Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi
Afyon Kocatepe University – Journal of Science and Engineering
https://dergipark.org.tr/tr/pub/akufemubid

Araştırma Makalesi / Research Article
DOI: https://doi.org/ 10.35414/akufemubid.1491469
AKÜ FEMÜBİD **24** (2024) 065104 (1364-1381)
AKU J. Sci. Eng. **24** (2024) 065104 (1364-1381)

Sağlık İstatistiklerinin Veri Madenciliği Teknikleri İle Analizi: Makine Öğrenmesi Algoritmaları Kullanılarak Genel Sağlık Durumunun Sınıflandırılması

Makale Bilgisi / Article Info
Alındı/Received: 29.05.2024
Kabul/Accepted: 31.08.2024
Yayımlandı/Published: xx.xx.xxxx

Analysis of Health Statistics With Data Mining Techniques: Classification of General Health Status Using Machine Learning Algorithms

Yunus Emre GÜR[1], Kamil Abdullah EŞİDİR[2], Ahmed İhsan ŞİMŞEK[3]*

[1] *Fırat Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü, Elazığ, Türkiye*
[2] *Fırat Kalkınma Ajansı, Elazığ, Türkiye*
[3] *Fırat Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü, Elazığ, Türkiye*

Öz

Sağlık sektörü, günümüzde devasa veri yığınlarıyla başa çıkmak zorundadır. Bu verilerin derinliklerindeki bilgileri çözümleyerek hastalıkları daha iyi anlama ve sağlık hizmetlerini geliştirme gibi birçok amaç doğrultusunda veri madenciliği metodolojileri kullanılmaktadır. Bu çalışma, bir veri madenciliği sınıflandırma tekniği olan Gradient Boosting'in, mikro sağlık verilerini nasıl kategorize edebileceğini araştırmaktadır. Çalışmada, Türkiye İstatistik Kurumu'ndan (TÜİK) elde edilen 2022 yılına ait mikro veriler kullanılmıştır. Araştırmada kullanılan soru formundan elde edilen 9 adet bağımsız değişken, analizlerde kullanılarak sağlık durum tespiti tahmin edilmiştir. Ayrıca, çeşitli sosyo-demografik faktörlerin (yaş, cinsiyet, medeni ve çalışma durumu) ve yaşam tarzı alışkanlıklarının (tütün kullanımı) genel sağlık üzerindeki etkileri analiz edilmiştir. Çalışmanın sonuçları, makine öğrenmesi metotlarının sağlık sektöründe ne denli etkili olabileceğini göstermektedir. Bu modeller arasında Gradient Boosting modeli, sınıflandırma performansında, doğruluk, hassasiyet, duyarlılık ve F1 skoru gibi metrikler dikkate alınarak yapılan incelemede ön plana çıkarak, sağlık politikalarının ve müdahalelerinin geliştirilmesine katkıda bulunacak önemli bilgiler sunmuştur. Özellikle, tütün kullanımının sağlık üzerindeki olumsuz etkilerinin belirlenmesi, bu faktörlerin insan sağlığına etkisinin büyük olduğunu ortaya koymaktadır. Bu bulgular, sağlık politikaları ve halk sağlığı programlarının geliştirilmesinde makine öğrenmesinin önemli faydalar sağlayabileceğine işaret etmektedir.

Anahtar Kelimeler: *Veri Madenciliği, Sınıflandırma, Mikro Sağlık Verileri, Makine Öğrenmesi, Gradient Boosting*

Abstract

The healthcare industry today has to deal with huge piles of data. Data mining methodologies are used for many purposes such as better understanding diseases and improving health services by analyzing the information in the depths of these data. This study investigates how Gradient Boosting, a data mining classification technique, can categorize micro health data. In the study, micro data for 2022 obtained from the Turkish Statistical Institute (TUIK) was used. Health status determination was estimated by using 9 independent variables obtained from the questionnaire used in the research in the analyses. Additionally, the effects of various socio-demographic factors (age, gender, marital and employment status) and lifestyle habits (tobacco use) on general health were analyzed. The results of the study show how effective machine learning methods can be in the healthcare sector. Among these models, the Gradient Boosting model came to the fore in the analysis of classification performance, taking into account metrics such as accuracy, sensitivity, sensitivity and F1 score, providing important information that will contribute to the development of health policies and interventions. In particular, determining the negative effects of tobacco use on health reveals that these factors have a great impact on human health. These findings indicate that machine learning can provide significant benefits in the development of health policies and public health programs.

Keywords: *Data Mining, Classification, Microhealth Data, Machine Learning, Gradient Boosting*

1. Giriş

Teknolojik gelişmeler ve büyüyen veri hacmi, insanların verileri daha anlamlı bir şekilde kullanma arzusunu artırmıştır. Bu bağlamda, veri madenciliği kavramı birçok kişi tarafından çeşitli şekillerde tanımlanabilir. Genel bir ifadeyle, veri madenciliği; depolanan verileri matematiksel ve istatistiksel yöntemler kullanarak işleyip, değerli bilgiler elde etme sürecidir (Altınbaş, 2010). Günümüzde sağlık sektörü, büyük miktarda veriyi ele almaktadır ve bu verilerden anlam çıkarmak, hastalıkları anlamak ve sağlık hizmetlerini iyileştirmek için veri madenciliği teknikleri kullanılmaktadır. Sağlık istatistikleri, hasta kayıtları, tedavi sonuçları ve genetik veriler gibi geniş veri setlerini içermektedir. Bu verilerin etkili bir şekilde analiz edilmesi, sağlık sistemlerinin daha iyi yönetilmesine ve hastaların daha etkili bir şekilde tedavi edilmesine olanak tanımaktadır (Terzi, 2019). Bununla

*Sorumlu Yazar/Corresponding Author: Ahmed İhsan Şimşek e-posta/e-mail: aisimsek@firat.edu.tr

birlikte, sağlık sektöründe veri madenciliğine yönelik çalışmalar gün geçtikçe artmaktadır. Veri madenciliği, bilgi keşfini beş temel metodoloji üzerinden gerçekleştirmektedir. Bunlar, genel çıkarımlar, ilişkisel kurallar, sınıflandırma ve kümeleme teknikleri, tahmin edici algoritmalar ve aykırı değer analizleri şeklindedir (Koçak ve Ergün, 2023).

Veri tabanlı bilgiler, sağlık sistemi politikalarının ve yönetim kararlarının temelini oluşturmaktadır. Sağlık politikalarının ve kararlarının hedeflerine ulaşması ve etkili olması, güvenilir, güncel ve doğru veriye dayanmaktadır. Sağlık bilgi sistemlerinin ana amacı, büyük miktardaki sağlık verilerini işleyerek kullanışlı bilgiler elde etmektir. Elde edilen bu bilgiler, hastalar için daha kaliteli sağlık hizmetleri sunumu, sağlık kuruluşlarının daha etkili yönetimi, kaynakların verimli kullanımı ve sağlık politikalarının geliştirilmesi için kullanılır. Sağlık verileri, hastaneler, diğer sağlık kuruluşları, sigorta firmaları ve ilgili devlet daireleri gibi çeşitli kurumlar tarafından toplanır. Dijital verilerin miktarındaki artış, yeni zorluklar doğurmuştur. Bunlar arasında, büyük, çok boyutlu ve karmaşık verileri işleyebilecek yöntem ve sistemler geliştirmek; yeni veri türlerini işlemek için metodolojiler ve sistemler oluşturmak; dağınık verileri işlemek için yöntemler, protokoller ve altyapılar kurmak; ve veri kullanımı ve güvenliğiyle ilgili modeller oluşturmak yer almaktadır (Koyuncugil ve Özgülbaş, 2009).

Bununla birlikte, bu çalışma kullanılan veri setinin temelini oluşturan ve TÜİK tarafından gerçekleştirilen Türkiye Sağlık Araştırması, ilk defa 2008 yılında yapılmıştır. Araştırma, bebeklerden yetişkinlere kadar geniş bir yaş grubunu kapsamaktadır. Bireylerin sağlık durumlarına ilişkin önemli veriler toplanmaktadır. 15 yaş ve üzeri bireylerin sağlık hizmetlerinden faydalanma durumları, günlük aktivitelerinde yaşadıkları zorluk dereceleri, sigara ve alkol kullanma alışkanlıkları gibi çeşitli sağlık göstergeleri araştırılmaktadır. Araştırma sayesinde, toplumun genel sağlık durumunun değerlendirmesi ve sağlık politikalarının yönlendirilmesine katkı sunulması hedeflenmektedir (Genç ve Kurutkan, 2021). Bu çalışmada kullanılan 2022 Yılı Türkiye Sağlık Araştırması verileri, ülke genelinde toplamda 11.170 adet haneden elde edilmiştir. Yapılan araştırma, ülkelerin gelişmişlik düzeylerini gösteren kalkınma göstergeleri içerisinde önemli paya sahip olan, sağlık göstergelerine ait bilgilerin elde edilmesine olanak tanımıştır. Araştırma, sadece ülke genelini yansıtmamakta, aynı zamanda uluslararası karşılaştırmalara olanak tanımaktadır. Araştırmanın kapsamı, Türkiye sınırları içinde yer alan tüm yerleşim yerlerinde bulunan hane halklarıdır.

Bu çalışmanın temel amacı, makine öğrenmesi gibi veri madenciliği sınıflandırma tekniklerinin, mikro sağlık verilerini nasıl kategorize edebileceğini araştırmaktır. Bu bağlamda, Türkiye İstatistik Kurumu (TÜİK) tarafından sağlanan 2022 yılına özgü mikro veri seti bu çalışmada kullanılmıştır. Kullanılan anket formundan elde edilen dokuz bağımsız değişken, sağlık durumu tahminlerinin yapılmasında analiz edilmiştir. Sonuçta, bu araştırma makine öğrenmesi sınıflandırma algoritmalarının sağlık istatistiklerinin detaylı analizinde ve politika oluşturma sürecinde ne kadar değerli bir araç olabileceğini göstermektedir. Bu tekniklerin kullanılması, sağlık sektörü analizlerinde önemli bir evrimi temsil ederek, daha bilinçli ve amaç odaklı politika kararları alınmasını mümkün kılmaktadır.

Çalışmanın ilerleyen bölümlerinde, sınıflandırma konusunda yapılan literatür taramasının ardından, çalışmanın metodolojisi ayrıntılı bir şekilde ele alınmıştır. Bu bölümde, analiz için kullanılan veri setinin nitelikleri, tercih edilen sınıflandırma metotlarının özellikleri ve yapılan analizin teknik detayları detaylandırılmıştır. Sonrasında, araştırmadan elde edilen sonuçlar, sağlık sektörünün güncel durumunu ve gelişmekte olan eğilimleri açıklayan kıymetli bilgilerle birlikte irdelenmiştir. Makalenin son kısmında, bu sonuçların sağlık politikalarına ve sektör analizlerine olan etkileri üzerine değerlendirmeler yapılmış ve gelecek çalışmalar için öneriler sunulmuştur.

2. Literatür Taraması

Çalışmanın bu bölümümde, sağlık sektöründe yapılan sınıflandırma çalışmaları ile ilgili literatür incelenmiştir. Bu inceleme sonucu elde edilen çalışmalar aşağıda sunulmuştur. Karakoyun ve Hacıbeyoğlu'nun (2014) çalışmasında, biyomedikal veri kümeleri kullanılarak altı farklı makine öğrenmesi sınıflandırma algoritmasının performansını test edilmiş ve sonuçlar istatistiksel olarak karşılaştırılmıştır. Çalışmada, Yapay Sinir Ağları (YSA) ve k-En Yakın Komşu (k-EYK) algoritmalarının küçük ve orta ölçekli veri kümeleri için sınıflandırma doğruluğu ve işlem hızı açısından etkili olduğu bulunmuştur. Çalışma, bu algoritmaların sağlık sektörü analizleri ve politika yapımında değerli araçlar olabileceğini göstermiştir. Bu bulgular, gelecekteki araştırmalar için temel oluşturmakta ve makine öğrenmesi algoritmalarının iyileştirilmesi veya hibrit kullanımlarının araştırılmasını önermektedir.

Alptekin ve Yeşilaydın (2015), OECD ülkelerinin belirlenen sağlık göstergeleri bazında bulanık kümeleme analizi ile sınıflandırmıştır. Çalışmanın amacı, Türkiye'nin dahil olduğu kümenin ve bu kümedeki diğer ülkelerin tespit edilmesi ve Türkiye'nin bu ülkelerle benzerliklerinin

belirlenmesidir. Toplam 34 OECD ülkesi, sağlıkla doğrudan ve dolaylı olarak etkileyen on değişken kullanılarak analiz edilmiştir. Bulgular, Türkiye'nin dördüncü kümede yer aldığını ve bu kümede Estonya, Macaristan, Meksika, Polonya ve Şili gibi ülkelerle benzer özellikler gösterdiğini ortaya koymuştur.

Akbar vd. (2020)'nin çalışmasında, sağlık sigortası dolandırıcılığının tahmininde karar ağacı sınıflandırıcı doğruluğunun Extreme Gradient Boosting (XGB) algoritması kullanılarak iyileştirilmesi incelenmiştir. Çalışma, Random Forest ve XGB Trees sınıflandırıcıları kullanarak potansiyel dolandırıcılık sağlayıcılarını tahmin etmeyi amaçlamaktadır. Toplamda, XGB Trees sınıflandırıcısı, rastgele alt örnekleme kullanarak %87 geri çağırma ve %86 doğruluk ile en iyi performansı göstermiştir. Bu sonuçlar, sağlık sigortası dolandırıcılığı tespitinde XGB Trees sınıflandırıcısının etkililiğini ortaya koymaktadır.

Doğan (2020)'ın çalışmasında, Türkiye'de sağlık hizmet talebinde gelir düzeyinin doğrudan ve zaman maliyeti nedeniyle gelirin dolaylı etkisini incelemeyi amaçlamıştır. Teorik inceleme, sağlık ekonomisi modellerinin karşılaştırılmasını içermekte olup, olgusal analiz Türkiye Sağlık Araştırması (2016) verileri üzerinden gerçekleştirilmiştir. Bulgular, gelirin hem doğrudan hem de dolaylı olarak sağlık hizmeti talebini etkilediğini göstermektedir. Çalışma, Türkiye'de sağlık düzenlemelerinde gelir grupları arasındaki farklılıkların dikkate alınmasının önemini vurgulamaktadır.

Genç ve Kurutkan (2021), çalışmalarında, sosyo-ekonomik faktörlerin karşılanmayan sağlık ihtiyaçlarına etkilerini ve bu ihtiyaçların sağlıkta eşitsizliklere neden olan faktörleri belirlemeyi amaçlamaktadır. "TÜİK Sağlık Araştırması" verileri (2014 ve 2016) üzerinden yapılan analizlerde, depresyon hastalığı ve sosyo-ekonomik değişkenler arasındaki ilişkiyi belirlemek için Binary Logit Regresyon analizi kullanılmıştır. Analiz sonuçlarına göre, yaş, gelir düzeyi, sağlık güvencesi, cinsiyet, medeni durum, eğitim, çalışma durumu, genel sağlık durumu, bedensel ağrı durumu ve depresyon gibi faktörler, karşılanmayan sağlık ihtiyaçlarına en fazla etki eden unsurlardır. Ayrıca, karşılanmayan sağlık ihtiyaçlarından kaynaklanan sağlık eşitsizliklerinin özellikle dezavantajlı grupları etkilediği belirlenmiştir.

Theerthagiri ve Vidya (2022)'nin çalışmasında, kardiyovasküler hastalık tahmini için yenilikçi bir Gradient Boosting tabanlı Rekürsif Özellik Eleme (RFE-GB) sınıflandırma algoritması önerilmiştir. Araştırma, bu algoritmayı geleneksel Lineer Diskriminant Analizi (LDA), K-En Yakın Komşu (KNN), Karar Ağacı (DT), Naive Bayes (NB) ve Çok Katmanlı Algılayıcı (MLP) algoritmalarıyla karşılaştırmıştır. RFE-GB algoritması, %89.78'lik bir doğruluk oranıyla en yüksek performansı göstermiştir. Bu sonuç, kardiyovasküler hastalıkların sınıflandırılmasında RFE-GB algoritmasının diğer makine öğrenmesi yöntemlerine kıyasla üstün olduğunu göstermektedir.
Wang vd. (2022)'nin çalışmasında, petrokimya tesislerinin insan sağlığı üzerindeki risklerini tespit etmek için Extreme Gradient Boosting (XGBoost) algoritması kullanılmıştır. 13 göstergeye dayanan bir risk tespit indeks sistemi oluşturulmuş ve farklı yapay zeka yöntemleriyle karşılaştırılmıştır. Sonuçlar, XGBoost'un diğer modellere göre daha yüksek doğruluk oranıyla en iyi performansı sergilediğini göstermiştir. Çalışma, site konumu, planlaması ve üretim süresi gibi faktörlerin risk tespitinde önemli olduğunu belirlemiştir. Bu bulgular, endüstriyel sitelerin sağlık risklerinin yönetiminde önemli katkılar sağlamaktadır.

Xu vd. (2022)'nin çalışmasında, EEG sinyallerinin doğrusal olmayan özellikleri ve Gradient Boosting Decision Tree (GBDT) kullanılarak epilepsi nöbetlerinin erken tahmini için bir yöntem geliştirilmiştir. EEG sinyalleri iki kategoriye ayrılmış ve çeşitli entropi özellikleri çıkarılmıştır. Geliştirilen GBDT sınıflandırıcısı, 10 kat çapraz doğrulama ile değerlendirilmiş ve ortalama %91.76 doğruluk oranı elde edilmiştir. Araştırma, bu yöntemin epilepsi nöbetlerini erken tahmin etmede etkili olduğunu ve düşük yanlış alarm oranları sağladığını göstermektedir. Bu sonuçlar, epilepsi hastaları için erken uyarı ve zarar azaltma potansiyeli taşımaktadır.

Chung ve Teo (2023), çalışmalarında, Gradient Boosting algoritması ile derin öğrenme yöntemleri kullanılarak zihinsel sağlık sorunlarının tahmin edilmesi üzerine odaklanmışlardır. Çeşitli makine öğrenme algoritmalarının performansları karşılaştırılmış ve Gradient Boosting, %88.80'lik bir doğruluk oranı ile en yüksek performansı göstermiştir. Çalışma, bu algoritmaların zihinsel sağlık sorunlarının tahmininde etkili olduğunu ve bu tür tahminlerin klinik teşhislerde kullanılabileceğini göstermektedir. Bu sonuçlar, zihinsel sağlık alanında makine öğrenmesi yaklaşımlarının potansiyelini ortaya koymaktadır. Tripathi vd. (2023)'nin çalışmasında, kolon kanseri dokusunun makine öğrenmesi algoritmaları kullanılarak sınıflandırılması incelenmiştir. Araştırmada, kolorektal kanserli dokulardan elde edilen 7180 görüntü üzerinde çeşitli makine öğrenimi yöntemleri (K-En Yakın Komşu, Destek Vektör Makinesi, Karar Ağacı, Rastgele Orman, Extreme Gradient Boosting, Gaussian Naive Bayes) kullanılmıştır. Sonuçlar, Extreme Gradient Boosting yönteminin en etkili ve uygulanabilir yaklaşım olduğunu göstermiştir. Bu

çalışma, kolon kanseri dokusunun sınıflandırılmasında yapay zekâ ve makine öğrenmesi tekniklerinin potansiyelini ortaya koymaktadır.

Yongcharoenchaiyasit vd. (2023), yaşlılarda kalp yetmezliği, aort stenozu ve demansın sınıflandırılması için Gradient Boosting tabanlı bir model geliştirilmiştir. Çalışma, bu hastalıkların tahmininde çeşitli makine öğrenmesi algoritmalarının performanslarını karşılaştırmış ve Gradient Boosting yönteminin diğerlerine göre daha yüksek doğruluk oranına sahip olduğunu bulmuştur. Araştırma, hastalıkların erken tanısında ve tedavisinde bu tür makine öğrenmesi yaklaşımlarının potansiyelini ortaya koymaktadır.

Yin vd. (2024), sağlık tesislerinin iklim üzerindeki etkilerini tahmin etmek için Gradient Boosting Machine (GBM) modelleri kullanmışlardır. GBM, 2020 yılında 283 hastanenin katıldığı bir anketten elde edilen verilerdeki eksiklikleri doldurmak için kullanılmıştır. GBM ile elektrik kullanımı, sığır eti tüketimi ve anestezik gaz desfluran kullanımı tahmin edilmiştir. Çalışma, bu üç alanın toplamda 3 milyon metrik ton CO2 eşdeğeri emisyon ürettiğini tespit etmiştir. Elektrik tüketimi, toplam karbon ayak izinin en büyük kısmını oluşturmuştur. Bu çalışma, sağlık tesislerinin karbon emisyonlarını tahmin etmede GBM yönteminin potansiyelini göstermektedir.

3. Materyal ve Metot

Akademik yöntem, bilimsel çalışmalarda kullanılan, hipotez oluşturma, veri toplama, analiz etme, yorum yapma ve sonuçları raporlama aşamalarını kapsayan sistematik ve disiplinli bir yaklaşımdır. Bu yöntem, bilimsel araştırmaların güvenilir, tekrar edilebilir ve geçerli sonuçlar üretmesine katkıda bulunur (Karaca, 2015). Sınıflandırma ise, belirli özelliklere dayanarak nesneleri veya bireyleri gruplara ayırmayı hedefler, amacı her durumda doğru sınıf tahminini yapmaktır (Kayakuş ve Yiğit Açıkgöz, 2023). Sınıflandırma yöntemlerinde, verinin bir bölümü eğitim, diğer bölümü test seti olarak kullanılır ve hata ile doğruluk oranları, sınıflandırmanın başarısını ölçmek için değerlendirilir (Yıldıztepe ve Kocataş, 2018).
Verilerin sınıflandırılması için öğrenme algoritmaları temel alınmaktadır. Var olan veri tabanının bir kısmı eğitim amacıyla kullanılarak sınıflandırma kuralları oluşturulur. Oluşturulan model yeni durumlar için nasıl karar vereceği belirlenir. Elde edilen modelin doğruluğu test edilerek onaylanırsa, bu model diğer veriler üzerinde de uygulanır (Yılmaz, 2012). Sınıflandırma yaparken kullanılan algoritmaların seçimi, kullanılacak veri tipiyle uyumlu olmalıdır. Böylece daha doğru sonuçlar elde edebilmemiz mümkündür (Çiçek ve Arslan, 2020).

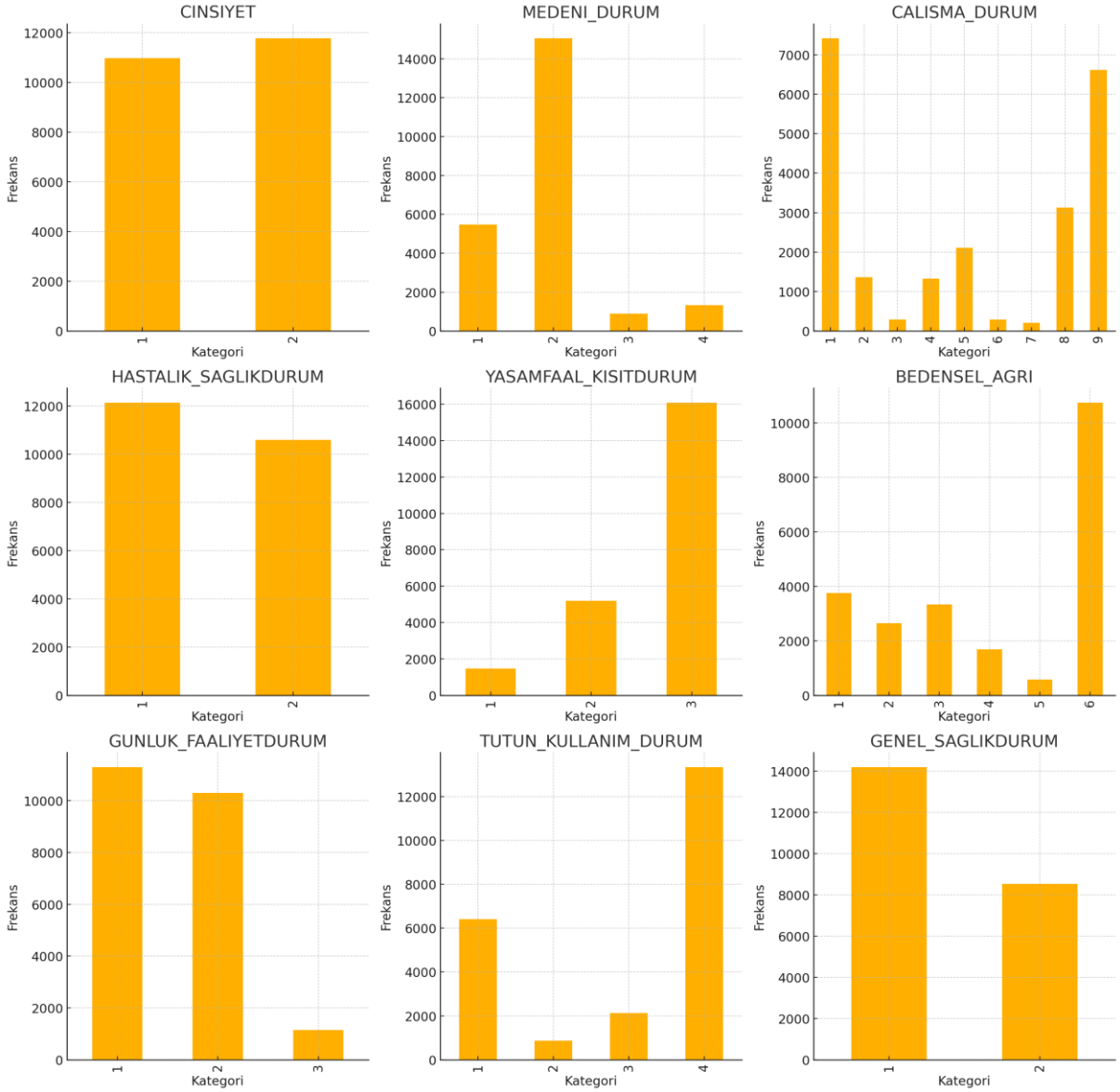
3.1 Veri Seti

Bu çalışmada Türkiye İstatistik Kurumu (TÜİK) tarafından gerçekleştirilen 2022 Yılı Türkiye Sağlık Araştırması verileri kullanılmıştır. Araştırmadan elde edilen veriler, ülke genelinde toplamda 11.170 adet haneden elde edilmiştir. Yapılan araştırma, ülkelerin gelişmişlik düzeylerini gösteren kalkınma göstergeleri içerisinde önemli paya sahip olan, sağlık göstergelerine ait bilgilerin elde edilmesine olanak tanımıştır. Araştırma, sadece ülke genelini yansıtmamakta, aynı zamanda uluslararası karşılaştırmalara olanak tanımaktadır. Araştırmanın kapsamı, Türkiye sınırları içinde yer alan tüm yerleşim yerlerinde bulunan hane halklarıdır. Bu bilgiler doğrultusunda, bu çalışmanın temel amacı, makine öğrenmesi tekniklerini kullanarak, mikro sağlık verilerinin nasıl kategorize edebileceğini araştırmaktır. Araştırmada kullanılan anket formundan elde edilen dokuz bağımsız değişken, sağlık durumu tahminlerinin yapılmasında analiz edilmiştir. Yapılan veri seti in inceleme adımında, eksik veriler, veri tabanından çıkarılmıştır. Veri setinde toplamda 22.742 adet hücre verisi mevcuttur.

Çalışmada kullanılan bağımsız değişkenler şunlardır: cinsiyet, yaş, medeni durum, çalışma durumu, hastalık ve sağlık durumu, yaşam faaliyetlerindeki kısıtlılıklar, bedensel ağrı, günlük faaliyetlerin durumu ve tütün kullanımı. Bu değişkenler, araştırmada bireylerin genel sağlık durumunu (Y olarak tanımlanan bağımlı değişken) etkileyen faktörler olarak incelenmiştir. Bu bağımsız değişkenler, genel sağlık durumunu anlamak ve değerlendirmek için kullanılmıştır. Kullanılan her bir kategorik değişkenin frekans dağılımları Şekil 1'de gösterilmektedir.

Veri setindeki tüm kategorik değişkenlerin frekans dağılımları Şekil 1'de gösterilen bar grafikler kullanılarak detaylı bir şekilde incelenmiştir. "Cinsiyet" değişkeni, kadın (2) ve erkek (1) bireylerin dengeli bir dağılım gösterdiğini ortaya koymaktadır. "Medeni Durum" değişkeninde, evli bireylerin (2) büyük bir çoğunluğu oluşturduğu, bekâr (1), dul (3) ve boşanmış (4) bireylerin ise daha düşük oranlarda temsil edildiği belirlenmiştir. "Çalışma Durumu" değişkeninde 1. (ücretli çalışan) ve 9. (Ev işleri ile meşgul) kategorilerinin en yüksek frekansa sahip olduğu, 7. (ev hanımı) ve 8. (öğrenci) kategorilerinin ise nispeten daha az gözlem içerdiği saptanmıştır. "Hastalık/Sağlık Durumu" değişkeni, büyük ölçüde sağlıklı bireylerden (1) oluşurken, "Yaşam Faaliyeti Kısıt Durumu" değişkeninde kısıtlılık yaşayan bireylerin (3) baskın olduğu gözlemlenmiştir. "Bedensel Ağrı" değişkeninde ise ağrısı yüksek olan bireylerin (6) yaygın olduğu, "Günlük Faaliyet Durumu" değişkeninde çoğunluğun (1) faaliyetsiz veya

kısıtlı faaliyet gösterdiği anlaşılmıştır. "Tütün Kullanım Durumu" değişkeninde ise yüksek oranda tütün kullanan bireylerin (4) mevcut olduğu görülürken, "Genel Sağlık Durumu" değişkeni ise bireylerin çoğunluğunun sağlıklı (1) olduğunu ortaya koymuştur. Bu analizler, verinin genel yapısını anlamaya yönelik önemli ipuçları sunmakta ve çalışmanın ilerleyen aşamaları için güçlü bir temel oluşturmaktadır.



Şekil 1. Kategorik Değişkenlerin Frekans Dağılımları

Tablo 1'de analizlerde kullanılan bağımsız değişkenler ve değişkenlerin açıklama ve tanımlamaları gösterilmiştir. İlgili tabloda yer alan veri seti, içerdiği bilgilerin türünü ve bu bilgilerin nasıl kodlandığını açıklamaktadır.

3.2. Yazılım ve Donanım

Bu çalışmada genel sağlık durumunun sınıflandırılması işlemi Python 3.6 yazılımında yapılmıştır. Sınıflandırma işleminde, Karar Ağaçları (Decision Trees), En Yakın Komşu Algoritması (K-Nearest Neighbors), Doğrusal Diskriminant Analizi (Linear Discriminant Analysis), Destek Vektör Makinesi (Support Vector Machine) ve Gradient Boosting modellerini kullanmak için yaygın olarak kullanılan scikit-learn kütüphanesi tercih edilmiştir. Scikit-learn, makine öğrenmesi modellerinin eğitimini, testini ve değerlendirilmesini kolaylaştıran zengin bir araç seti sunmaktadır. Decision Tree Classifier, K-Neighbors Classifier, Linear Discriminant Analysis, SVC (Support Vector Classification) ve Gradient Boosting Classifier sınıfları, bu modellerin scikit-learn kütüphanesindeki karşılıklarıdır. Ayrıca, veri işleme ve model değerlendirme süreçlerinde pandas ve numpy kütüphaneleri veri manipülasyonu ve matematiksel hesaplamalar için; matplotlib ve seaborn kütüphaneleri ise veri görselleştirme için kullanılmıştır. Bu kütüphaneler birlikte kullanıldığında, makine öğrenmesi projelerinin baştan sona etkili bir şekilde gerçekleştirilmesine imkân tanımaktadır.

Tablo 1. Bağımsız Değişkenler ve Tanımlamaları

Değişken Ad	Değişken Tanımı	Açıklama
Cinsiyet	Kişinin Cinsiyeti	1. Erkek 2. Kadın
Yaş	Bitirilen yaş	
Medeni Durum	Medeni durumunuz nedir?	1. Hiç evlenmedi 2. Evli 3. Boşanmış 4. Eşi vefat etmiş
Çalışma Durumu	Çalışma durumunuz nedir?	1. Ücretli çalışan 2. İşveren 3. Ücretsiz aile işçisi 4. İş arayan 5. Eğitime devam eden 6. Emekli 7. Yaştan ötürü çalışamayan 8. Engelli 9. Ev işleri ile meşgul
Hastalık ve Sağlık Durumu	6 aydan uzun süren hastalığınız var mı?	1. Evet 2. Hayır
Yaşam Faaliyet Kısıt Durumu	Sağlık probleminden ötürü günlük faaliyetlerinizin kısıtlanma durumu nedir?	1.Ciddi biçimde kısıtlandı 2.Kısıtlandı (ciddi değil) 3.Kısıtlanmadı
Bedensel Ağrı	Son 4 haftadaki bedensel ağrınız nedir?	1.Çok az 2.Az 3.Orta 4.Fazla 5.Çok fazla 6. Hiç
Günlük Faaliyet Durumu	Hangisi durumunuzu en iyi şekilde tanımlamaktadır?	1.Çoğunlukla oturan 2.Orta derecede fiziksel işler 3.Genelde ağır iş
Tütün Kullanım Durumu	Tütün mamulleri kullanıyor musunuz?	1.Evet, her gün 2.Evet, ara sıra 3.Hayır içiyordum bıraktım 4.Hayır hiç kullanmadım
Genel Sağlık Durumu	Ferdin Genel Sağlık Durumu	1.Sağlığı iyi 2.Sağlığı kötü

Bununla birlikte, sınıflandırma yöntemlerinde, kullanılan veri seti genellikle eğitim ve test setleri olarak ikiye ayrılır. Hata oranı ve doğruluk oranı, yanlış ve doğru sınıflandırılan kayıtları belirlemek için kullanılır (Yıldıztepe ve Kocataş, 2018). Söz konusu ayrım, bu çalışmada, verinin %80'i eğitim ve %20'si test seti olacak şekilde "train_test_split" komutu ile gerçekleştirilmiştir. Veri bölümlemesi sırasında "random_state" parametresi de kullanılarak, bu işlemde tutarlılık sağlanmış ve veri setleri her seferinde aynı şekilde bölünerek model performanslarının kıyaslanmasına olanak sağlanmıştır. Random_state, rastgele sayı tohumu olarak işlev görerek, her çalışmada aynı veri bölümlemesinin tekrarlanmasını ve sonuçların karşılaştırılmasını sağlar. Bu sayede, farklı denemeler arasında tutarlılık korunmuş olur (Bisht ve Bisht, 2022).

Tüm sınıflandırma işlemleri için kullanılan bilgisayar donanımları şu özelliklere sahiptir: İşlemci olarak, yüksek performanslı 8 çekirdekli bir Intel Core i7-10700K bulunmaktadır, bu da büyük veri kümeleriyle çalışırken ve yoğun hesaplamalar yaparken üstün performans sağlamaktadır. 32 GB DDR4 RAM, veri işleme ve model eğitimi sırasında bellek yetersizliği yaşanmaması için yeterli kapasite sunmaktadır. Grafik kartı olarak, NVIDIA GeForce RTX 3080 tercih edilmiş; bu güçlü GPU, özellikle derin öğrenme modelleriyle çalışırken hesaplama gücünü artırmıştır. Depolama birimi olarak, 1 TB kapasiteli NVMe SSD kullanılmış, bu da hızlı veri erişimi ve büyük dosyaların saklanması için idealdir. Ayrıca, bilgisayar iyi bir soğutma sistemi ve yüksek çözünürlüklü 27 inç bir monitör ile donatılmış, bu da uzun süreli çalışma koşullarında verimliliği artırmıştır. Bu donanım özellikleri, sınıflandırma işlemlerini hızlı ve verimli bir şekilde gerçekleştirmek için mükemmel bir altyapı sunmaktadır.

3.3. Makine Öğrenmesi Modelleri

3.3.1. Karar Ağaçları

Karar ağaçları, makine öğrenimi, görüntü işleme ve örüntü tanıma dahil olmak üzere çeşitli alanlarda sıklıkla

kullanılan etkili bir tekniktir (Stein vd., 2005). Karar ağacı modeli, adından da anlaşılacağı üzere ağaç benzeri bir yapı sergiler. Hiyerarşik bir ağaç yapısına benzeyen kökler, dallar ve yapraklardan oluşur. Karar ağaçları verileri kökte sınıflandırarak başlar, ardından düğümler, dallar ve yapraklar (karar sınıfı) aracılığıyla devam eder. Bu süreç ağaç yapısı yapraklar elde edene kadar devam eder. Düğüm noktaları, karar oluşturma sürecinin gerçekleştiği belirli konumlardır. Elde edilen sonuca göre dallar oluşturulur ve oluşturulan dallara uygun olarak yapraklar elde edilir. Bir yaprak oluşturulmamışsa bir düğüm oluşturulur ve bu süreç karar sınıfını temsil eden bir yaprak oluşana kadar devam eder (Kızgın vd., 2023). Bu çalışmada, İnce Ağaç (Fine Tree) kullanılmıştır. İnce Ağaç sınıflandırması, ince taneli sınıfların daha kaba üst sınıflar halinde kümelendiği ve nesne örneklerinin üst sınıfları aracılığıyla ayrıntılı kategorilere ayrıştırılmasına yardımcı olan bir sınıflandırma ağacı oluşturan bir yöntemdir (Wu vd., 2020). Bu yaklaşım, özellikle geniş kelime dağarcığına sahip uzun kuyruklu nesne algılama ve örnek segmentasyon görevlerinin üstesinden gelme ihtiyacının olduğu senaryolarda kullanışlıdır (Wu vd., 2020). İnce Ağaç sınıflandırması, çeşitli çalışmalarda Doğrusal Diskriminant Analizi (LDA) ve K-En Yakın Komşular (KNN) gibi diğer yöntemlerle karşılaştırılmış ve farklı uygulamalardaki etkinliği gösterilmiştir (Kaya, 2021).

3.3.2. En Yakın Komşular (K-Nearest Neighbours-KNN)

KNN algoritması kullanım kolaylığı ve etkinliği ile en çok kullanılan sınıflandırma algoritmalarından biridir (Wu vd., 2008). Ayrıca KNN algoritmaları regresyon ve kayıp değer girişi gibi çeşitli veri madenciliği uygulamalarında da sıklıkla kullanılmaktadır. k-En Yakın Komşular (KNN), birden fazla uygulamada hem basit hem de oldukça verimli olan parametrik olmayan bir sınıflandırma tekniğidir. Standart KNN yaklaşımının arkasındaki temel kavram, çoğunluk kuralına dayalı olarak bir test veri noktasının etiketini tahmin etmektir. Başka bir deyişle, test veri noktasının etiketi, özellik uzayındaki en benzer k eğitim veri noktası arasında en yaygın sınıf dikkate alınarak tahmin edilir (Cheng vd., 2015).

Bir t veri kaydını sınıflandırmak için, önce t'nin bir komşuluğunu oluşturan k en yakın komşusuna ulaşılır. t için sınıflandırma, mesafeye dayalı ağırlıklandırmanın dikkate alınıp alınmadığına bakılmaksızın, genellikle komşuluktaki veri kayıtları arasında oy çokluğuyla belirlenir. Bununla birlikte, k-en yakın komşu (KNN) algoritmasını uygulamak için, sınıflandırmanın doğruluğu büyük ölçüde bu parametreye bağlı olduğundan, k için uygun bir değer seçmek çok önemlidir. KNN yöntemi k değerinden etkilenir ve bu da belirli bir düzeyde önyargıya neden olabilir. K değerini seçmek için çok sayıda yöntem vardır, ancak basit olanı algoritmayı farklı k değerleriyle birden çok kez çalıştırmak ve en uygun performansı göstereni seçmektir (Guo vd., 2003). KNN tekniği tipik olarak sırasıyla Denklem 1-4'te verilen Öklid, Chebyshev, Manhattan ve Mahalanobis mesafe ölçümlerini kullanarak verilerin yakınlık ölçümlerini hesaplar.

$$d_{Euclidean}(x_i, y_i) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (1)$$

$$d_{Chebishev}(X_i, Y_i) = max_{i=1,2,3,\ldots,m}|X_i, Y_i| \quad (2)$$

$$d_{Manhattan}(X_i, Y_i) = \sum_{i=1}^{n}|X_i - Y_i| \quad (3)$$

$$d_{Mahalabonis}(x_i, y_i) = \sqrt{(X_i - Y_i)^T \Sigma^{-1}(X_i - Y_i)} \quad (4)$$

3.3.3. Doğrusal Diskriminant Analizi (Linear Discriminant Analysis-LDA)

Doğrusal Diskriminant Analizinin (LDA) amacı, orijinal belirleyicilerin bir bileşimi olan yeni bir değişken üretmektir. Bu, yeni faktörle ilgili olarak önceden belirlenmiş kategoriler arasındaki farklılıkların optimize edilmesiyle elde edilir. Amaç, tahmin edici puanları, diskriminant puanı olarak bilinen tek bir yeni bileşik değişkenin oluşturulmasıyla sonuçlanacak şekilde birleştirmektir. Bu, verilerin boyutlarını azaltmaya yönelik bir yaklaşım olarak görülebilir ve tahmin edicileri p boyuttan tek boyutlu bir çizgiye sıkıştırır. Sürecin sonunda amaç, sınıflar arasındaki ortalama puan farkını en üst düzeye çıkarırken, her sınıf için normal bir diskriminant puanı dağılımına sahip olmaktır (Subaşı ve Gürsoy, 2010). LDA yaklaşımı, yordayıcı değişkenlerin aralık veya oran değişkenleri olduğunu varsayar, yani nesne büyüklükleri açısından sıralanabilir ve karşılaştırılabilirler. Oran ölçekleri aralık ölçekleriyle aynı özelliklere sahiptir, ancak aynı zamanda mutlak bir sıfır noktasına sahiptirler. İkinci olarak LDA, yordayıcı değişkenlerin puanlarının çok değişkenli normal dağılım izleyen bir popülasyondan bağımsız ve rastgele seçildiğini varsayar. Üçüncü olarak, gruplar aynı varyans/kovaryans matrislerine sahiptir. Görsel olarak, her bir gruba ait öğeler kendilerini aynı şekil, boyut ve yönelimi paylaşan çok boyutlu bir 'bulut' halinde düzenlemelidir (Worth ve Cronin, 2003). LDA, olasılıksal yorumlar sağlayan bir sınıflandırma yaklaşımıdır. Basit bir ifadeyle, çeşitli sınıf etiketleri arasında optimum farklılaşma sağlamak için doğrusal bir kombinasyon oluşturan faktörlerin belirlenmesine dayanır (Mandelkow vd., 2016). LDA açıkça farklı veri sınıfları arasındaki ayrımı yakalamayı amaçlamaktadır. LDA, altta yatan uzayda farklı sınıflar arasında en yüksek düzeyde ayırt edilebilirlik sergileyen vektörleri tanımlamaya çalışır. Resmi olarak LDA, mevcut verilere dayanarak hedef sınıflar arasındaki ortalama farkları maksimize eden bağımsız özelliklerin doğrusal bir

kombinasyonunu oluşturur (Rathi ve Palani, 2012). İlgili denklemler 5 ve 6 'da verilmiştir:

$$S_W = \sum_{j=1}^{c}\sum_{i=1}^{N_j}(x_i^J - \mu_j)(x_i^J - \mu_j)^T \quad (5)$$

"x_i^J" ifadesi j sınıfının birinci örneğini, "μ_j" j sınıfının ortalamasını, "c" sınıf sayısını temsil etmektedir. Ayrıca, "sınıflar arası dağılım matrisi" terimi belirli bir matematiksel kavramı ifade eder. μ tüm sınıfların ortalamasını gösterir.

$$S_b = \sum_{j=1}^{c}(x_i^J - \mu_j)(x_i^J - \mu_j)^T \quad (6)$$

3.3.4. Destek Vektör Makinesi (Support Vector Machine-SVM)

Destek Vektör Makineleri (SVM) teorisi, geleneksel örüntü tanıma tekniklerinin eğitim seti üzerindeki performansı optimize ederek ampirik riski en aza indirmeyi amaçladığını belirtmektedir. Ancak SVM, sabit ancak bilinmeyen bir veri olasılık dağılımına dayalı olarak görülmeyen örüntüleri yanlış sınıflandırma olasılığı olan yapısal riski en aza indirmeye odaklanır. Genelleme hatasında bir üst sınırı en aza indirmeye eşdeğer olan bu yeni tümevarım ilkesi, olasılıkta tekdüze yakınsama teorisine dayanmaktadır (Yue vd., 2003). Destek Vektör Makinelerinin (SVM) amacı, marjı maksimize eden ve sınıfları doğrusal bir şekilde etkili bir şekilde ayırabilen hiper düzlemi tanımlamaktır (Abdullah ve Abdulazeez, 2021). Destek Vektör Makineleri (SVM'ler), doğrusal olmayan durumların sınıflandırılmasını kolaylaştırmak için girdi uzayından özellik uzayına bir eşleme gerçekleştirir. Çekirdek yöntemi, eşleme fonksiyonunun açık bir şekilde tanımlanması ihtiyacını ortadan kaldırarak boyutluluk laneti sorununu ele almada faydalıdır. Özellik uzayı olarak da bilinen yeni uzaydaki doğrusal sınıflandırma, girdi uzayı olarak da bilinen orijinal uzaydaki doğrusal olmayan sınıflandırmaya eşdeğerdir. SVM'ler bunu, girdi vektörlerini maksimum ayırma hiper düzleminin oluşturulduğu daha yüksek boyutlu bir uzaya (özellik uzayı olarak da bilinir) dönüştürerek başarır (Yu ve Kim, 2012).

3.3.5. Gradient Boosting

Gradient Boosting, makine öğrenimi alanında yaygın olarak kullanılan bir ensemble öğrenme tekniğidir. Algoritma, hatalara odaklanarak çalışmaktadır. İlk öğrenici genellikle basit bir tahmin yapar, ardından sonraki modeller öncekilerin hatalarını düzeltmeye odaklanarak öğrenir. Bu şekilde, her adımda daha doğru tahminler elde edilir. Gradient Boosting'in belki de en önemli özelliği, her adımda modelin önceki hatalarını minimize etmek için gradient (eğim) bilgisini kullanmasıdır. Bu, her bir öğrenicinin eklenirken tahmin performansını optimize etmeye yöneliktir. Algoritma, belirli bir hata eşiğine ulaşılıncaya veya belirli bir iterasyon sayısına kadar devam eden iteratif bir süreçtir. Her adımda, bir önceki modelin hatası üzerine yeni bir model eklenir ve bu süreç tekrarlanır (Bentéjac vd., 2021).

Gradient Boosting'in Temel Adımları şu şekilde özetlenebilir: Veri seti için ilk basit tahmin yapılır. Genellikle bu, sınıflandırma için en yaygın sınıfı veya regresyon için ortalama bir değeri temsil etmektedir. Bu süreç Denklem 7'de gösterilmiştir.

$$F_0(x) = arg\,min_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \quad (7)$$

Burada, $F_0(x)$, başlangıç modelidir, L kayıp fonksiyonu, y_i gerçek değerler ve γ modelin çıktısıdır. Bu işlemden sonra, ilk tahminle gerçek değerler arasındaki fark (hata) hesaplanmaktadır. Yani, her iterasyonda, modelin çıktısı ile gerçek değer arasındaki kayıp fonksiyonunun gradyanı hesaplanır. Bu süreç Denklem 8'de gösterilmiştir.

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad (8)$$

r_{im} değerleri, m-inci iterasyonda modelin tahmin hatalarının gradyanlarıdır ve $F_{m-1}(x)$ bir önceki adımda elde edilen modeldir. Bu işlemin ardından, hataları minimize edecek bir zayıf öğrenici $h_m(x)$ (örneğin, bir karar ağacı) eğitilir. Bu öğrenici, hataların gradyanlarını (r_{im}) tahmin etmeye çalışır. Ardından, modeli güncellemek için bir çarpan γ_m bulunur. Bu süreç Denklem 9'da gösterilmiştir.

$$\gamma_m = \arg min_\gamma \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \quad (9)$$

Bu adım, yeni öğrenicinin katkısının boyutunu ayarlar. Son olarak model, yeni öğrenicinin katkısıyla güncellenir. Bu süreç Denklem 10'da gösterilmiştir.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (10)$$

Burada, $F_m(x)$ güncellenmiş modeldir. Bu işlem, belirlenen sayıda iterasyon boyunca veya hatalar bir eşik değerin altına düşene kadar tekrarlanır (Mayr vd. 2014).

3.4. Performans Değerlendirme

Bununla birlikte, sınıflandırma başarısını değerlendirmek için bir dizi metrik bulunmaktadır, bunlar arasında kesinlik, hassasiyet, f1-skor, karışıklık matrisi, doğruluk ve özgünlük bulunmaktadır (Hossin ve Sulaiman, 2015).

Performans ölçütleri hata matrisi kullanılarak türetilmiştir ve 0 ile 1 arasında bir değer aralığına sahiptir. Hesaplamalara göre, ölçümlerin değerleri 1'e yaklaştığında üstün performans sergilediği iddia edilmektedir. Gerçek pozitif (TP), beklenen sonuç pozitif olduğunda ve gerçekleşen sonuç da pozitif olduğunda ortaya çıkar. Yanlış pozitif (FP), bir tahmin pozitif olduğunda, ancak gerçek durum negatif olduğunda,

örneğin sağlıklı bir bireyin hasta olacağı tahmin edildiğinde ortaya çıkar. Yanlış negatif (FN), tahmin olumsuz bir sonuca işaret ettiğinde, ancak gerçek durum olumlu olduğunda, örneğin bir hastanın sağlıklı olacağı tahmin edildiğinde ortaya çıkar. Tahmin olumsuz olduğunda ve gerçek durum da olumsuz olduğunda gerçek bir olumsuzluk (TN) meydana gelir. Çalışmada doğruluk, duyarlılık, hassasiyet ve F1 skoru metrikleri kullanılmıştır. Bu metriklerin hesaplanışı sırasıyla Denklem 11-14 arasında verilmiştir:

$$\text{Doğruluk} = \frac{Gerçek\ Pozitifler\ (TP) + Gerçek\ Negatifler\ (TN)}{Toplam\ Örnek\ Sayısı} \quad (11)$$

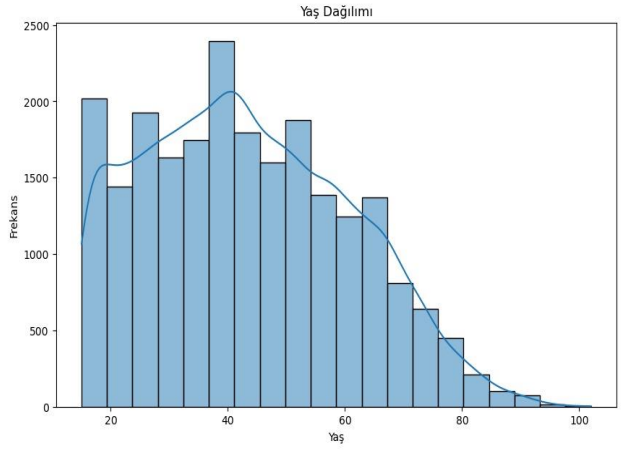
$$\text{Duyarlılık} = \frac{Gerçek\ Pozitifler\ (TP)}{Gerçek\ Pozitifler\ (TP) + Yanlış\ Negatifler\ (FN)} \quad (12)$$

$$\text{Hassasiyet} = \frac{Gerçek\ Pozitifler\ (TP)}{Gerçek\ Pozitifler\ (TP) + Yanlış\ Pozitifler\ (FP)} \quad (13)$$

$$\text{F1 Skoru} = 2 \times \frac{(Hassasiyet \times Duyarlılık)}{(Hassasiyet + Duyarlılık)} \quad (14)$$

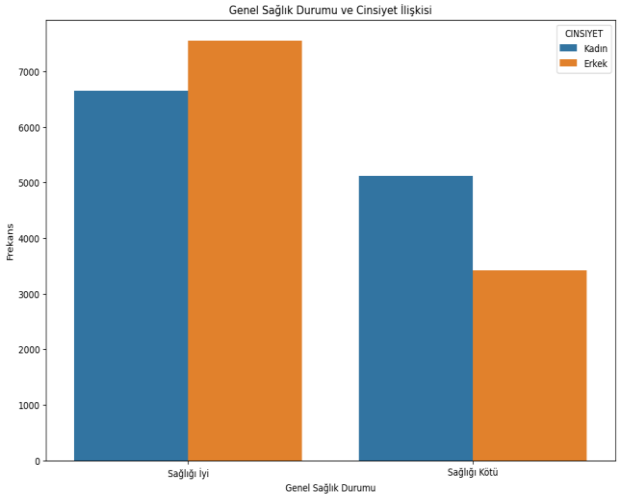
4. Bulgular

Çalışmadan elde edilen bulgular incelendiğinde, Şekil 2'de, veri setindeki kişilerin yaş dağılımını gösteren bir histogram bulunmaktadır. Histogram, veri noktalarını belirli aralıklara veya "kutulara" bölen bir grafik türüdür. Her bir aralıktaki gözlemlerin frekansını veya sayısını göstererek, veri setindeki yaş dağılımını görsel olarak ifade etmektedir.



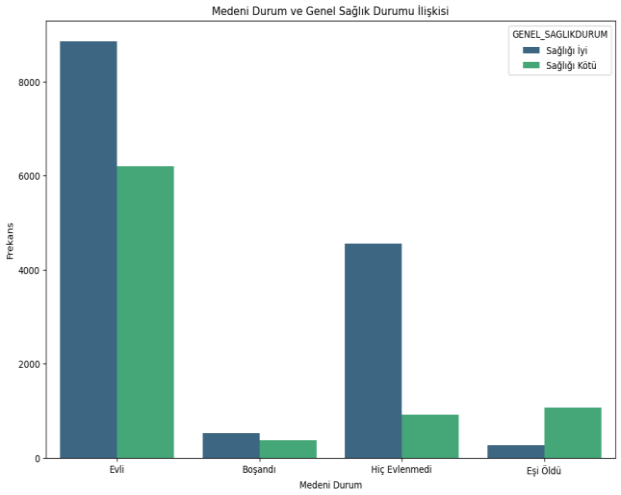
Şekil 2. Yaş Dağılımı

Grafikte, yaş dağılımının genellikle 20 ila 60 yaş arasında yoğunlaştığı görülmektedir. Genç yetişkinlerin sayısı fazla iken yaşın artışıyla birlikte birey sayısı azalmaktadır, 60 yaşından sonra ise hızla düşmektedir. Bu durum, yaşlı nüfusun az olduğuna işaret etmektedir. Grafik, nüfusun orta yaş grubunda yoğunlaştığını ve yaş dağılımının sağlık durum analizi için önemli olduğunu göstermektedir. Buna ek olarak, Şekil 3'te genel sağlık durumu ile cinsiyet ilişkisi grafik ortamında gösterilmiştir.



Şekil 3. Genel Sağlık Durumu ve Cinsiyet İlişkisi

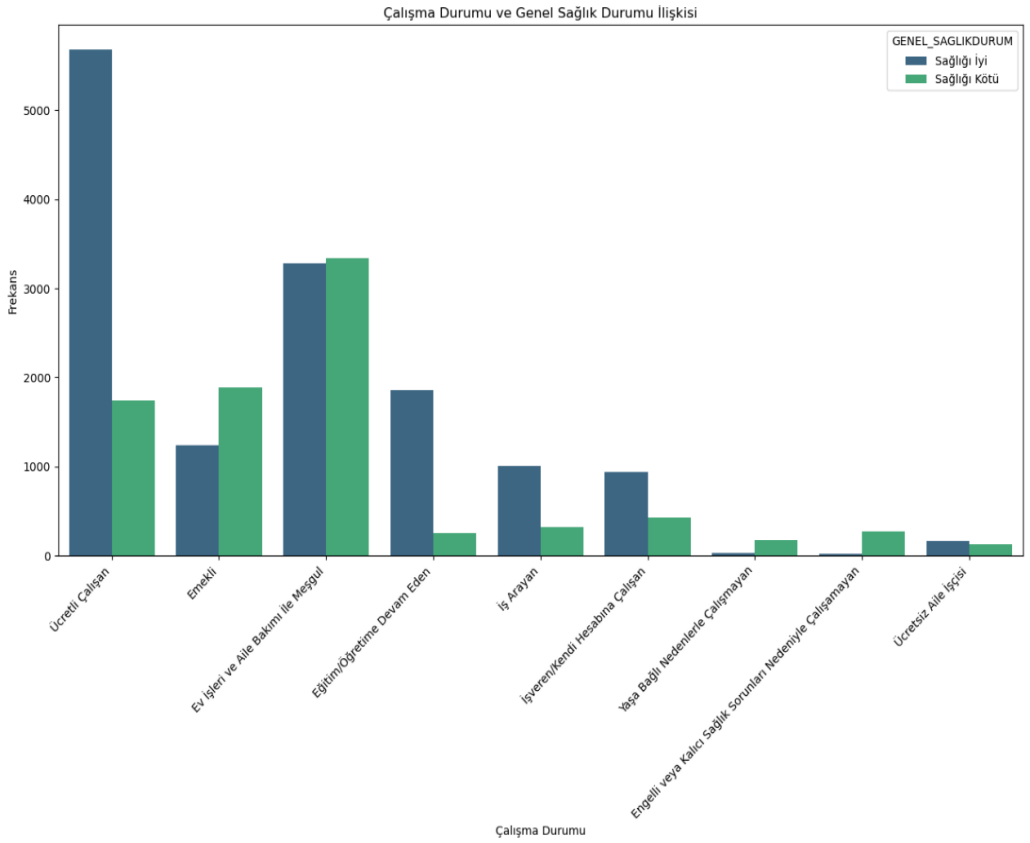
Şekil 3'te, genel sağlık durumunun iyi ya da kötü olduğunu belirten kadın ve erkek katılımcıların sayısı karşılaştırılmaktadır. Sağlıklı olduğunu belirtenler arasında kadınların sayısı erkeklere göre daha fazla, sağlığı kötü olanlar arasında ise erkeklerin sayısı kadınlardan yüksektir. Bu, sağlık durumunun cinsiyete göre değişiklik gösterebileceğine dair bir gösterge olabilir ve belki de cinsiyete özgü sağlık riskleri veya sağlık hizmetlerine erişim farklılıklarını yansıtabilir. Her iki kategoride de katılımcıların sayısı, sağlık algısının cinsiyetlere göre nasıl değişebileceği hakkında önemli veriler sunmaktadır. Öte yandan, Şekil 4'te medeni durum ve genel sağlık durumu arasındaki ilişki gösterilmiştir.



Şekil 4. Medeni Durum ve Genel Sağlık Durumu İlişkisi

Şekil 4, farklı medeni durumlara sahip bireylerin genel sağlık durumu ile ilişkisini göstermektedir. Evli bireyler arasında sağlıklı olanların sayısı oldukça yüksektir ve bu grup sağlığı kötü olanların sayısını da en çok barındırıyor. Boşanmış bireylerde ise sağlıklı olanların sayısı, sağlığı kötü olanlara göre daha az farkla öndedir. Hiç evlenmemiş kişilerde sağlıklı olanların sayısı, sağlığı kötü olanlardan

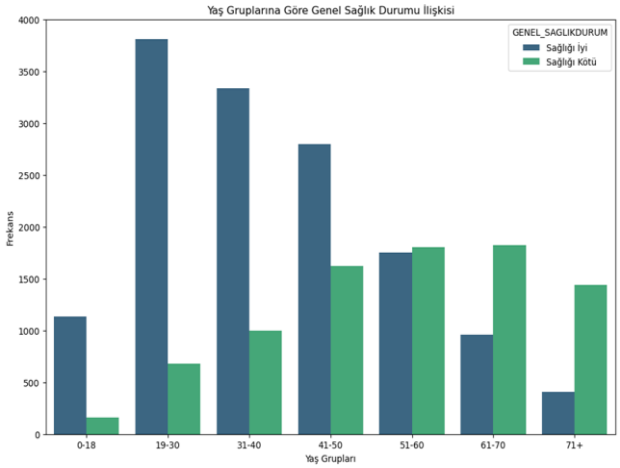
daha fazla, ancak bu grup en az sayıda sağlığı kötü bireyi içermektedir. Eşi ölmüş bireylerin sağlığı genel olarak kötü olarak raporlanmıştır ve bu kategorideki sağlıklı olanların sayısı en düşüktür. Bu veriler, medeni durumun genel sağlık üzerindeki olası etkilerini yansıtmakta ve sosyal destek ile sağlık arasındaki ilişkiyi anlamak için önemli ipuçları sunmaktadır. Bununla birlikte, Şekil 5'te çalışma durumu ve genel sağlık durumu arasındaki ilişki gösterilmektedir.



Şekil 5. Çalışma Durumu ve Genel Sağlık Durumu İlişkisi

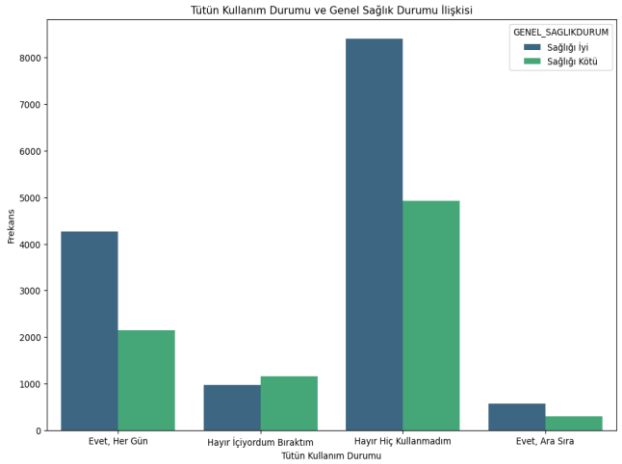
İlgili şekildeki bulgular incelendiğinde, ücretli çalışanlar arasında kendini sağlıklı olarak rapor edenlerin sayısı belirgin bir şekilde daha yüksektir. Ev hanımları ve emekliler arasında sağlığı iyi ve kötü olarak rapor edenler arasındaki fark daha azdır. İş arayanlar ve öğrenciler genellikle sağlıklı olduklarını belirtmişlerdir. Freelance (genellikle bağımsız çalışan ve bir organizasyona sabit bir şekilde bağlı olmayan profesyoneller için kullanılan bir terimdir) çalışanlar ve serbest meslek sahipleri genellikle sağlıklı olarak rapor edilmişken, evden çalışanlar ve uzaktan çalışanlar arasında sağlığı iyi olanların oranı daha düşüktür. Ücretsiz izinde olanlar ve diğerleri kategorisinde sağlığı iyi olanların sayısı oldukça azdır. Bu veriler, iş ve yaşam tarzı koşullarının sağlık üzerindeki etkilerini yansıtmakta ve sosyoekonomik faktörlerle sağlık arasındaki ilişkiyi gözler önüne sermektedir. Bununla birlikte Şekil 6, farklı yaş gruplarına göre bireylerin genel sağlık durumlarını karşılaştırmaktadır.

Bulgular incelendiğinde, genç yaş gruplarındaki (0-18, 19-30) bireylerin büyük çoğunluğu kendilerini sağlıklı olarak nitelendirirken, orta yaş grubu (31-40, 41-50) da önemli bir sağlıklı popülasyona sahip. 51-60 yaş aralığındaki sağlık durumu dengeli bir dağılım gösterirken, 61-70 ve 71+ yaş gruplarında sağlıklı bireylerin sayısı sağlığı kötü olanlara göre daha azdır.



Şekil 6. Yaş Gruplarına Göre Genel Sağlık Durumu İlişkisi

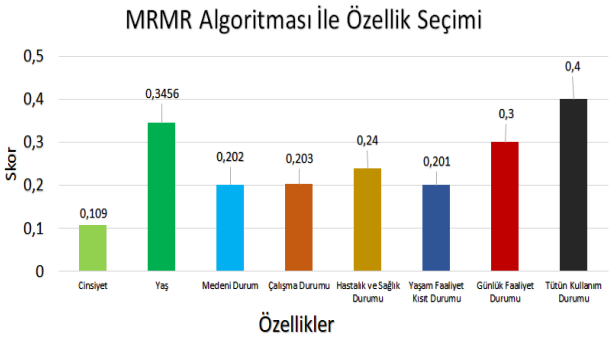
Bu eğilim, yaşla birlikte sağlık sorunlarının artışını ve genel sağlık durumunun genç bireylerde daha iyi olduğunu göstermektedir. Buna ek olarak Şekil 7'de tütün kullanım durumu ve genel sağlık durumu ilişkisi karşılaştırılmıştır.



Şekil 7. Tütün Kullanım Durumu ve Genel Sağlık Durumu İlişkisi

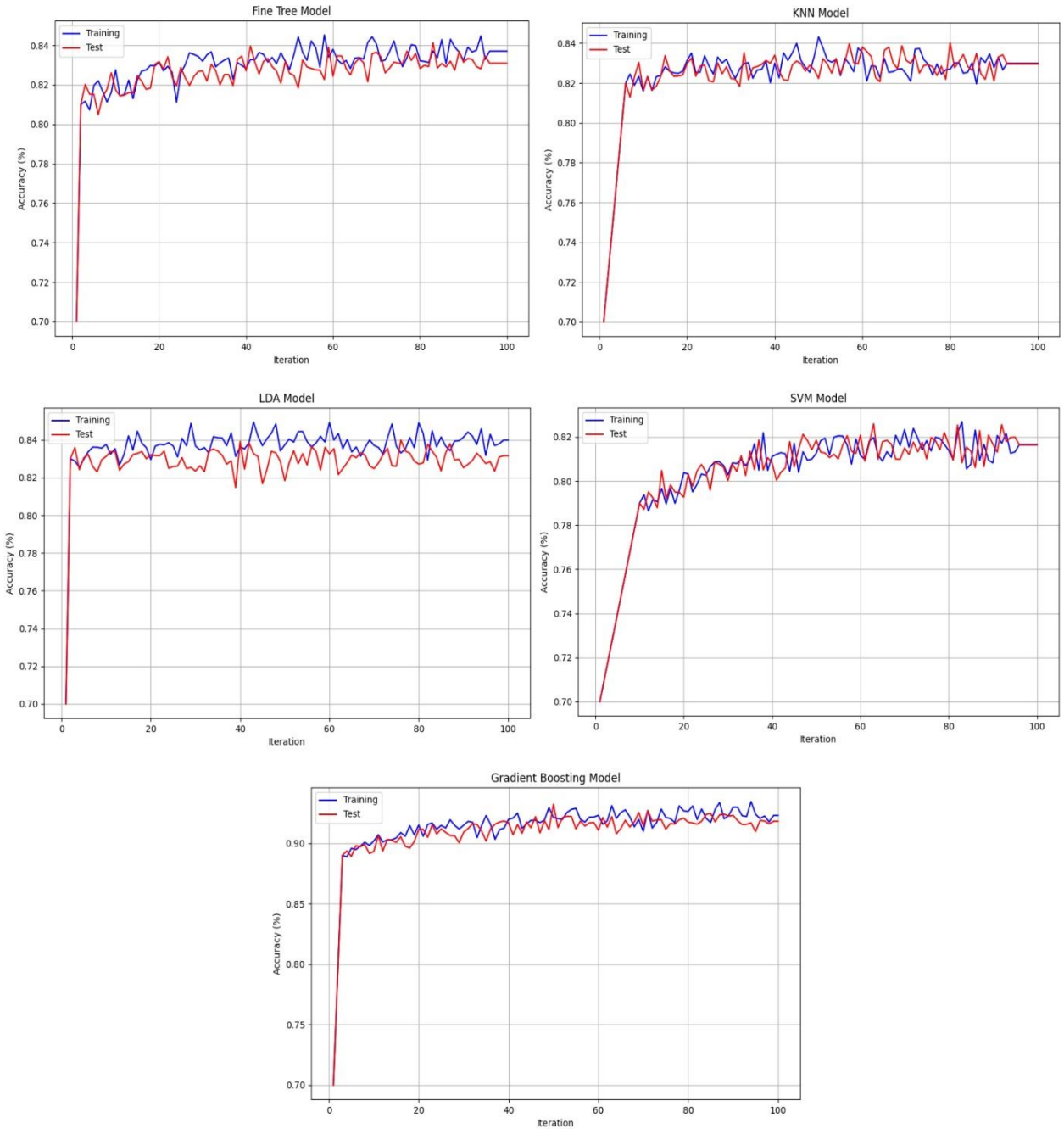
İlgili şekil, tütün kullanım durumunun bireylerin genel sağlık durumu ile ilişkisini incelemektedir. Veriler, her gün tütün kullananlar arasında sağlığı kötü olan bireylerin sayısının, sağlığı iyi olanlara kıyasla daha fazla olduğunu göstermektedir. Tütün kullanımını bırakan bireylerde, sağlıklı olma oranı daha yüksek görünmekte, bu durum tütün kullanımını bırakmanın potansiyel olarak olumlu sağlık etkileri olabileceğini düşündürmektedir. Tütün hiç kullanmamış bireyler genel olarak daha sağlıklı görünürken, ara sıra tütün kullananlarda sağlığı iyi olanların oranı nispeten daha düşük gözlemlenmiştir. Bu bulgular, düzensiz tütün kullanımının bile sağlık üzerinde olumsuz etkiler yaratabileceğini ima edebilir, ancak bu ilişkilerin daha derinlemesine analizlerle desteklenmesi gerekmektedir.

Bununla birlikte, sınıflandırma işlemlerinde özellik seçimi, model performansını artırmak ve gereksiz veya alakasız özellikleri elemek için önemlidir. MRMR (Minimum Redundancy Maximum Relevance) algoritması, bu amaçla kullanılan popüler bir yöntemdir ve adından da anlaşılacağı gibi, seçilen özelliklerin hem birbirleriyle minimum düzeyde tekrar eden bilgi (redundancy) içermesini hem de sınıflandırma hedefi ile maksimum düzeyde ilişkili (relevance) olmasını sağlamayı hedefler (Peng vd., 2005). MRMR algoritması, maksimum ilgililik (relevance) ile hedef değişkenle olan ilişkiyi ve minimum tekrar (redundancy) ile seçilen özelliklerin kendi aralarındaki korelasyonu hesaplayarak çalışır (Estévez vd., 2009). Bu süreçte, mutual information (karşılıklı bilgi) gibi ölçütler kullanılır. Her adımda en fazla bilgi sağlayan ve en az tekrar eden özellikleri seçerek, MRMR algoritması modelin genel performansını artırır ve aşırı öğrenme (overfitting) riskini azaltır. Yüksek boyutlu veri kümelerinde etkili bir şekilde çalışan MRMR, sınıflandırma işlemlerinde dengeli ve verimli sonuçlar üretir (Zhou vd., 2020). Bu çalışmada makine öğrenmesi modelleri ile sınıflandırma işlemi yapılmadan önce, MRMR algoritması ile özellik seçimi yapılmış ve bu özellik dereceleri kullanılarak sınıflandırma işlemleri gerçekleştirilmiştir. Şekil 8'de bu algoritma kullanılarak elde edilen bağımsız değişkenlerin özellik önem dereceleri gösterilmiştir.



Şekil 8. MRMR Algoritmasına Göre Bağımsız Değişkenlerin Önem Sıralamaları

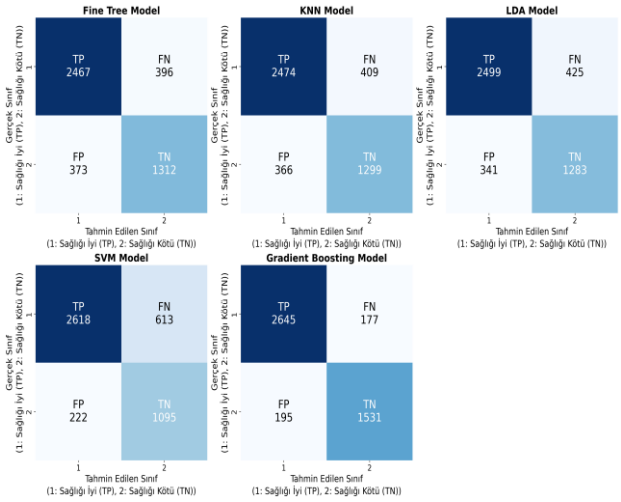
Şekil 8, MRMR algoritmasına göre çeşitli özelliklerin sınıflandırma işlemine olan katkısını skorlar aracılığıyla ifade etmektedir. En yüksek skora sahip olan özellik, 0,4 ile Tütün Kullanım Durumu olarak belirlenmiştir, bu da bu değişkenin bireylerin genel sağlık durumunu sınıflandırmada en etkili faktör olduğunu göstermektedir. İkinci sırada 0,3456 skoruyla Yaş özelliği gelmektedir, bu da yaşın sağlık durumu üzerinde önemli bir etkisi olduğunu vurgular. Günlük Faaliyet Durumu (0,3) ve Hastalık ve Sağlık Durumu (0,24) da yüksek skorlarıyla dikkat çekmektedir, bu da bu özelliklerin sağlık durumu sınıflandırmasında önemli rol oynadığını göstermektedir. Diğer özellikler arasında Medeni Durum (0,202), Çalışma Durumu (0,203), Yaşam Faaliyet Kısıt Durumu (0,201) ve Cinsiyet (0,109) yer almaktadır. Bu sonuçlar, tüm bu özelliklerin makine öğrenmesi modellerinde girdi olarak kullanıldığını ve bireylerin genel sağlık durumunun sınıflandırılmasında etkili olduğunu ortaya koymaktadır. Tütün Kullanım Durumu ve Yaş gibi değişkenlerin yüksek skorları, bu faktörlerin sağlık durumu üzerinde daha belirleyici olduğunu gösterirken, Cinsiyet ve Yaşam Faaliyet Kısıt Durumu gibi değişkenlerin daha düşük skorları, bu faktörlerin sınıflandırma üzerindeki etkisinin nispeten daha az olduğunu ima etmektedir. Bu bulgular, sağlık durumu sınıflandırma modellerinde hangi özelliklerin daha kritik olduğunu anlamada önemli bilgiler sunmaktadır. MRMR algoritması ile özellik skorları değerlendirildikten sonra, bu özelliklerin tamamı makine öğrenmesi modellerine girdi olarak verilmiş ve sınıflandırma işlemleri için tüm modeller 100 iterasyon boyunca eğitilip test edilmiştir. Şekil 9'da tüm modellerin sınıflandırma işlemlerine ilişkin eğitim ve test aşamasında, doğruluk grafikleri gösterilmektedir.



Şekil 9. Sınıflandırma İşleminde Kullanılan Tüm Modellerin Eğitim ve Test Doğruluk Grafikleri

Şekil 9'daki grafikler, genel sağlık durumu sınıflandırması için kullanılan modellerin eğitim ve test doğruluk oranlarını göstermektedir. Fine Tree modeli, 100 iterasyon boyunca eğitimde %83.70, testte %83.09 doğrulukla kararlı bir performans sergilemiştir. KNN modeli, 20. iterasyonda test doğruluğunda hızlı bir artış göstererek %82.96 test doğruluğuna ulaşmış, eğitim doğruluğu ise %82.99 seviyesindedir. LDA modeli, eğitimde %83.99 ve testte %83.16 doğruluk ile en yüksek eğitim doğruluğunu göstermiş, ilk 20 iterasyonda hızlı bir artışın ardından stabil bir performans sergilemiştir. SVM modeli, başlangıçta daha düşük doğrulukla başlamış, 20. iterasyona kadar hızlı bir artış göstermiş, ardından %81.66 eğitim ve %81.64 test doğruluğuyla stabil kalmıştır. En dikkat çekici model olan Gradient Boosting modeli, ilk 10 iterasyonda hızlı bir doğruluk artışı göstererek eğitimde %92.30, testte %91.82 doğruluğa ulaşmış ve iterasyon boyunca bu yüksek doğruluk oranını korumuştur.

Bu sonuçlar, Gradient Boosting modelinin diğer modellere kıyasla daha hızlı ve yüksek doğruluk oranlarına ulaşarak genel sağlık durumu sınıflandırmasında en etkili yöntem olduğunu ve iterasyon boyunca tutarlı bir performans sergilediğini göstermektedir. Bununla birlikte, Şekil 10'da tüm modellerin test setine ilişkin sınıflandırma işlemlerine ait karışıklık matrisleri gösterilmiştir. Karışıklık matrisi, model doğruluğunu değerlendirmek için kullanılan temel ve anlaşılır metriklerden biridir (Cengil ve Çınar, 2020).



Şekil 10. Modellerin Test Sonuçlarına İlişkin Karışıklık Matrisleri

Şekil 10'da sunulan karışıklık matrisleri, beş farklı modelin (Fine Tree, KNN, LDA, SVM ve Gradient Boosting) test sonuçlarını ayrıntılı bir şekilde karşılaştırmaktadır. Fine Tree Modeli, 2467 doğru pozitif (TP) ve 1312 doğru negatif (TN) tahmin ile ortalama bir performans sergilemiştir; ancak 373 yanlış pozitif (FP) ve 396 yanlış negatif (FN) tahmin ile hatalı sınıflandırmalarda belirli bir düzeyde zayıflık göstermiştir. KNN Modeli, benzer bir performansla 2474 TP ve 1299 TN elde etmiş, ancak 366 FP ve 409 FN değerleriyle hala dikkate değer bir hata oranına sahiptir. LDA Modeli ise 2499 TP ve 1283 TN ile doğru sınıflandırma oranlarında nispeten iyi bir performans göstermiştir, fakat 341 FP ve 425 FN oranları modelin yanlış negatif sınıflandırmalarda geliştirilmesi gerektiğini ortaya koymaktadır. SVM Modeli, 2618 TP ve 1095 TN ile diğer modellere kıyasla en yüksek doğru pozitif tahmin oranını yakalamış; ancak 613 FN ve 222 FP ile özellikle negatif sınıfların doğru tahmin edilmesinde diğer modellere göre daha düşük bir performans sergilemiştir. Bu durum, modelin pozitif sınıflandırmalar açısından başarılı olmasına karşın, negatif sınıflandırmalarda zayıf kaldığını göstermektedir. Gradient Boosting Modeli ise 2645 TP ve 1531 TN ile en yüksek genel doğruluk oranını elde etmiştir. Sadece 195 FP ve 177 FN ile yanlış sınıflandırmaları en aza indiren bu model, özellikle hatalı sınıflandırma oranlarını azaltmada diğer modellere kıyasla belirgin bir üstünlük sağlamıştır. Bu sonuçlar, Gradient Boosting Modeli'nin diğer modellere kıyasla daha etkili bir sınıflandırma performansı sergilediğini ve genel olarak daha güvenilir sonuçlar verdiğini göstermektedir. Diğer modellerin ise özellikle yanlış negatif (FN) sınıflandırmaların azaltılması için iyileştirilmesi gerektiği anlaşılmaktadır. Buna ek olarak, modellerin test performansını değerlendiren sınıflandırma raporu Çizelge 2'de sunulmuştur.

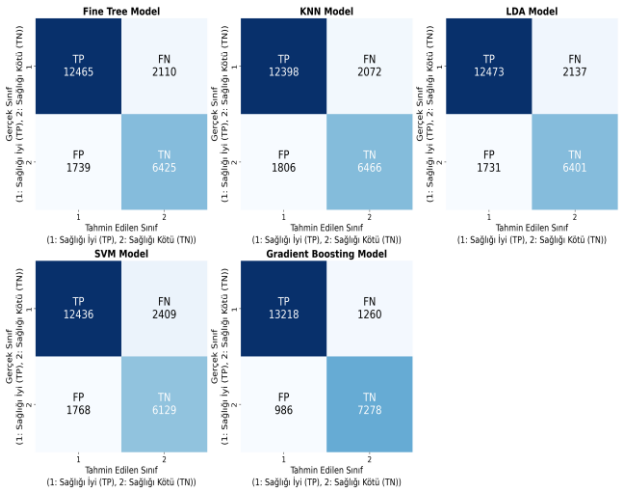
Çizelge 2. Test Sonuçları Sınıflandırma Raporu

Model	Genel Doğruluk	Sınıf	Hassasiyet	Duyarlılık	F1-skoru	Destek
Fine Tree	0.8309	1	0.8687	0.8617	0.865	2840
		2	0.768	0.7786	0.773	1708
		Makro Ort.			0.8193	4548
		Ağırlıklı Ort.			0.8311	4548
KNN	0.8296	1	0.8711	0.8581	0.8646	2840
		2	0.7605	0.7802	0.7702	1708
		Makro Ort.			0.8174	4548
		Ağırlıklı Ort.			0.8300	4548
LDA	0.8316	1	0.8799	0.8547	0.8671	2840
		2	0.7512	0.7900	0.7701	1708
		Makro Ort.			0.8186	4548
		Ağırlıklı Ort.			0.8325	4548
SVM	0.8164	1	0.9218	0.8103	0.8671	2840
		2	0.6411	0.8314	0.7240	1708
		Makro Ort.			0.7932	4548
		Ağırlıklı Ort.			0.8224	4548
GB	0.9182	1	0.9313	0.9373	0.9343	2840
		2	0.8964	0.8870	0.8917	1708
		Makro Ort.			0.9130	4548
		Ağırlıklı Ort.			0.9181	4548

Tablo 2'deki modellerin test sonuçlarına göre, genel sağlık durumu sınıflandırmasında Gradient Boosting modeli, %91.82 genel doğruluk oranıyla en yüksek performansı sergilemiştir. Sınıf 1 (sağlığı iyi) tahminlerinde SVM modeli, en yüksek hassasiyeti (%92.18) ve F1 skorunu (%86.71) sağlarken, LDA modeli en yüksek duyarlılığı (%87.99) göstermiştir. Sınıf 2 (sağlığı kötü) tahminlerinde ise Gradient Boosting modeli, hem hassasiyet (%89.64),

duyarlılık (%88.70) hem de F1 skoru (%89.17) açısından diğer modellere üstün gelmiştir. Genel olarak, Gradient Boosting modeli, her iki sınıfta da yüksek doğruluk, hassasiyet, duyarlılık ve F1 skoru ile en iyi performansı göstermiş, özellikle sınıf 2 tahminlerinde başarılı olmuştur. Diğer modeller de makul performans sergilemiş, ancak Gradient Boosting modeli, genel doğruluk ve dengeli sınıflandırma yetenekleriyle öne çıkmıştır.

Tablo 2'deki sonuçlara ek olarak, modellerin performansını daha güvenilir bir şekilde değerlendirmek amacıyla 5 katlı çapraz doğrulama uygulanmıştır. Çapraz doğrulama, modellerin genelleme yeteneklerini teyit etmek ve overfitting (aşırı öğrenme) sorunlarını tespit etmek için kritik bir yöntemdir. Bu yöntemle, tüm veri seti 5 eşit parçaya bölünmüş ve her seferinde bir parça test seti olarak kullanılarak model eğitilmiştir. Bu süreç 5 kez tekrarlanmış ve her bir tekrarda farklı bir bölüm test seti olarak kullanılmıştır. Böylece her modelin performansı, veri setinin farklı bölümlerinde test edilerek ortalama performans değerleri elde edilmiştir. 5 katlı çapraz doğrulama, modellerin daha geniş veri kümesi üzerindeki tutarlılıklarını ve güvenilirliklerini artırarak, sonuçların genel geçerliliğini sağlamıştır. Bu analiz, her bir modelin sadece belirli bir veri kümesi üzerinde değil, tüm veri setinde nasıl performans gösterdiğini değerlendirmeye olanak tanımıştır. Bu sayede, modellerin gerçek dünya verileri üzerinde ne kadar başarılı olabilecekleri daha iyi anlanabilir ve model seçimi daha doğru yapabilir. Şekil 10'da modellerin çapraz doğrulama sonuçlarına ilişkin karışıklık matrisleri gösterilmiştir.



Şekil 11. Tüm Modellerin Çapraz Doğrulama Sonuçlarına İlişkin Karışıklık Matrisleri

Şekil 11'de sunulan karışıklık matrisleri, Fine Tree, KNN, LDA, SVM ve Gradient Boosting modellerinin 5 katlı çapraz doğrulama aşamasındaki performanslarını karşılaştırmaktadır. Fine Tree Modeli, 12465 doğru pozitif (TP) ve 6425 doğru negatif (TN) tahminlerle makul bir performans sergilemiştir; ancak 1739 yanlış pozitif (FP) ve 2110 yanlış negatif (FN) tahminle, modelin hatalı sınıflandırmalarda belirli bir zayıflığa sahip olduğu görülmektedir. KNN Modeli, benzer şekilde, 12398 TP ve 6466 TN ile doğru sınıflandırmalarda başarılı olmuş, ancak 1806 FP ve 2072 FN değerleriyle hala iyileştirme gerektiren bir hata oranına sahiptir. LDA Modeli, 12473 TP ve 6401 TN ile doğru sınıflandırmalarda tutarlı bir performans göstermiş olsa da, 1731 FP ve 2137 FN değerleri, özellikle negatif sınıflandırmalar açısından modelin performansının artırılabileceğini göstermektedir. SVM Modeli, 12436 TP ve 6129 TN değerleri ile doğru sınıflandırmalarda nispeten güçlü bir performans sergilemiş, ancak 2409 FN ve 1768 FP ile diğer modellere kıyasla daha fazla yanlış sınıflandırma yapmıştır. Bu, özellikle yanlış negatif sınıflandırmalarda modelin zayıf kaldığını ortaya koymaktadır. Gradient Boosting Modeli ise 13218 TP ve 7278 TN ile tüm modeller arasında en yüksek doğruluğa ulaşmıştır; sadece 986 FP ve 1260 FN değerleriyle diğer modellere göre çok daha düşük bir hata oranı sunmuştur. Bu sonuçlar, Gradient Boosting Modeli'nin diğer modellere kıyasla veriyi daha etkili bir şekilde modellediğini ve özellikle hatalı sınıflandırmaları minimize etmede üstün bir performans sergilediğini göstermektedir. Genel olarak, Gradient Boosting Modeli'nin, sınıflandırma doğruluğu ve hata oranlarını minimize etme açısından en başarılı model olduğu sonucuna varılabilir. Diğer modellerin ise özellikle negatif sınıfların doğru tahmin edilmesi konusunda geliştirilmesi gerekmektedir.

5. Sonuçlar ve Tartışma

Günümüzde sağlık sektörü, büyük veri analizi ve veri madenciliği uygulamalarıyla önemli bir dönüşüm yaşamaktadır. Sağlık istatistikleri, bu alandaki kilit bilgileri içermekte olup, doğru analiz ve sınıflandırma stratejileri kullanılarak bu verilerden elde edilen bilgiler, sağlık hizmetlerinin planlanması, hastalıkların öngörülmesi ve genel sağlık politikalarının geliştirilmesi konularında önemli katkılarda bulunabilir. Mikro veri setleri ve veri madenciliği yaklaşımları, büyük bir potansiyel taşımaktadır. Bu makalede kullanılan veri seti, birçok farklı disiplindeki araştırmacılar, analistler ve karar alıcılar için önemli bilgilere sahip bir kaynaktır.

Bu çalışma, sağlık istatistiklerinin veri madenciliği teknikleri kullanılarak analiz edilmesi ve genel sağlık durumunun makine öğrenmesi yöntemleri ile sınıflandırılması üzerine odaklanmıştır. Bu bağlamda, Türkiye İstatistik Kurumu (TÜİK) tarafından sağlanan 2022 yılı mikro verileri kullanılarak çeşitli makine öğrenmesi

modelleri değerlendirilmiştir. Çalışmanın bulguları, Gradient Boosting algoritmasının sağlık durumu sınıflandırmasında diğer modellere kıyasla üstün performans gösterdiğini ortaya koymaktadır. Bu model, yüksek doğruluk, hassasiyet, duyarlılık ve F1 skoru ile öne çıkmıştır ve özellikle sağlığı kötü olan bireylerin doğru bir şekilde tespit edilmesinde etkili olmuştur. Gradient Boosting modelinin öne çıkmasının yanı sıra, diğer modellerin performansları da dikkate değer bulunmuştur. Fine Tree, KNN, LDA ve SVM modelleri de makul doğruluk oranlarına ulaşmış, ancak Gradient Boosting modeli ile karşılaştırıldığında daha düşük performans sergilemişlerdir. Bu sonuçlar, Gradient Boosting modelinin sağlık durumu gibi karmaşık ve çok boyutlu veri setlerinde güvenilir ve doğru tahminler yapabilme kapasitesini göstermektedir.

Bu çalışmanın sonuçları, çeşitli demografik ve sosyal faktörlerin bireylerin genel sağlık durumları üzerinde belirgin etkileri olduğunu göstermektedir. Yaş, cinsiyet, medeni durum, tütün kullanımı ve çalışma durumu gibi değişkenler, sağlık durumu üzerinde önemli rol oynamaktadır. Özellikle tütün kullanımının azaltılması ve yaşam tarzı değişiklikleri konusunda politika yapıcılar ve sağlık profesyonelleri için faydalı veriler sunulmaktadır. Bunun yanı sıra, bulgular, sağlık eğitimi programlarının geliştirilmesi ve hedeflenen müdahaleler için bir temel oluşturabilir. Örneğin, yaşlı bireyler ve kronik rahatsızlıkları olanlar için özel sağlık hizmetleri ve destek programları tasarlanabilir. Ayrıca, genç yetişkinler ve çalışan nüfus için sağlıklı yaşam tarzı alışkanlıkları teşvik eden kampanyalar düzenlenmelidir.

Öte yandan, literatürde yer alan ve bu çalışmada incelenen makaleler, sağlık sektöründe veri madenciliği ve sınıflandırma yöntemlerinin çeşitli uygulamalarını ele almaktadır. Örneğin, biyomedikal veri kümeleri üzerinde farklı makine öğrenmesi algoritmalarının etkinliğinin karşılaştırıldığı çalışmalar, sağlık sigortası dolandırıcılığının tespiti için algoritmaların kullanıldığı analizler, ve farklı demografik faktörlerin sağlık hizmeti talebi üzerindeki etkisini inceleyen araştırmalar bulunmaktadır. Ayrıca, sosyo-ekonomik faktörlerin sağlık ihtiyaçlarına etkilerini belirlemeye yönelik çalışmalar ve Gradient Boosting yöntemi kullanılarak yapılan sınıflandırma analizleri gibi çeşitli konulara değinilmiştir. Literatürde yapılan diğer çalışmalarla karşılaştırıldığında, bu çalışma, Gradient Boosting algoritmasının sağlık istatistikleri veri setlerinde etkili bir sınıflandırma aracı olduğunu doğrulamaktadır. Örneğin, Akbar ve arkadaşlarının (2020) sağlık sigortası dolandırıcılığı tahmininde, Wang ve arkadaşlarının (2022) petrokimya tesislerinin insan sağlığı üzerindeki risklerini tespit etmede, ve Theerthagiri ve Vidya'nın (2022) kardiyovasküler hastalık tahmininde Gradient Boosting algoritmasının üstün performans gösterdiği çalışmalarla uyumlu bulgular elde edilmiştir.

Bu çalışma, sağlık alanında makine öğrenmesi ve veri madenciliği tekniklerinin uygulanabilirliğini ortaya koymaktadır. Özellikle Gradient Boosting modelinin sağlık verilerinde yüksek performans göstermesi, literatürdeki benzer çalışmalara paralel olarak değerlendirilebilir. Örneğin, Kardiyotokogram verileri kullanılarak fetal sağlığın sınıflandırılması, XGBoost'un en yetkin sınıflandırıcı olarak ortaya çıktığını ve birden fazla metrikte sürekli olarak diğerlerinden daha iyi performans gösterdiğini tespit eden araştırmacılar tarafından araştırılmıştır (Alkurdi, 2024). Bu çalışma, fetal sağlık takibinde devrim yaratarak fetal koşulların daha güvenilir ve objektif bir şekilde değerlendirilmesini sağlayan makine öğrenimi vaadinin altını çizmektedir. Sağlık sonuçlarının tahmin edilmesinde makine öğreniminin uygulanması, Etiyopya'da beş yaş altı çocukların yetersiz beslenmesine odaklanan Fenta vd. (2021)'in çalışmasıyla da desteklenmektedir. Çalışmada, beslenme durumunun belirleyicilerini tespit etmek için rastgele orman sınıflandırıcıları kullanılmış ve bu da makine öğreniminin halk sağlığı araştırmalarındaki etkinliğini ortaya koymuştur (Fenta vd., 2021). Bulgular, sosyo-ekonomik faktörlerin çocuk sağlığı sonuçlarını etkilemedeki önemini vurgulamaktadır. Bu çalışmalar, mevcut çalışmanın sonuçlarını desteklemekte ve makine öğrenmesi tekniklerinin sağlık verileri üzerinde uygulanabilirliğini güçlendirmektedir.

Bu çalışmanın bulguları, politika yapıcılar ve sağlık profesyonelleri için önemli çıkarımlar sunmaktadır. Tütün kullanımının azaltılması, sağlıklı yaşam tarzı alışkanlıklarının teşvik edilmesi ve yaşlı bireyler ile kronik hastalıkları olanlara yönelik özel sağlık hizmetleri ve destek programları tasarlanması gerekmektedir. Ayrıca, genç yetişkinler ve çalışan nüfus için sağlıklı yaşam tarzı alışkanlıklarını teşvik eden kampanyalar düzenlenmeli ve sağlık eğitimi programları geliştirilmelidir. Bu politikalar, toplum sağlığının iyileştirilmesine ve sağlık hizmetlerinin daha verimli kullanılmasına katkı sağlayacaktır.

Çalışmanın bazı kısıtlılıkları bulunmaktadır. Veri seti yalnızca belirli demografik ve sosyo-ekonomik değişkenleri içermektedir, bu nedenle daha geniş kapsamlı verilerin analizi ile sonuçlar daha da güçlendirilebilir. Ayrıca, çalışmada yalnızca belirli makine öğrenmesi modelleri değerlendirilmiştir; farklı ve daha yeni algoritmaların kullanımı ile daha yüksek performans elde edilebilir. Gelecek çalışmalarda, farklı veri setlerinin ve daha geniş değişken yelpazesinin kullanılması, ayrıca

derin öğrenme yöntemlerinin dahil edilmesi önerilmektedir. Bu yaklaşımlar, sağlık durumunun daha doğru ve kapsamlı bir şekilde sınıflandırılmasına olanak tanıyacaktır. Bununla birlikte, bu faktörlerin sağlık üzerindeki etkilerini daha derinlemesine incelemeli ve özellikle düşük sağlık durumu bildiren gruplara yönelik müdahaleleri geliştirmek hedeflenmelidir. Ayrıca, farklı yaş grupları için sağlık hizmetlerine erişim ve kullanımını iyileştirmeye yönelik politikaların oluşturulması gerekmektedir. Sonuç olarak, bu çalışma, veri madenciliği ve makine öğrenmesi tekniklerinin sağlık sektöründe ne denli etkili olabileceğini göstermiş ve Gradient Boosting algoritmasının genel sağlık durumu sınıflandırmasında üstün performansını ortaya koymuştur. Bu bulgular, sağlık politikalarının geliştirilmesi ve halk sağlığı programlarının iyileştirilmesi için değerli bilgiler sunmaktadır.

Etik Standartlar Bildirgesi

Bu çalışmanın hazırlanma sürecinde bilimsel ve etik ilkelere uyulduğu ve yararlanılan tüm çalışmaların kaynakçada belirtildiği beyan olunur.

Yazarlık Katkı Beyanı

Yazar 1: Kavramsallaştırma, Veri İyileştirme, Analiz ve Yorumlama, Görselleştirme, Yazma – orijinal taslak

Yazar 2: Araştırma, Fikir Sahibi, Analiz ve Yorumlama, Doğrulama, Görselleştirme, Yazma – orijinal taslak,

Yazar 3: Metodoloji, Doğrulama, Biçimsel Analiz, Görselleştirme, Denetleme, Yazma –orijinal taslak

Çıkar Çatışması Beyanı

Yazarların bu makalenin içeriğiyle ilgili olarak beyan edecekleri hiçbir çıkar çatışması yoktur.

Verilerin Kullanılabilirliği/ Data Availability

Bu çalışma sırasında oluşturulan veya analiz edilen tüm veriler, yayınlanan bu makaleye dahil edilmiştir.

6. Kaynaklar

Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine learning applications based on SVM classification a review. *Qubahan Academic Journal*, **1(2),** 81-90.

Akbar, N. A., Sunyoto, A., Arief, M. R., & Caesarendra, W. (2020). *Improvement of Decision Tree Classifier Accuracy for Healthcare Insurance Fraud Prediction by Using Extreme Gradient Boosting Algorithm*. 2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, pp. 110-114. IEEE.

Alkurdi, A. and Abdulazeez, D. A. M., 2024. Comprehensive classification of fetal health using cardiotocogram data based on machine learning. *Indonesian Journal of Computer Science*, **13(1)**. https://doi.org/10.33022/ijcs.v13i1.3718.

Alptekin, N., & Yeşilaydın, G., 2015. OECD ülkelerinin sağlık göstergelerine göre bulanık kümeleme analizi ile sınıflandırılması. *İşletme Araştırmaları Dergisi*, **7(4)**, 137-155.

Altıntaş YY. 2010. Veri madenciliğinin tıpta kullanımı ve bir uygulama: hemodiyaliz hastaları için risk seviyelerine göre risk faktörlerinin etkileşimlerinin incelenmesi. Ulusal Tez Merkezi, 269710: 1-3.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. 2021. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, **54,** 1937-1967.

Bisht, R. K., & Bisht, I. P. (2022). *Investigation of the Role of Test Size, Random State, and Dataset in the Accuracy of Classification Algorithms*. International Conference on Communication and Intelligent Systems, Singapore, pp. 715-726. Springer Nature Singapore.

Cengil, E. & Çınar A., 2020. Göğüs Verileri Metrikleri Üzerinden Kanser Sınıflandırılması, *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi,* **11(2)**, ss. 513-519.

Cheng, D., Zhang, S., Liu, X., Sun, K., & Zong, M., 2017. Feature selection by combining subspace learning with sparse representation. *Multimedia Systems,* **23,** 285-291.

Chung, J., & Teo, J., 2023. Single classifier vs. ensemble machine learning approaches for mental health prediction. *Brain informatics,* **10(1)**, 1-10.

Çiçek, A. ve Arslan, Y., 2020. Müşteri Kayıp Analizi İçin Sınıflandırma Algoritmalarının Karşılaştırılması. *İleri Mühendislik Çalışmaları Ve Teknolojileri Dergisi*, **1(1),** 13-19.

Doğan, E., 2020. Gelir Düzeyi ve Sağlık Hizmet Talebi İlişkisi: Mikro Veriler ile Türkiye Örneği. *MANAS Sosyal Araştırmalar Dergisi*, **9(4),** 2376-2392. https://doi.org/10.33206/mjss.705718

Estévez, P. A., Tesmer, M., Perez, C. A., & Żurada, J. M. 2009. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, **20(2)**,189-201. https://doi.org/10.1109/tnn.2008.2005601

Fenta, H. M., Zewotir, T., & Muluneh, E. K. 2021. A machine learning classifier approach for identifying the determinants of under-five child undernutrition in ethiopian administrative zones. *BMC Medical Informatics and Decision Making*, **21(1),** 291. https://doi.org/10.1186/s12911-021-01652-1.

Genç, B. U. G., & Kurutkan, M. N. (2021). Eşitsizlik Bağlamında Karşılanmayan Sağlık İhtiyacı: Türkiye Sağlık Araştırması Verilerinden Kanıtlar. *SDÜ Sağlık Yönetimi Dergisi,* **3(1**), 34-51.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). *KNN Model-Based Approach in Classification.* OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy,

November 3-7, 2003, pp. 986-996. Springer Berlin Heidelberg.

Hossin M., and Sulaiman M. N., (2015). A review on evaluation metrics for data classification evaluations, *International Journal of Data Mining & Knowledge Management Process*, **5(2)**, ss. 1.

Karaca İ. (2015). Büyük Veri Analizlerinin Kurumsal Faaliyetlerde Kullanım Alanları, Lisans Tezi, Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Bilgi ve Belge Yönetimi Bölümü, Ankara.

Karakoyun, M., & Hacibeyoğlu, M. (2014). Biyomedikal Veri Kümeleri İle Makine Öğrenmesi Siniflandirma Algoritmalarinin İstatistiksel Olarak Karşilaştirilmasi. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, **16(48**), 30-42.

Kaya, I. (2021). Comparison of spectral and template matching features for ssvep bci target frequency classification. *International Journal of Intelligent Systems and Applications in Engineering*, **9(2**), 64-68. https://doi.org/10.18201/ijisae.2021.235.

Kayakuş, M. & Yiğit Açıkgöz, F. (2023). Twitter'da Makine Öğrenmesi Yöntemleriyle Sahte Haber Tespiti. *Abant Sosyal Bilimler Dergisi*, **23(2)**, 1017-1027. https://doi.org/10.11616/asbi.1266179

Kızgın, M. S., Çambay, Z., Sepet, H., Özçelik, S. T. A., & Uyanık, H. (2023). Onobrychis Bitkisine Ait Meyve Tiplerinin Makine Öğrenmesi Yaklaşımıyla Sınıflandırılması. Fırat Üniversitesi Fen Bilimleri Dergisi, **35(2),** 87-96.

Koçak, A., & Ergün, M. A. (2023). Sağlıkta veri kalitesi ve veri madenciliği uygulamaları. *Disiplinlerarası Yenilik Araştırmaları Dergisi,* **3(1),** 23-30.

Koyuncugil, A., & Özgülbaş, N. (2009). Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. *Bilişim Teknolojileri Dergisi*, **2(2)**, 21-32

Mandelkow, H., De Zwart, J. A., & Duyn, J. H. (2016). Linear discriminant analysis achieves high classification accuracy for the BOLD fMRI response to naturalistic movie stimuli. *Frontiers in human neuroscience,* **10,** 128.

Mayr, A., Binder, H., Gefeller, O., & Schmid, M. (2014). The evolution of boosting algorithms. *Methods of information in medicine,* **53(6)**, 419-427.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27(8)**, 1226-1238. https://doi.org/10.1109/tpami.2005.159

Rathi, V. P., & Palani, S. (2012). Brain tumor MRI image classification with feature selection and extraction using linear discriminant analysis. arXiv preprint arXiv:1208.2128.

Stein, G., Chen, B., Wu, A. S., & Hua, K. A. (2005). Decision Tree Classifier for Network Intrusion Detection with GA-Based Feature Selection. 43rd Annual Southeast Regional Conference, Kennesaw, GA, USA, pp. 136-141.

Subasi, A., & Gursoy, M. I. (2010). EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert systems with applications*, **37(12),** 8659-8666.

Terzi, M. (2019). Türkiye'de Sağlık Sektöründe Veri Madenciliği Kullanım Alanları. *Black Sea Journal of Health Science,* **2(2)**, 45-48.

Theerthagiri, P., & Vidya, J. (2022). Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques. *Expert Systems*, **39(9)**, e13064.

Tripathi, A., Kumar, K., Misra, A., & Chaurasia, B. K. (2023). Colon Cancer Tissue Classification Using ML. 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, pp. 1-6.

Türkiye Sağlık Araştırması 2022 Yılı Mikro Veri Seti, (2023). Yayın No: 4702, ISBN: 978-625-8368-43-7, Yayın Tarihi: Temmuz 2023, *Türkiye İstatistik Kurumu*, Ankara.

Wang, M., Li, X., Lei, M., Duan, L., & Chen, H. (2022). Human health risk identification of petrochemical sites based on extreme gradient boosting. *Ecotoxicology and Environmental Safety*, **233,** 113332.

Worth, A. P., & Cronin, M. T. (2003). The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *Journal of Molecular Structure: THEOCHEM*, **622(1-2),** 97-111.

Wu, J., Song, L., Wang, T., Zhang, Q., & Yuan, J. (2020). F*orest r-cnn: large-vocabulary long-tailed object detection and instance segmentation*. 28th ACM International Conference on Multimedia, Seattle, WA, USA (Online).

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and information systems,* **14**, 1-37.

Xu, X., Lin, M., & Xu, T. (2022). Epilepsy seizures prediction based on nonlinear features of EEG signal and gradient boosting decision tree. International Journal of *Environmental Research and Public Health*, **19(18)**, 11326.

Yıldıztepe, E. ve Kocataş, A. (2018). Türkiye işgücü verilerinin karar ağacı yöntemleriyle analizi. Çankırı *Karatekin Üniversitesi İİBF Dergisi.* **8** (2), 91-114.

Yılmaz, E. (2012). İstatiksel Analiz Yöntemi Olarak Veri Madenciliğinde Chaıd Algoritması ve Türkiye'de İşgücü Piyasasının Durumunun Ve Bunun Nedenlerinin Belirlenmesine İlişkin Bir Uygulama, Yüksek Lisans

Tezi, Yıldız Teknik Üniversitesi Sosyal Bilimler Enstitüsü İşletme Ana Bilim Dalı, İstanbul.

Yin, H., Sharma, B., Hu, H., Liu, F., Kaur, M., Cohen, G., ... & Eckel, S. P. (2024). Predicting the climate impact of healthcare facilities using gradient boosting machines. *Cleaner Environmental Systems*, **12**, 100155.

Yongcharoenchaiyasit, K., Arwatchananukul, S., Temdee, P., & Prasad, R. (2023). Gradient Boosting Based Model for Elderly Heart Failure, Aortic Stenosis, and Dementia Classification. *IEEE Access*. 11, 48677-48696, https://doi.org/10.1109/ACCESS.2023.3276468

Yu, H., & Kim, S. (2012). SVM Tutorial-Classification, Regression and Ranking. *Handbook of Natural computing,* **1**, 479-506.

Yue, S., Li, P., & Hao, P. (2003). SVM classification: Its contents and challenges. *Applied Mathematics-A Journal of Chinese Universities,* **18**, 332-342.

Zhou, H., Wang, X., & Zhang, Y. (2020). Feature selection based on weighted conditional mutual information. *Applied Computing and Informatics,* **20**(1/2), 55-68. https://doi.org/10.1016/j.aci.2019.12.003