin tüm karmaşıklığını yansıtmayabilir. Belirli kanser türlerinin yanlış sınıflandırılması veya eksik raporlanması bulguları etkileyebilir.

4. Bazı Değişkenlerin Hariç Tutulması:

- Detaylı tedavi geçmişleri, belirli ilaç kullanımları, aile öyküsü, fiziksel aktivite, sigara, alkol, madde kullanımı ve sosyoekonomik durum gibi bazı potansiyel olarak ilgili değişkenler analize dahil edilememiştir. Bu faktörlerin rolünü anlamadaki eksiklik sınırlı kalabilir.

5. İstatistiksel Sınırlamalar:

- Lojistik regresyon ve ki-kare testleri gibi birden fazla istatistiksel teknik kullanılmış olsa da, çalışma hala tip I ve tip II hatalara tabi olabilir. Ayrıca, etki büyüklükleri ve odds oranlarının kullanımı bilgilendirici olmakla birlikte, ilişkilerin tüm kapsamını yansıtmaz.

Bu sınırlılıklara rağmen, çalışma Türk nüfusunda komorbidite ve kanser riski arasındaki ilişkiye dair önemli bilgiler sunmaktadır. Gelecekteki araştırmalar, bu sınırlılıkları uzunlamasına veri içerecek şekilde ele almalı, dahil edilen değişkenlerin kapsamını genişletmeli ve bulguları çeşitli popülasyonlar ve ortamlarda doğrulamalıdır.

7. SONUÇ ve ÖNERİLER

Sonuç:

Çalışmamızda, Türkiye'de 2018-2022 yılları arasında kanser tanısı almış bireyler ile kanser tanısı almamış bireylerin verilerini analiz ederek, kanser risk faktörlerini ve makine öğrenimi modellerinin performansını değerlendirmeyi amaçlamıştır. Çeşitli makine öğrenme modelleri kullanılarak yapılan analizler, kanser tanısında potansiyel etkileri olan faktörleri ortaya koymuş ve bazı önemli bulgular elde edilmiştir.

Analizler sonucunda, diğer bölgelerin İstanbul'a göre daha düşük risk taşıdığı görülmüştür; en düşük risk ise Orta Anadolu'da tespit edilmiştir ($p<0.001$). Bu bölgesel farklılıklar, çevresel ve genetik faktörlerin kanser gelişimindeki rolünü işaret etmektedir.

Cinsiyet analizi, kadınların kontrol grubunda %62.9, deney grubunda %55.9 olduğunu; erkeklerin ise kontrol grubunda %37.1, deney grubunda %44.1 olduğunu göstermiştir. Ancak, lojistik regresyon analizinde cinsiyetin kanser riski üzerindeki etkisi istatistiksel olarak anlamlı bulunmamıştır ($p=0.075$). Bu, cinsiyetin kanser riski üzerindeki etkisinin küçük olduğunu ve tek değişkenli testlerde anlamlı olarak görülebileceğini göstermektedir.

EOS seviyeleri, kontrol ve deney grupları arasında anlamlı bir farklılık göstermemiştir ($p=0.718$, odds oranı 0.999, %95 GA: 0.992-1.006). Bu, EOS seviyelerinin kanser riski ile ilişkili olmadığını ortaya koymaktadır.

Çalışmamızda, kanser tahmini için çeşitli makine öğrenme modellerinin performansları incelenmiştir. XGBoost modeli, doğruluk, MCC, AUC-ROC ve PRC metriklerinde en yüksek değerleri elde etmiştir ve bu bulgu, literatürdeki benzer çalışmalarla uyumludur. Lojistik regresyon, SVM, karar ağaçları, rastgele ormanlar ve yapay sinir ağları (ANN) gibi diğer modeller de yüksek doğruluk oranları sunmakla birlikte, XGBoost'un özellikle duyarlılık, F1 skoru ve MCC gibi kritik metriklerde üstünlük sağlaması, bu modeli kanser tahmini için ideal bir seçenek haline getirmektedir. Ancak, ANN gibi derin öğrenme modellerinin potansiyeli de göz ardı edilmemelidir. Gelecekteki çalışmalar, daha büyük

ve dengeli veri setleri ile yapılmalı ve farklı makine öğrenme modelleri arasındaki karşılaştırmalar daha derinlemesine incelenmelidir.

Çalışmada ayrıca, demir eksikliği anemisi, kronik böbrek yetmezliği gibi komorbiditelerin kanser riski üzerindeki etkileri de incelenmiştir. Bu hastalıkların kanser riskini azalttığı tespit edilmiştir. Çalışma bulguları, komorbiditelerin yönetimi ve izlenmesinin kanser riskini azaltmada önemli olabileceğini göstermektedir. Ayrıca, bölgesel farklılıklar ve yaş faktörlerinin de kanser riski üzerinde önemli etkileri olduğu tespit edilmiştir. Batı Marmara ve Orta Anadolu bölgelerinde kanser riski diğer bölgelere kıyasla daha düşük bulunmuştur. Kadınlarda belirli kanser türleri, erkeklere kıyasla daha yaygındır ve yaş ilerledikçe kanser riski artmaktadır.

Öneriler:

1. Erken Teşhis ve Önleme Stratejileri:

- Komorbiditelerin kanser riski üzerindeki etkileri dikkate alınarak, bu hastalıkların yönetimi ve izlenmesi kanser riskini azaltmada önemli olabilir. Bu nedenle, sağlık politikalarında komorbiditelerin yönetimine daha fazla önem verilmelidir.
- Özellikle yüksek risk taşıyan bölgelerde ve yaş gruplarında düzenli tarama ve erken teşhis programları uygulanmalıdır.

2. Makine Öğrenimi Modellerinin Kullanımı:

- XGBoost gibi yüksek performanslı makine öğrenme modelleri, sağlık verilerinin analizinde ve hastalık tahmininde daha yaygın kullanılmalıdır. Bu modellerin kullanımı, sağlık hizmetlerinin kalitesini artırabilir ve hastalıkların erken teşhisine katkıda bulunabilir.
- Makine öğrenimi ve yapay zekâ teknikleri, sağlık verilerinin daha etkili bir şekilde analiz edilmesi için sağlık çalışanlarına ve araştırmacılara daha fazla eğitim verilmelidir.

3. Veri Toplama ve Analiz:

- Sağlık verilerinin daha kapsamlı ve doğru bir şekilde toplanması, analizlerin güvenilirliğini artıracaktır. Bu nedenle, veri toplama süreçlerinde standartlaştırılmış yöntemler kullanılmalı ve veri kalitesine önem verilmelidir.
- Büyük veri analitiği ve yapay zekâ teknikleri, sağlık verilerinin daha etkili bir şekilde işlenmesi ve analiz edilmesi için kullanılmalıdır.

4. Gelecek Araştırmalar:

- Bu çalışmanın bulguları, daha geniş ve çeşitli veri setleri ile doğrulanmalı ve genişletilmelidir. Özellikle farklı coğrafi bölgelerde ve demografik gruplarda benzer analizler yapılmalıdır.
- Komorbiditelerin kanser riskini nasıl etkilediği konusunda daha derinlemesine araştırmalar yapılmalı ve bu hastalıkların biyolojik mekanizmaları daha iyi anlaşılmalıdır.
- Mevcut veri setinde, en sık görülen kanser tiplerini etkileyen faktörler detaylı olarak çalışılacaktır.

5. Politika ve Uygulamalar:

- Sağlık politikaları ve uygulamaları, bu çalışmanın bulgularını dikkate alarak güncellenmeli ve kanser tarama programları optimize edilmelidir.
- Özellikle yüksek risk gruplarında (belirli yaş grupları, komorbiditesi olan bireyler, belirli coğrafi bölgeler) yoğunlaştırılmış sağlık taramaları yapılmalı ve önleyici tedbirler alınmalıdır.

Bu öneriler, kanser tahmininde ve önlenmesinde daha etkili stratejilerin geliştirilmesine katkıda bulunabileceği ve sağlık hizmetlerinin genel kalitesini artırabileceği değerlendirilmektedir.

Sonu olarak, bu alıřma, belirli blgeler, cinsiyet ve bazı biyolojik belirtelerin kanser riski zerindeki potansiyel etkilerini ortaya koymuřtur. Makine ğrenme modelleri kullanarak elde edilen bu bulgular, kanserin erken teřhisi ve nlenmesine ynelik stratejilerin geliřtirilmesine katkıda bulunabilir. Gelecekteki arařtırmalarda, daha geniř veri setleri ve farklı biyolojik belirtelerin dahil edilmesi, kanser risk faktrlerinin daha ayrıntılı bir řekilde anlařılmasına yardımcı olacaktır.



KAYNAKLAR

- Aamir, S., Rahim, A., Aamir, Z., Farooq, S., M, A. (2022). Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques. *Computational and Mathematical Methods in Medicine*, p. 13. doi:<https://doi.org/10.1155/2022/5869529>
- Abadi, M., Barham, P., Chen, J. M., Chen, Z. F., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X. Q. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of OSDI'16: 12th USENIX Symposium on Operating Systems Design and Implementation*, 265-283.
- Benzekry, S., Grangeon, J. M., Karlsen, M., Alexa, M. (2021). Machine Learning for Prediction of Immunotherapy Efficacy in Non-Small Cell Lung Cancer from Simple Clinical and Biological Data. *Cancers* 2021(6210), p. 13(24). doi:<https://doi.org/10.3390/cancers13246210>
- Birks, J., Bankhead, C., Holt, T., Fuller, A. (2017). .Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. *Cancer medicine*(6(10)), pp. 2453–2460. doi:<https://doi.org/10.1002/cam4.1183>
- Bray, F., Ferlay, F., Soerjomataram, J., Siegel, I., Siegel, R. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, pp. 394–424. doi:<https://doi.org/10.3322/caac.21492>
- Breiman, L. (2001) Random Forests. *Machine Learning*, 45, 5-32. doi:<https://dx.doi.org/10.1023/A:1010933404324>
- Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Burkov, Andriy.
- Cai, Q., Zhu, C., Yuan, Y., Feng, Q., Feng, Y., Hao, Y., & Li, J. (2019). Gastrointestinal Early Cancer Prevention & Treatment Alliance of China (GECA). Development and validation of a prediction rule for estimating gastric cancer risk in the Chinese high-risk population: a nationwide multicentre study. *Gut*, pp. 1576-1587. doi:[doi:10.1136/gutjnl-2018-317556](https://doi.org/10.1136/gutjnl-2018-317556).
- Cao, X. H. (2016). A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC bioinformatics*, *BMC*, 17(1), 359. doi:<https://doi.org/10.1186/s12859-016-1236-x>

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
- Cohen, J., Li, L., Thoburn, Y., Afsar, C. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, pp. 926–930.
doi:<https://doi.org/10.1126/science.aar3247>
- Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, 20, 273-297.
<http://dx.doi.org/10.1007/BF00994018>
- Cowling, T. E. (2020). A novel approach selected small sets of diagnosis codes with high prediction performance in large healthcare datasets. *Journal of clinical epidemiology*., 128, 20–28. doi:<https://doi.org/10.1016/j.jclinepi.2020.08.001>
- Cubiella, J., Vega, P., Salve, M., Díaz-Ondina, M. (2016). Development and external validation of a faecal immunochemical test-based prediction model for colorectal cancer detection in symptomatic patients. *BMC Med.*, 14, p. 128. doi:<https://doi.org/10.1186/s12916-016-0668-5>
- Detopoulou, P., Panoutsopoulos, G. I., Mantoglou, M., Michailidis, P., Pantazi, I., Papadopoulos, S., & Rojas Gil, A. P. (2023). Relation of Mean Platelet Volume (MPV) with Cancer: A Systematic Review with a Focus on Disease Outcome on Twelve Types of Cancer. *Current oncology (Toronto, Ont.)*, 30(3), 3391–3420.
<https://doi.org/10.3390/curroncol30030258>
- Dong, Y. (2013). Principled missing data methods for researchers. *SpringerPlus*, s. 2(1), 222. doi:
<https://doi.org/10.1186/2193-1801-2-222>
- Dormann, C. F. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, s. 36(1), 27-46.
doi:<https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*(27), pp. 861-878. doi:<https://doi.org/10.1016/j.patrec.2005.10.010>
- Ferlay, J., Colombet, M., Soerjomataram (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer*(144), pp. 1941–1953. doi:10.1002/ijc.31937
- Ferrucci, L., Corsi, A., Lauretani, F., Bandinelli, S., Bartali, B., Taub, D. D., & Guralnik, J. M. (2005). The origins of age-related proinflammatory state. *Blood*, 105(6), 2294-2299.
- García-Jiménez, C., Gutiérrez-Salmerón, M., Chocarro-Calvo, A., García-Martínez, J. M., Castaño, A., & De la Vieja, A. (2016). From obesity to diabetes and cancer: epidemiological links and role of therapies. *British Journal of Cancer*, 114(7), 716-722.

- Geman, S., Bienenstock, E., Doursat, R. (1997). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*(4(1)), pp. 1-58. Retrieved from <https://www.dam.brown.edu/people/geman/Homepage/Essays%20and%20ideas%20about%20neurobiology/bias-variance.pdf>
- Geron, A. (2019). *Hands-on machine learning with scikit-learn, Keras, and tensorflow: Concepts, tools and techniques to build Intelligent Systems*. Beijing ; Boston ; Farnham etc..
- GLOBOCAN. (2020). Retrieved from Uluslararası Kanser Araştırma Ajansı Küresel Kanser Gözlemevi: <https://gco.iarc.fr/>
- Goshen, R., Barak, M., Pini, A., Yaron, K.(2017). Predicting the presence of colon cancer in members of a health maintenance organisation by evaluating analytes from standard laboratory records. *British journal of cancer*(116(7)), pp. 944–950. doi:<https://doi.org/10.1038/bjc.2017.53>
- Gunčar, G., Kukar, M., Nota, M., Brvar, M. (2018). An Application of Machine Learning to Haematological Diagnosis. *Sci. Rep*, pp. 1-12. doi:<https://doi.org/10.1038/s41598-017-18564-8>
- Guo, Y., Li, Z., Du , C., Gong, J., Liao, P., Zhang , J.(2022). *Machine learning for identifying benign and malignant of thyroid tumors: A retrospective study of 2,423 patients*. (Vol. 10). Front Public Health. doi:doi: 10.3389/fpubh.2022.960740. PMID: 3618
- Gupta, S., Tran, S., Luo, T. (2014). Machine-learning prediction of cancer survival: a retrospective study using electronic. *BMJ Open*. doi:<https://doi.org/10.1136/bmjopen-2013-004007>
- Hart, G., Roffman, D., Decker, R., Deng, J. (2019). A multi-parameterized artificial neural network for lung cancer risk prediction. *PLoS One*(13), p. 10. doi: 10.1371/journal.pone.0205264. PMID: 30356283; PMCID: PMC6200229.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall.
- Hilsden, R., Heitman, ,. R., Mizrahi, S., Narod, B. (2018). Prediction of findings at screening colonoscopy using a machine learning algorithm based on complete blood counts (ColonFlag). *PloS one*(,e0207848.), p. 13(11). doi:<https://doi.org/10.1371/journal.pone.0207848>
- Hornbrook, M., Goshen, R., Choman, E. (2017). Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data. *Data. Dig Dis Sci* (62), pp. 2719–2727. doi:<https://doi.org/10.1007/s10620-017-4722-8>

- Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. Vol. 398, John Wiley & Sons.
<https://doi.org/10.1002/9781118548387>
- Huang, S., Yang, J., Fong, S., Zhao, Q. (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer letters*(471), pp. 61–71.
[doi:https://doi.org/10.1016/j.canlet.2019.12.007](https://doi.org/10.1016/j.canlet.2019.12.007)
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95. <https://doi.org/10.1109/MCSE.2007.55>
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.
- Kinar, Y., Kalkstein, N., Akiva, P., Levin, B., Half, E., Goldshtein, I., & Chodick, G. (2016). Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. *Journal of the American Medical Informatics Association : JAMIA*, (23(5)), pp. 879–890.
[doi:https://doi.org/10.1093/jamia/ocv195](https://doi.org/10.1093/jamia/ocv195)
- Konstantina , K., Themis , P., Exarchos, K. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*(13), pp. 8-17. [doi:https://doi.org/10.1016/j.csbj.2014.11.005](https://doi.org/10.1016/j.csbj.2014.11.005)
- Kovalszki, A. &. (2016). Eosinophilia. Primary care. 43(4), 607–617. [doi:https://doi.org/10.1016/j.pop.2016.07.010](https://doi.org/10.1016/j.pop.2016.07.010)
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
<https://doi.org/10.1038/nature14539>
- Lee , E., Jung , S., Hwang , H. (2021). Patient-Level Cancer Prediction Models From a Nationwide Patient Cohort: Model Development and Validation. *JMIR Med Inform*.
[doi:10.2196/29807](https://doi.org/10.2196/29807)
- Li, H., Lin, J., Xiao, Y., Zheng, W., Zhao, L. (2021). Colorectal Cancer Detected by Machine Learning Models Using Conventional Laboratory Test Data. *Technology in cancer research & treatment*(20). [doi:https://doi.org/10.1177/15330338211058352](https://doi.org/10.1177/15330338211058352)
- Lind, M., Vernon, C., Cruickshank, D., Wilkinson, P., Littlewood, T., Stuart, N., Jenkinson, C., Grey-Amante, P., Doll, H., & Wild, D. (2002). The level of haemoglobin in anaemic cancer patients correlates positively with quality of life. *British journal of cancer*, 86(8), 1243–1249. <https://doi.org/10.1038/sj.bjc.6600247>
- Lu, M., Fan, Z., Xu, B., Chen, L., Zheng, X., Li, J., Mi, Q. (2020). Using machine learning to predict ovarian cancer. *International Journal of Medical Informatics*,(141).
[doi:https://doi.org/10.1016/j.ijmedinf.2020.104195](https://doi.org/10.1016/j.ijmedinf.2020.104195)

- Ludwig, H., Van Belle, S., Barrett-Lee, P., Birgegård, G., Bokemeyer, C., Gascón, P., Kosmidis, P., Krzakowski, M., Nortier, J., Olmi, P., Schneider, M., & Schrijvers, D. (2004). The European Cancer Anaemia Survey (ECAS): a large, multinational, prospective survey defining the prevalence, incidence, and treatment of anaemia in cancer patients. *European journal of cancer* (Oxford, England : 1990), 40(15), 2293–2306. <https://doi.org/10.1016/j.ejca.2004.06.019>
- McConnell, B. B., & Yang, V. W. (2009). The Role of Inflammation in the Pathogenesis of Colorectal Cancer. *Current colorectal cancer reports*, 5(2), 69–74. <https://doi.org/10.1007/s11888-009-0011-z>
- McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830
- Plotly Technologies Inc. (2015). *Collaborative Data Science*. Montréal, QC: Plotly Technologies Inc. Retrieved from <https://plotly.com>
- Waskom, M. L. (2021). *Seaborn: Statistical Data Visualization*. Retrieved from <https://seaborn.pydata.org>
- Zorbozan, N., Akarken, I., & Zorbozan, O. (2021). The Performance of the Urine Strip Test for Predicting Microscopic Urine Analysis. *Türk hijyen ve deneysel biyoloji dergisi. Turkish bulletin of hygiene and experimental biology*. <https://doi.org/10.5505/TurkHijyen.2020.98105>

EKLER

1. Çalışmada Kullanılan Hastalıkların ICD-10 Kodları

Hastalık adı	ICD 10 Kodu	Hastalık adı	ICD 10 Kodu	Hastalık adı	ICD 10 Kodu
Dermatofitoz	B35	Periferik vasküler hastalıklar diğer	I73	Omuz lezyonları	M75
Kandidiyaz	B37	Venöz embolizm ve trombozis diğer	I82	Yumuşak doku bozuklukları diğer, başka yerde sınıflanmamış	M79
Demir eksikliği anemisi	D50	Hemoroidler	I84	Osteoporoz, patolojik kırksız	M81
Vitamin B12 eksikliği anemisi	D51	Venlerin diğer bozuklukları	I87	Akut böbrek yetmezliği	N17
Anemi, diğer	D64	Akut nazofarenjit [nezle]	J00	Kronik böbrek yetmezliği	N18
Hipotiroidizm, diğer	E03	Akut sinüzit	J01	Böbrek yetmezlikleri, tanımlanmamış	N19
Toksik olmayan guatr, diğer	E04	Akut farenjit	J02	Böbrek ve üreter taşı	N20
Tiroid bezi diğer bozuklukları	E07	Akut tonsillit	J03	Renal kolik, tanımlanmamış	N23
İnsulin-bağımlı diabetes mellitus	E10	Akut üst solunum yolu enfeksiyonları, birden fazla ve tanımlanmamış yerlerin	J06	Sistit	N30
İnsulin-bağımlı olmayan diabetes mellitus	E11	Bakteriyel pnömoniler, başka yerde sınıflanmamış	J15	Mesenenin nöromusküler fonksiyon bozukluğu, başka yerde sınıflanmamış	N31
Diabetes mellitus, diğer tanımlanmış	E13	Pnömoni, tanımlanmamış organizmalar	J18	Üriner sistemin diğer bozuklukları	N39
Diabetes mellitus, tanımlanmamış	E14	Akut bronşit	J20	Prostat hiperplazisi	N40
Endokrin bozukluklar, diğer	E34	Akut alt solunum yolu enfeksiyon, tanımlanmamış	J22	Memenin diğer bozuklukları	N64
Vitamin D eksikliği	E55	Vazomotor ve allerjik rinit	J30	Vajen ve vulvanın diğer enflamasyonu	N76
Beslenme öğelerinin eksikliği, diğer	E61	Üst solunum yolu diğer hastalıkları	J39	Menstruasyonun yokluğu, azlığı ve seyrekliği	N91
Lipoprotein metabolizması bozuklukları ve diğer lipidemiler	E78	Kronik obstrüktif akciğer hastalığı diğer	J44	Menstruasyonun aşırı, sık ve düzensiz olması	N92
Depresif nöbet	F32	Astım	J45	Kalp atım anormallikleri	R00
Anksiyete bozuklukları, diğer	F41	Solunum yolu yetmezliği, başka yerde sınıflanmamış	J96	Öksürük	R05
Epilepsi	G40	Diş çürükleri	K02	Solunum anormallikleri	R06
Baş ağrısı diğer sendromları	G44	Pulpa ve periapikal dokuların hastalığı	K04	Boğaz ve göğüste ağrı	R07
Serebrovasküler hastalıklarda beyin vasküler sendromları	G46	Gingivit ve periodontal hastalıklar	K05	Abdominal ve pelvik ağrı	R10
Uyku bozuklukları	G47	Dişler ve destekleyici yapıların diğer hastalıkları	K08	Bulantı ve kusma	R11
Polinöropatiler, diğer	G62	Stomatit ve ilgili lezyonlar	K12	Üriner inkontinans tanımlanmamış	R32
Polinöropati, başka yerde sınıflanmış hastalıklarda	G63	Gastro-özefageal reflü hastalığı	K21	Baş dönmesi ve sersemlik	R42
Göz kapağının diğer enflamasyonu	H01	Peptik ülser, yeri tanımlanmamış	K27	Sebebi bilinmeyen ateş	R50
Lakrimal sistem bozuklukları	H04	Gastrit ve duodenit	K29	Baş ağrısı	R51
Konjunktivit	H10	Dispepsi	K30	Ağrı, başka yerde sınıflanmamış	R52
Senil katarakt	H25	Enfektif olmayan diğer gastroenterit ve kolit	K52	Kırgınlık ve yorgunluk	R53

Kataraktlar, diğer	H26	İrritabl barsak sendromu	K58	Senkop ve bayılma	R55
Retinal bozuklukları, diğer	H35	Fonksiyonel barsak bozuklukları diğer	K59	Ödem, başka yerde sınıflanmamış	R60
Glokom	H40	Safra taşı	K80	Genel semptom ve diğer belirtiler	R68
Refraksiyon ve akomodasyon bozuklukları	H52	Kutenöz apse, furonkül ve karbonkül	L02	Akciğer tanısı görüntülemesinde anormal bulgular	R91
Otit eksterna	H60	Derinin ve subkütanöz dokunun diğer lokal enfeksiyonları	L08	Fonksiyon çalışmalarının anormal sonuçları	R94
Dış kulağın diğer bozuklukları	H61	Allerjik kontakt dermatit	L23	Ters etkiler, başka yerde sınıflanmamış	T78
İşitme kaybı, diğer	H91	Pruritus	L29	Düşme, tanımlanmamış	W19
Kulağın diğer bozuklukları, başka yerde sınıflanmış	H93	Dermatit, diğer	L30	Yakınma veya bilinen teşhisi olmayan kişilerin genel muayene ve incelemesi	Z00
Esansiyel (primer) hipertansiyon	I10	Epidermal kalınlaşmalar, diğer	L85	Yakınma veya bilinen teşhisi olmayan kişilerin diğer özel muayene ve incelemeleri	Z01
Angina pectoris	I20	Romatoid artrit, diğer	M06	İdari amaçlar için muayene	Z02
Akut miyokard enfarktüsü	I21	Artrit, diğer	M13	Muayene ve gözlem için diğer nedenler	Z04
Kronik iskemik kalp hastalığı	I25	Gonartroz (diz ekleminin artrozo)	M17	Viral diğer hastalıklara karşı bağışıklama	Z25
Atrial fibrilasyon ve flutter	I48	Artrozlar, diğer	M19	Gebelik durumu	Z33
Kardiyak aritmiler diğer	I49	Eklem diğer bozuklukları, başka yerde sınıflanmamış	M25	Normal gebeliğin gözlemi	Z34
Kalp yetmezliği	I50	Servikal disk bozuklukları	M50	İyileşme dönemi	Z54
Serebrovasküler hastalıklar, diğer	I67	İntervertebral disk bozuklukları, diğer	M51	Sağlık servislerine diğer danışma ve tıbbi tavsiye için gelen kişiler, başka yerde sınıflanmamış	Z71
Periferik vasküler hastalıklar diğer	I73	Dorsalji	M54	Kalp ve damar implant ve greftleri	Z95

2. Kontrol ve Deney Gruplarının Özelliklerine Göre İstatistiksel Analizler

N:209464						Lojistik Regresyon		
Özellik		Kontrol (%n)	Deney (%n)	Etki Büyüklüğü	p	p	Odds Oranı	%95 GA
Dermatofitoz	Var	31413 (%20.6)	11906 (%20.8)	0.00	0.540	<0.001	1.084	1.052 - 1.117
	Yok	120726 (%79.4)	45419 (%79.2)					
Kandidiyaz	Var	11262 (%7.4)	2728 (%4.8)	0.09	<0.001	0.498	0.982	0.933 - 1.034
	Yok	140877 (%92.6)	54597 (%95.2)					
Demir eksikliği anemisi	Var	40897 (%26.9)	9510 (%16.6)	0.22	<0.001	<0.001	0.879	0.853 - 0.907
	Yok	111242 (%73.1)	47815 (%83.4)					
Vitamin B12 eksikliği anemisi	Var	34427 (%22.6)	12273 (%21.4)	0.03	<0.001	<0.001	1.068	1.037 - 1.099
	Yok	117712 (%77.4)	45052 (%78.6)					
Anemi, diğer	Var	60538(%39.8)	13917 (%24.3)	0.29	<0.001	<0.001	0.739	0.719 - 0.759
	Yok	91601 (%60.2)	43408 (%75.7)					

Hipotiroidizm, diğer	Var	27277 (%17.9)	9770 (%17)	0.02	<0.001	0.129	0.975	0.943 - 1.008
	Yok	124862 (%82.1)	47555 (%83)					
Toksik olmayan guatr, diğer	Var	12823 (%8.4)	5077 (%8.9)	0.01	0.002	<0.001	1.15	1.101 - 1.201
	Yok	139316 (%91.6)	52248 (%91.1)					
Tiroid bezi diğer bozuklukları	Var	11483 (%7.5)	4054 (%7.1)	0.02	<0.001	<0.001	1.109	1.058 - 1.162
	Yok	140656 (%92.5)	53271 (%92.9)					
İnsulin-bağımlı diabetes mellitus	Var	18972 (%12.5)	3335 (%5.8)	0.19	<0.001	<0.001	0.701	0.668 - 0.736
	Yok	133167 (%87.5)	53990 (%94.2)					
İnsulin-bağımlı olmayan diabetes mellitus	Var	54729 (%36)	17552 (%30.6)	0.10	<0.001	0.102	0.973	0.942 - 1.005
	Yok	97410 (%64)	39773 (%69.4)					
Diabetes mellitus, diğer tanımlanmış	Var	40258(%26.5)	11877 (%20.7)	0.12	<0.001	<0.001	1.051	1.002 - 1.106
	Yok	111881 (%73.5)	45448 (%79.3)					
Diabetes mellitus, tanımlanmamış	Var	32880 (%21.6)	9646 (%16.8)	0.11	<0.001	0.186	0.897	0.866 - 0.929
	Yok	119259 (%78.4)	47679 (%83.2)					
Endokrin bozukluklar, diğer	Var	9415 (%6.2)	2618 (%4.6)	0.06	<0.001	0.068	0.976	0.942 - 1.012
	Yok	142724 (%93.8)	54707 (%95.4)					
Vitamin D eksikliği	Var	38844 (%25.5)	12760 (%22.3)	0.07	<0.001	0.006	0.951	0.902 - 1.004
	Yok	113295 (%74.5)	44565 (%77.7)					
Beslenme öğelerinin eksikliği, diğer	Var	12480 (%8.2)	3664 (%6.4)	0.06	<0.001	0.783	0.96	0.933 - 0.988
	Yok	139659 (%91.8)	53661 (%93.6)					
Lipoprotein metabolizması bozuklukları ve diğer lipidemiler	Var	54004 (%35.5)	18718 (%32.7)	0.05	<0.001	0.688	1.006	0.961 - 1.054
	Yok	98135 (%64.5)	38607 (%67.3)					
Depresif nöbet	Var	34835 (%22.9)	10113 (%17.6)	0.11	<0.001	0.001	0.949	0.919 - 0.979
	Yok	117304 (%77.1)	47212 (%82.4)					
Anksiyete bozuklukları, diğer	Var	56054 (%36.8)	15865 (%27.7)	0.17	<0.001	<0.001	0.902	0.878 - 0.928
	Yok	96085 (%63.2)	41460 (%72.3)					
Epilepsi	Var	9970 (%6.6)	1819 (%3.2)	0.13	<0.001	<0.001	0.693	0.652 - 0.735
	Yok	142169 (%93.4)	55506 (%96.8)					
Baş ağrısı diğer sendromları	Var	18926 (%12.4)	5247 (%9.2)	0.09	<0.001	0.266	1.022	0.983 - 1.063
	Yok	133213 (%87.6)	52078 (%90.8)					
Serebrovasküler hastalıklarda	Var	8822 (%5.8)	1747 (%3)	0.11	<0.001	<0.001	0.754	0.708 - 0.804

beynin vasküler sendromları								
	Yok	143317 (%94.2)	55578 (%97)					
Uyku bozuklukları	Var	8466 (%5.6)	2451 (%4.3)	0.05	<0.001	0.928	0.997	0.944 - 1.054
	Yok	143673 (%94.4)	54874 (%95.7)					
Polinöropatiler, diğer	Var	20515 (%13.5)	5866 (%10.2)	0.09	<0.001	0.164	0.973	0.935 - 1.011
	Yok	131624 (%86.5)	51459 (%89.8)					
Polinöropati, başka yerde sınıflanmış hastalıklarda	Var	20247 (%13.3)	5324 (%9.3)	0.11	<0.001	<0.001	0.927	0.889 - 0.966
	Yok	131892 (%86.7)	52001 (%90.7)					
Göz kapağının diğer enflamasyonu	Var	9547 (%6.3)	3516 (%6.1)	0.00	0.232	0.23	1.031	0.981 - 1.083
	Yok	142592 (%93.7)	53809 (%93.9)					
Lakrimal sistem bozuklukları	Var	39582 (%26)	14822 (%25.9)	0.00	0.454	0.057	1.031	0.999 - 1.064
	Yok	112557 (%74)	42503 (%74.1)					
Konjunktivit	Var	45911 (%30.2)	16153 (%28.2)	0.04	<0.001	0.686	0.994	0.966 - 1.023
	Yok	106228 (%69.8)	41172 (%71.8)					
Senil katarakt	Var	9450 (%6.2)	3433 (%6)	0.00	0.058	0.952	0.998	0.948 - 1.052
	Yok	142689 (%93.8)	53892 (%94)					
Kataraktlar, diğer	Var	9775 (%6.4)	3134 (%5.5)	0.04	<0.001	0.027	0.942	0.893 - 0.993
	Yok	142364 (%93.6)	54191 (%94.5)					
Retinal bozuklukları, diğer	Var	12399 (%8.1)	3448 (%6)	0.07	<0.001	<0.001	0.869	0.828 - 0.913
	Yok	139740 (%91.9)	53877 (%94)					
Glokom	Var	10519 (%6.9)	3888 (%6.8)	0.00	0.288	0.219	0.97	0.925 - 1.018
	Yok	141620 (%93.1)	53437 (%93.2)					
Refraksiyon ve akomodasyon bozuklukları	Var	48335 (%31.8)	21187 (%37)	0.10	<0.001	<0.001	1.185	1.155 - 1.216
	Yok	103804 (%68.2)	36138 (%63)					
Otit eksterna	Var	10373 (%6.8)	3795 (%6.6)	0.00	0.108	0.012	1.062	1.013 - 1.112
	Yok	141766 (%93.2)	53530 (%93.4)					
Dış kulağın diğer bozuklukları	Var	15085 (%9.9)	6249 (%10.9)	0.03	<0.001	0.004	1.058	1.018 - 1.100
	Yok	137054 (%90.1)	51076 (%89.1)					
İşitme kaybı, diğer	Var	10110 (%6.6)	3647 (%6.4)	0.01	0.020	0.999	1	0.953 - 1.049
	Yok	142029 (%93.4)	53678 (%93.6)					
Kulağın diğer bozuklukları, başka yerde sınıflanmış	Var	8111 (%5.3)	3246 (%5.7)	0.01	0.003	0.059	1.051	0.998 - 1.106

	Yok	144028 (%94.7)	54079 (%94.3)					
Esansiyel (%primer) hipertansiyon	Var	105154 (%69.1)	37600 (%65.6)	0.07	<0.001	0.545	0.99	0.960 - 1.022
	Yok	46985 (%30.9)	19725 (%34.4)					
Angina pectoris	Var	12090 (%7.9)	2668 (%4.7)	0.11	<0.001	0.852	0.995	0.944 - 1.049
	Yok	140049 (%92.1)	54657 (%95.3)					
Akut myokard enfarktüsü	Var	13496 (%8.9)	1878 (%3.3)	0.19	<0.001	<0.001	0.66	0.622 - 0.699
	Yok	138643 (%91.1)	55447 (%96.7)					
Kronik iskemik kalp hastalığı	Var	66718 (%43.9)	18078 (%31.5)	0.23	<0.001	<0.001	0.915	0.886 - 0.944
	Yok	85421 (%56.1)	39247 (%68.5)					
Atrial fibrilasyon ve flutter	Var	23082 (%15.2)	3459 (%6)	0.25	<0.001	<0.001	0.623	0.594 - 0.653
	Yok	129057 (%84.8)	53866 (%94)					
Kardiyak aritmiler diğer	Var	10906 (%7.2)	2298 (%4)	0.12	<0.001	0.107	0.955	0.903 - 1.010
	Yok	141233 (%92.8)	55027 (%96)					
Kalp yetmezliği	Var	33847 (%22.2)	4372 (%7.6)	0.34	<0.001	<0.001	0.583	0.558 - 0.608
	Yok	118292 (%77.8)	52953 (%92.4)					
Serebrovasküler hastalıklar, diğer	Var	25208 (%16.6)	5629 (%9.8)	0.17	<0.001	<0.001	0.736	0.707 - 0.766
	Yok	126931 (%83.4)	51696 (%90.2)					
Periferik vasküler hastalıklar diğer	Var	14722 (%9.7)	3280 (%5.7)	0.13	<0.001	<0.001	0.834	0.795 - 0.875
	Yok	137417 (%90.3)	54045 (%94.3)					
Venöz embolizm ve trombozis diğer	Var	9508 (%6.2)	1424 (%2.5)	0.15	<0.001	<0.001	0.641	0.601 - 0.684
	Yok	142631 (%93.8)	55901 (%97.5)					
Hemoroidler	Var	15519 (%10.2)	4995 (%8.7)	0.04	<0.001	0.001	1.07	1.027 - 1.114
	Yok	136620 (%89.8)	52330 (%91.3)					
Venlerin diğer bozuklukları	Var	23093 (%15.2)	6313 (%11)	0.11	<0.001	0.781	0.995	0.959 - 1.032
	Yok	129046 (%84.8)	51012 (%89)					
Akut nazofarenjit [nezle]	Var	25481 (%16.7)	10555 (%18.4)	0.04	<0.001	<0.001	1.184	1.148 - 1.221
	Yok	126658 (%83.3)	46770 (%81.6)					
Akut sinüzit	Var	30547 (%20.1)	10588 (%18.5)	0.04	<0.001	<0.001	1.071	1.039 - 1.105
	Yok	121592 (%79.9)	46737 (%81.5)					
Akut farenjit	Var	46936 (%32.6)	17222 (%30)	0.05	<0.001	<0.001	1.049	1.021 - 1.076
	Yok	102503 (%67.4)	40103 (%70)					
Akut tonsillit	Var	30336 (%19.9)	8422 (%14.7)	0.12	<0.001	0.035	1.035	1.002 - 1.069
	Yok	121803 (%80.1)	48903 (%85.3)					
Akut üst solunum yolu	Var	70467 (%46.3)	22138 (%38.6)	0.14	<0.001	0.447	1.01	0.985 - 1.034

enfeksiyonları, birden fazla ve tanımlanmamış yerlerin								
	Yok	81672 (%53.7)	35187 (%61.4)					
Bakteriyel pnömoniler, başka yerde sınıflanmamış	Var	17331 (%11.4)	2893 (%5)	0.19	<0.001	<0.001	1.05	1.021 - 1.080
	Yok	134808 (%88.6)	54432 (%95)					
Pnömoni, tanımlanmamış organizmalar	Var	46232 (%30.4)	9353 (%16.3)	0.29	<0.001	<0.001	0.666	0.635 - 0.699
	Yok	105907 (%69.6)	47972 (%83.7)					
Akut bronşit	Var	55335 (%36.4)	16103 (%28.1)	0.16	<0.001	0.037	0.709	0.686 - 0.732
	Yok	96804 (%63.6)	41222 (%71.9)					
Akut alt solunum yolu enfeksiyon, tanımlanmamış	Var	18764 (%12.3)	4752 (%8.3)	0.11	<0.001	<0.001	0.971	0.945 - 0.998
	Yok	133375 (%87.7)	52573 (%91.7)					
Vazomotor ve allerjik rinit	Var	41225 (%27.1)	14666 (%25.6)	0.03	<0.001	<0.001	0.922	0.885 - 0.959
	Yok	110914 (%72.9)	42659 (%74.4)					
Üst solunum yolu diğer hastalıkları	Var	47050 (%30.9)	14034 (%24.5)	0.13	<0.001	0.333	1.118	1.086 - 1.151
	Yok	105089 (%69.1)	43291 (%75.5)					
Kronik obstrüktif akciğer hastalığı diğer	Var	40467 (%26.6)	9865 (%17.2)	0.20	<0.001	<0.001	0.987	0.961 - 1.014
	Yok	111672 (%73.4)	47460 (%82.8)					
Astım	Var	47207 (%31)	12115 (%21.1)	0.20	<0.001	0.002	0.91	0.879 - 0.944
	Yok	104932 (%69)	45210 (%78.9)					
Solunum yolu yetmezliği, başka yerde sınıflanmamış	Var	10055 (%6.6)	653 (%1.1)	0.22	<0.001	<0.001	0.95	0.920 - 0.981
	Yok	142084 (%93.4)	56672 (%98.9)					
Diş çürükleri	Var	40987 (%26.9)	15153 (%26.4)	0.01	0.002	<0.001	1.09	1.060 - 1.122
	Yok	111152 (%73.1)	42172 (%73.6)					
Pulpa ve periapikal dokuların hastalığı	Var	21767 (%14.3)	7913 (%13.8)	0.01	0.003	<0.001	1.096	1.057 - 1.135
	Yok	130372 (%85.7)	49412 (%86.2)					
Gingivitis ve periodontal hastalıklar	Var	15779 (%10.4)	6218 (%10.8)	0.001	0.002	<0.001	1.086	1.046 - 1.129
	Yok	136360 (%89.6)	51107 (%89.2)					
Dişler ve destekleyici yapıların diğer hastalıkları	Var	14340 (%9.4)	6601 (%11.5)	0.06	<0.001	<0.001	1.152	1.108 - 1.199
	Yok	137899 (%90.6)	50724 (%88.5)					
Stomatit ve ilgili lezyonlar	Var	11450 (%7.5)	3827 (%6.7)	0.03	<0.001	0.774	0.993	0.949 - 1.040

	Yok	140689 (%92.5)	53498 (%93.3)					
Gastro-özefageal reflü hastalığı	Var	110732 (%72.8)	36226 (%63.2)	0.19	<0.001	<0.001	0.862	0.839 - 0.886
	Yok	41407 (%27.2)	21099 (%36.8)					
Peptik ülser, yeri tanımlanmamış	Var	53602 (%35.2)	15540 (%27.1)	0.16	<0.001	0.002	0.958	0.933 - 0.984
	Yok	98537 (%64.8)	41785 (%72.9)					
Gastrit ve duodenit	Var	51897 (%34.1)	14077 (%24.6)	0.18	<0.001	<0.001	0.907	0.883 - 0.932
	Yok	100242 (%65.9)	43248 (%75.4)					
Dispepsi	Var	48035 (%31.6)	13720 (%23.9)	0.15	<0.001	<0.001	0.94	0.914 - 0.966
	Yok	104104 (%68.4)	43605 (%76.1)					
Enfektif olmayan diğer gastroenterit ve kolit	Var	35992 (%23.7)	8913 (%15.5)	0.18	<0.001	<0.001	0.882	0.856 - 0.910
	Yok	116147 (%76.3)	48412 (%84.5)					
İrritabl barsak sendromu	Var	26147 (%17.2)	7402 (%12.9)	0.10	<0.001	0.093	1.03	0.995 - 1.066
	Yok	125992 (%82.8)	49923 (%87.1)					
Fonksiyonel barsak bozuklukları diğer	Var	43428 (%28.5)	11046 (%19.3)	0.19	<0.001	<0.001	0.861	0.836 - 0.887
	Yok	108711 (%71.5)	46279 (%80.7)					
Safra taşı	Var	12851 (%8.4)	2601 (%4.5)	0.13	<0.001	<0.001	0.653	0.621 - 0.686
	Yok	139288 (%91.6)	54724 (%95.5)					
Kutenöz apse, furonkül ve karbonkül	Var	9216 (%6.1)	2592 (%4.5)	0.06	<0.001	0.003	0.924	0.876 - 0.974
	Yok	142923 (%93.9)	54733 (%95.5)					
Derinin ve subkutenöz dokunun diğer lokal enfeksiyonları	Var	22774 (%15)	6166 (%10.8)	0.11	<0.001	<0.001	0.87	0.839 - 0.902
	Yok	129365 (%85)	51159 (%89.2)					
Allerjik kontakt dermatit	Var	7979 (%5.2)	2696 (%4.7)	0.02	<0.001	0.056	0.948	0.898 - 1.001
	Yok	144160 (%94.8)	54629 (%95.3)					
Pruritus	Var	12671 (%8.3)	3477 (%6.1)	0.08	<0.001	<0.001	0.913	0.870 - 0.959
	Yok	139468 (%91.7)	53848 (%93.9)					
Dermatit, diğer	Var	53177 (%35)	18521 (%32.3)	0.05	<0.001	<0.001	1.073	1.045 - 1.101
	Yok	98962 (%65)	38804 (%67.7)					
Epidermal kalınlaşmalar, diğer	Var	15207 (%10)	4845 (%8.5)	0.05	<0.001	0.602	0.989	0.947 - 1.032
	Yok	136932 (%90)	52480 (%91.5)					
Romatoid artrit, diğer	Var	9707 (%6.4)	2170 (%3.8)	0.10	<0.001	<0.001	0.583	0.551 - 0.616
	Yok	142432 (%93.6)	55155 (%96.2)					
Artrit, diğer	Var	23174 (%15.2)	7873 (%13.7)	0.04	<0.001	0.314	0.983	0.950 - 1.017

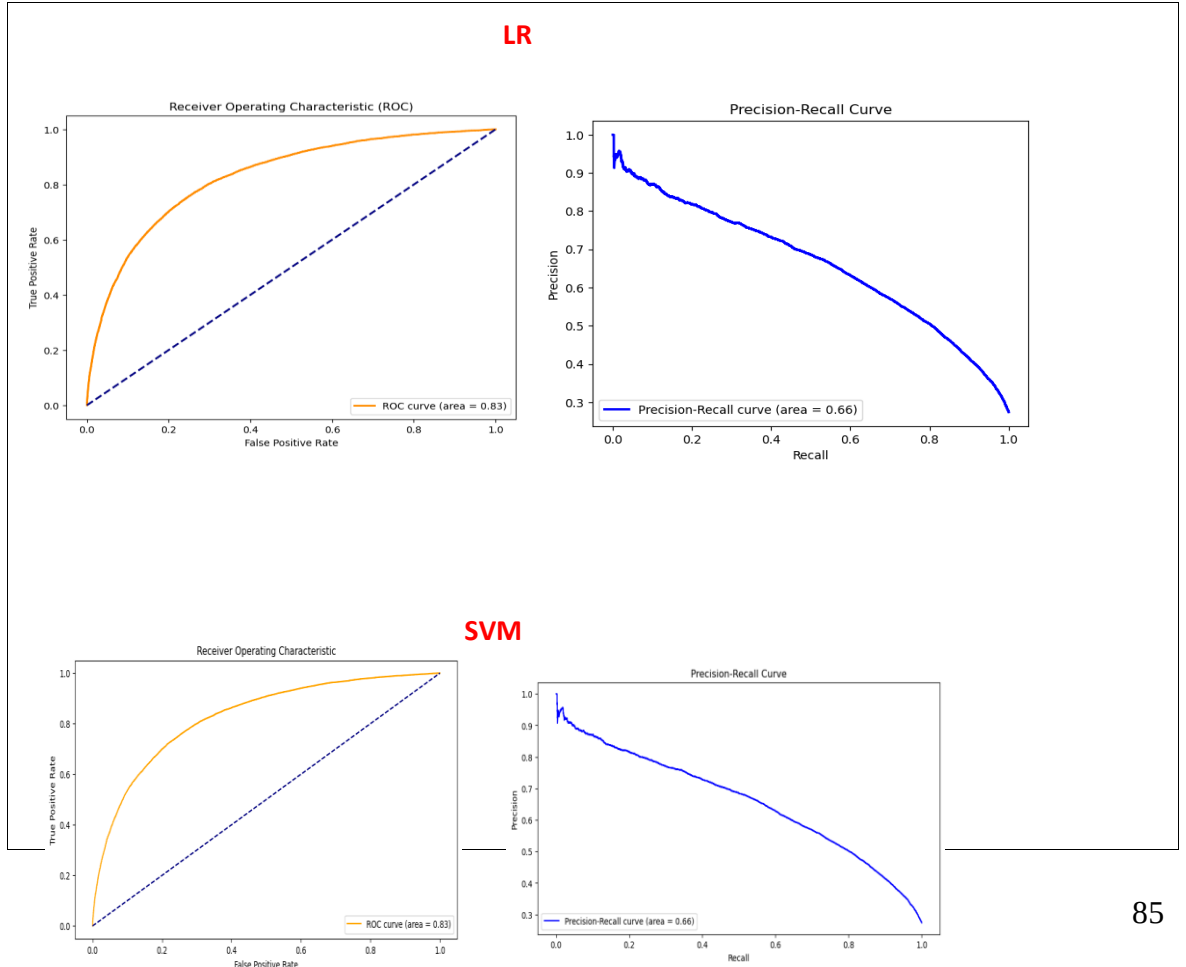
	Yok	128965 (%84.8)	49452 (%86.3)					
Gonartroz (%diz eklemının artrazo)	Var	26290 (%17.3)	10251 (%17.9)	0.01	0.001	0.067	1.031	0.998 - 1.065
	Yok	125849 (%82.7)	47074 (%82.1)					
Artrozlar, diğér	Var	12600 (%8.3)	4999 (%8.7)	0.01	0.001	0.461	1.016	0.974 - 1.060
	Yok	139539 (%91.7)	52326 (%91.3)					
Eklem diğér bozuklukları, başka yerde sınıflanmamış	Var	55002 (%36.2)	17741 (%30.9)	0.10	<0.001	<0.001	0.884	0.861 - 0.908
	Yok	97137 (%63.8)	39584 (%69.1)					
Servikal disk bozuklukları	Var	12450 (%8.2)	4479 (%7.8)	0.01	0.006	0.057	1.043	0.999 - 1.090
	Yok	139689 (%91.8)	52846 (%92.2)					
İntervertebral disk bozuklukları, diğér	Var	25229 (%16.6)	9189 (%16)	0.01	0.002	<0.001	1.066	1.031 - 1.102
	Yok	126910 (%83.4)	48136 (%84)					
Dorsalji	Var	61550 (%40.5)	22097 (%38.5)	0.03	<0.001	<0.001	1.051	1.025 - 1.078
	Yok	90589 (%59.5)	35228 (%61.5)					
Omuz lezyonları	Var	10152 (%6.7)	4062 (%7.1)	0.01	0.001	0.001	1.078	1.030 - 1.128
	Yok	141987 (%93.3)	53263 (%92.9)					
Yumuşak doku bozuklukları diğér, başka yerde sınıflanmamış	Var	119859 (%78.8)	41248 (%72)	0.14	<0.001	<0.001	0.925	0.899 - 0.951
	Yok	32280 (%21.2)	16077 (%28)					
Osteoporoz, patolojik kırksız	Var	22864 (%15)	7464 (%13)	0.05	<0.001	<0.001	0.851	0.821 - 0.882
	Yok	129275 (%85)	49861 (%87)					
Akut böbrek yetmezliğı	Var	13514 (%8.9)	745 (%1.3)	0.27	<0.001	<0.001	0.363	0.334 - 0.395
	Yok	138625 (%91.1)	56580 (%98.7)					
Kronik böbrek yetmezliğı	Var	25273 (%16.6)	2333 (%4.1)	0.34	<0.001	<0.001	0.325	0.309 - 0.343
	Yok	126866 (%83.4)	54992 (%95.9)					
Böbrek yetmezlikleri, tanımlanmamış	Var	10843 (%7.1)	935 (%1.6)	0.21	<0.001	<0.001	0.563	0.520 - 0.609
	Yok	141296 (%92.9)	56390 (%98.4)					
Böbrek ve üreter taşı	Var	11535 (%7.6)	2782 (%4.9)	0.10	<0.001	<0.001	0.887	0.842 - 0.934
	Yok	140604 (%92.4)	54543 (%95.1)					
Renal kolik, tanımlanmamış	Var	22576 (%14.8)	5132 (%9)	0.16	<0.001	0.01	0.95	0.913 - 0.988
	Yok	129563 (%85.2)	52193 (%91)					
Sistit	Var	43891 (%28.8)	10327 (%18)	0.22	<0.001	<0.001	0.843	0.818 - 0.868
	Yok	108248 (%71.2)	46998 (%82)					

Mesenenin nöromusküler fonksiyon bozukluğu, başka yerde sınıflanmamış	Var	14665 (%9.6)	3418 (%6)	0.12	<0.001	<0.001	0.814	0.776 - 0.854
	Yok	137474 (%90.4)	53907 (%94)					
Üriner sistemin diğer bozuklukları	Var	74040 (%48.7)	17496 (%30.5)	0.33	<0.001	<0.001	0.738	0.719 - 0.758
	Yok	78099 (%51.3)	39829 (%69.5)					
Prostat hiperplazisi	Var	24204 (%15.9)	10860 (%18.9)	0.07	<0.001	<0.001	1.397	1.344 - 1.452
	Yok	127935 (%84.1)	46465 (%81.1)					
Memenin diğer bozuklukları	Var	10855 (%7.1)	4209 (%7.3)	0.00	0.101	<0.001	1.207	1.154 - 1.262
	Yok	141284 (%92.9)	53116 (%92.7)					
Vajen ve vulvanın diğer enflamasyonu	Var	19633 (%12.9)	4717 (%8.2)	0.13	<0.001	<0.001	1.104	1.057 - 1.153
	Yok	132506 (%87.1)	52608 (%91.8)					
Menstruasyonun yokluğu, azlığı ve seyrekliği	Var	9450 (%6.2)	1435 (%2.5)	0.15	<0.001	0.014	0.917	0.856 - 0.982
	Yok	142689 (%93.8)	55890 (%97.5)					
Menstruasyonun aşırı, sık ve düzensiz olması	Var	15396 (%10.1)	3228 (%5.6)	0.14	<0.001	0.274	0.972	0.925 - 1.022
	Yok	136743 (%89.9)	54097 (%94.4)					
Kalp atım anormallikleri	Var	30869 (%20.3)	6815 (%11.9)	0.20	<0.001	<0.001	0.88	0.850 - 0.911
	Yok	121270 (%79.7)	50510 (%88.1)					
Öksürük	Var	36456 (%24)	9896 (%17.3)	0.14	<0.001	0.105	1.026	0.995 - 1.058
	Yok	115683 (%76)	47429 (%82.7)					
Solunum anormallikleri	Var	42545 (%28)	7219 (%12.6)	0.33	<0.001	<0.001	0.684	0.661 - 0.708
	Yok	109594 (%72)	50106 (%87.4)					
Boğaz ve göğüste ağrı	Var	77480 (%50.9)	18680 (%32.6)	0.33	<0.001	<0.001	0.738	0.718 - 0.757
	Yok	74659 (%49.1)	38645 (%67.4)					
Abdominal ve pelvik ağrı	Var	87738 (%57.7)	19399 (%33.8)	0.44	<0.001	<0.001	0.64	0.624 - 0.656
	Yok	64401 (%42.3)	37926 (%66.2)					
Bulantı ve kusma	Var	51547 (%33.9)	9456 (%16.5)	0.35	<0.001	<0.001	0.718	0.697 - 0.740
	Yok	100592 (%66.1)	47869 (%83.5)					
Üriner inkontinans tanımlanmamış	Var	13597 (%8.9)	2861 (%5)	0.13	<0.001	<0.001	0.713	0.678 - 0.751
	Yok	138542 (%91.1)	54464 (%95)					
Baş dönmesi ve sersemlik	Var	45476 (%29.9)	13409 (%23.4)	0.13	<0.001	<0.001	0.944	0.918 - 0.971
	Yok	106663 (%70.1)	43916 (%76.6)					
Sebebi bilinmeyen ateş	Var	24852 (%16.3)	3565 (%6.2)	0.27	<0.001	<0.001	0.613	0.588 - 0.639
	Yok	127287 (%83.7)	53760 (%93.8)					

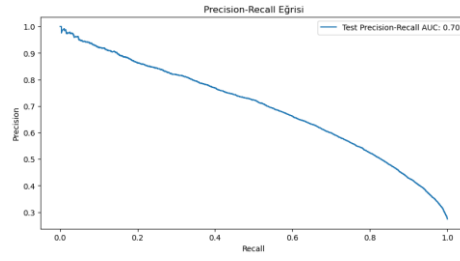
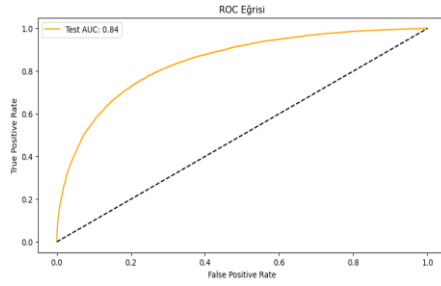
Baş ağrısı	Var	52877 (%34.8)	14679 (%25.6)	0.18	<0.001	<0.001	0.954	0.929 - 0.980
	Yok	99262 (%65.2)	42646 (%74.4)					
Ağrı, başka yerde sınıflanmamış	Var	67219 (%44.2)	19091 (%33.3)	0.20	<0.001	<0.001	0.833	0.813 - 0.854
	Yok	84920 (%55.8)	38234 (%66.7)					
Kırgınlık ve yorgunluk	Var	38811 (%25.5)	9376 (%16.4)	0.19	<0.001	<0.001	0.806	0.783 - 0.830
	Yok	113328 (%74.5)	47949 (%83.6)					
Senkop ve bayılma	Var	9937 (%6.5)	1555 (%2.7)	0.15	<0.001	<0.001	0.686	0.644 - 0.731
	Yok	142202 (%93.5)	55770 (%97.3)					
Ödem, başka yerde sınıflanmamış	Var	9175 (%6)	1523 (%2.7)	0.14	<0.001	<0.001	0.788	0.739 - 0.841
	Yok	142964 (%94)	55802 (%97.3)					
Genel semptom ve diğer belirtiler	Var	16652 (%10.9)	3953 (%6.9)	0.12	<0.001	<0.001	0.807	0.773 - 0.842
	Yok	135487 (%89.1)	53372 (%93.1)					
Akciğer tanısı görüntülemesinde anormal bulgular	Var	10636 (%7)	2471 (%4.3)	0.10	<0.001	<0.001	0.889	0.843 - 0.938
	Yok	141503 (%93)	54854 (%95.7)					
Fonksiyon çalışmalarının anormal sonuçları	Var	11695 (%7.7)	1752 (%3.1)	0.17	<0.001	<0.001	0.68	0.641 - 0.721
	Yok	140444 (%92.3)	55573 (%96.9)					
Ters etkiler, başka yerde sınıflanmamış	Var	26712 (%17.6)	8719 (%15.2)	0.06	<0.001	0.022	0.963	0.932 - 0.995
	Yok	125427 (%82.4)	48602 (%84.8)					
Düşme, tanımlanmamış	Var	11514 (%7.6)	2457 (%4.3)	0.12	<0.001	<0.001	0.711	0.675 - 0.749
	Yok	140625 (%92.4)	54868 (%95.7)					
Yakınma veya bilinen teşhisi olmayan kişilerin genel muayene ve incelemesi	Var	129222 (%84.9)	40118 (%70)	0.34	<0.001	<0.001	0.651	0.633 - 0.669
	Yok	22917 (%15.1)	17207 (%30)					
Yakınma veya bilinen teşhisi olmayan kişilerin diğer özel muayene ve incelemeleri	Var	69130 (%45.4)	21492 (%37.5)	0.14	<0.001	<0.001	0.795	0.775 - 0.815
	Yok	83009 (%54.6)	35833 (%62.5)					
İdari amaçlar için muayene	Var	19646 (%12.9)	5444 (%9.5)	0.09	<0.001	<0.001	0.848	0.816 - 0.881
	Yok	132493 (%87.1)	51881 (%90.5)					
Muayene ve gözlem için diğer nedenler	Var	16903 (%11.1)	3579 (%6.2)	0.15	<0.001	<0.001	0.717	0.687 - 0.749
	Yok	135236 (%88.9)	53746 (%93.8)					
Viral diğer hastalıklara karşı bağışıklama	Var	13597 (%8.9)	5081 (%8.9)	0.00	0.598	<0.001	1.136	1.089 - 1.185
	Yok	138542 (%91.1)	52244 (%91.1)					

Gebelik durumu	Var	17623 (%11.6)	1366 (%2.4)	0.29	<0.001	<0.001	0.316	0.293 - 0.341
	Yok	134516 (%88.4)	55959 (%97.6)					
Normal gebeliğin gözlemi	Var	11297 (%7.4)	803 (%1.4)	0.23	<0.001	<0.001	0.589	0.537 - 0.647
	Yok	140842 (%92.6)	56522 (%98.6)					
İyileşme dönemi	Var	10420 (%6.8)	2296 (%4)	0.11	<0.001	<0.001	0.692	0.656 - 0.730
	Yok	141719 (%93.2)	55029 (%96)					
Sağlık servislerine diğer danışma ve tıbbi tavsiye için gelen kişiler, başka yerde sınıflanmamış	Var	7946 (%5.2)	2492 (%4.3)	0.04	<0.001	0.776	0.992	0.939 - 1.048
	Yok	144193 (%94.8)	54833 (%95.7)					
Kalp ve damar implant ve greftleri	Var	14240 (%9.4)	1894 (%3.3)	0.20	<0.001	<0.001	0.392	0.37 - 0.416
	Yok	137899 (%90.6)	55431 (%96.7)					

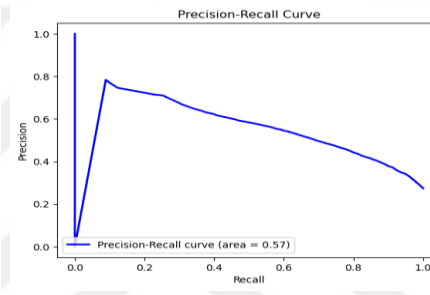
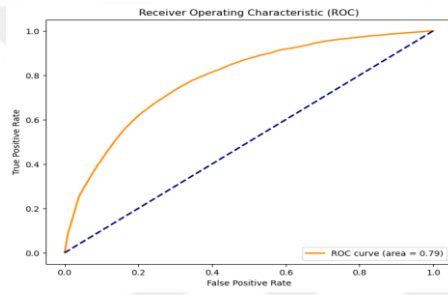
3. Makine Öğrenmesi Modellerinin ROC ve Precision-Recall Curve Grafikleri



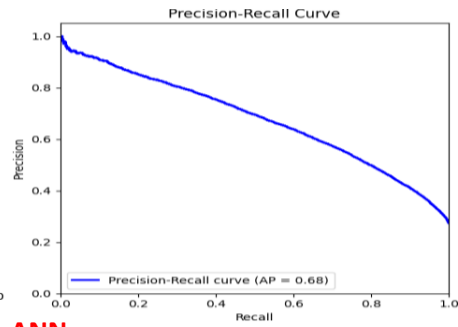
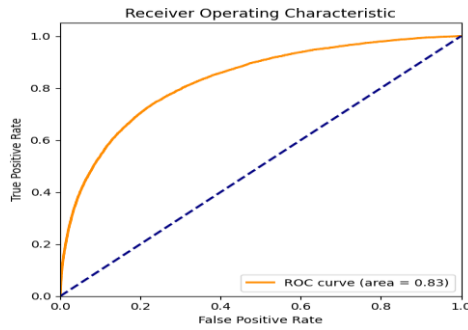
XGBOOST



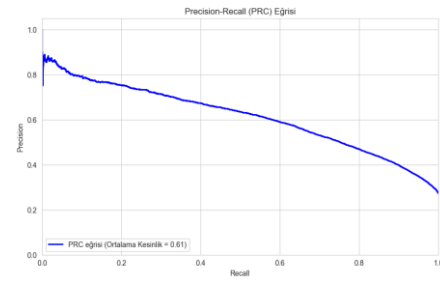
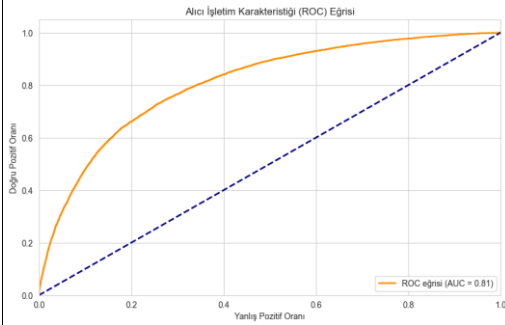
DT



RF

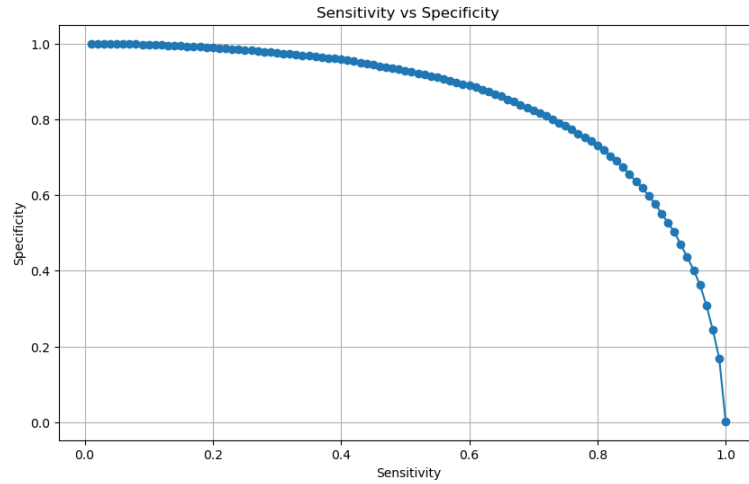


ANN



4. XGBoost Modelinin Farklı Eşik Değerlerinde Sensitivite ve Spesifisite Değerleri

Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
0.01	1	0.26	0.98	0.51	0.92	0.76	0.77
0.02	1	0.27	0.98	0.52	0.92	0.77	0.76
0.03	1	0.28	0.98	0.53	0.92	0.78	0.75
0.04	1	0.29	0.98	0.54	0.91	0.79	0.74
0.05	1	0.3	0.98	0.55	0.91	0.8	0.73
0.06	1	0.31	0.97	0.56	0.91	0.81	0.72
0.07	1	0.32	0.97	0.57	0.9	0.82	0.7
0.08	1	0.33	0.97	0.58	0.9	0.83	0.69
0.09	1	0.34	0.97	0.59	0.89	0.84	0.67
0.1	1	0.35	0.97	0.6	0.89	0.85	0.65
0.11	1	0.36	0.97	0.61	0.88	0.86	0.64
0.12	1	0.37	0.96	0.62	0.88	0.87	0.62
0.13	1	0.38	0.96	0.63	0.87	0.88	0.6
0.14	0.99	0.39	0.96	0.64	0.87	0.89	0.58
0.15	0.99	0.4	0.96	0.65	0.86	0.9	0.55
0.16	0.99	0.41	0.96	0.66	0.85	0.91	0.53
0.17	0.99	0.42	0.95	0.67	0.85	0.92	0.5
0.18	0.99	0.43	0.95	0.68	0.84	0.93	0.47
0.19	0.99	0.44	0.95	0.69	0.83	0.94	0.44
0.2	0.99	0.45	0.94	0.7	0.82	0.95	0.4
0.21	0.99	0.46	0.94	0.71	0.82	0.96	0.36
0.22	0.99	0.47	0.94	0.72	0.81	0.97	0.31
0.23	0.99	0.48	0.94	0.73	0.8	0.98	0.24
0.24	0.98	0.49	0.93	0.74	0.79	0.99	0.17
0.25	0.98	0.5	0.93	0.75	0.78	1	0



ÖZGEÇMİŞ

Kişisel Bilgiler

Adı	Esra	Uyruğu	TC
Soyadı	Tokur Sonuvar	Tel no	
Doğum tarihi		e-posta	

Eğitim Bilgileri

Mezun olduğu kurum		Mezuniyet yılı
Lisans, İstatistik	Hacettepe Üniversitesi	2004
Yüksek Lisans, İstatistik	Yıldız Teknik Üniversitesi	2007

İş Deneyimi

Görevi	Kurum	Süre (yıl-yıl)
İstatistik Uzmanı	Hava Kuvvetleri Komutanlığı	2004-2021
İstatistik Uzmanı	Akdeniz Üniversitesi	2021-devam ediyor