Fırat Üniversitesi Müh. Bil. Dergisi
35(2), 433-442, 2023

*Araştırma Makalesi*
https://doi.org/10.35234/fumbd.1234161

# Komşuluk Bileşen Analizi Tabanlı Makine Öğrenimi Yöntemleri ile Obezite Seviyelerinin Tahmini

**Çağla DANACI[1,2*], Derya AVCI[3], Seda ARSLAN TUNCER[4]**
[1] Yazılım Mühendisliği, Fen Bilimleri Enstitüsü, Fırat Üniversitesi, Elazığ, Türkiye
[2]Yazılım Mühendisliği, Teknoloji Fakültesi, Sivas Cumhuriyet Üniversitesi, Sivas, Türkiye
[3] Yazılım Mühendisliği, Teknoloji Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye
[4] Yazılım Mühendisliği, Mühendislik Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye
[*1,2] cdanaci@firat.edu.tr, [3] davci@firat.edu.tr, [4] satuncer@firat.edu.tr

**(Geliş/Received: 14/01/2023; Kabul/Accepted: 01/04/2023)**

**Öz:** Obezite, gelişmiş ülkelerde daha çok görülmekle birlikte gelişmekte olan ülkelerde de yaşam kalitelerini olumsuz yönde etkileyen bir hastalıktır. Obeziteyi tetikleyen birden çok etmen bulunmakla beraber bu etmenlerin en somut örneklerinden bazıları hareketsiz yaşam, dengesiz beslenme olarak sıralanabilir. Obezite, hastalar için farklı düzeylerde görülebilmektedir. Her düzey, tedavi aşamasında arz ettiği önem ile obezite tanısının erken aşamada belirlenme ihtiyacını doğurmaktadır. Bu doğrultuda uzmanlara karar aşamasında yardımcı olabilecek otonom bir sistem bu ihtiyaca destek niteliğinde tasarlanmıştır. Çalışmada obezite hastalarını, obezite düzeylerine göre sınıflandırabilmek amacıyla makine öğrenimi tabanlı bir yaklaşım önerilmiştir. UCI makine öğrenimi deposundan 16 özelliğe sahip 2111 hasta verisi üzerinde Komşuluk Bileşen Analizi (KBA) yöntemi ile özellik seçimi yapılarak özellikler Aşırı Gradyan Artırma (XGBoost) ve Karar Ağacı algoritmaları ile sınıflandırılmıştır. Sonuçlar incelendiğinde özellik seçimi sonrası doğruluk oranlarında iki algoritma için de %1 artış gözlemlenmiştir. Sistemin amaca uygun olarak performans sergilemesi sebebiyle, obezite düzey tahmininde optimum özellik sayısı ile uzmanlara yardımcı bir çalışma olacağı öngörülmektedir.

**Anahtar kelimeler:** Obezite, Yapay Zekâ, Makine Öğrenmesi, Özellik Seçimi, Sağlık.

# Prediction of Obesity Levels by Neighborhood Component Analysis Based Machine Learning Methods

**Abstract:** Although obesity is more common in developed countries, it is also a disease in developing countries that negatively affects people's quality of life. Although there are several factors that trigger obesity, some of the most concrete examples of these factors can be mentioned: sedentary lifestyle, unbalanced diet. Obesity can show up in patients at different levels. Each of these levels is important for treatment, so the diagnosis of obesity must be made early. With this in mind, an autonomous system was developed that can assist experts in decision making. In the study, a machine learning-based approach was proposed to classify obese patients according to their degree of obesity. Feature selection was made with Neighborhood Component Analysis (NCA) method on 2111 patient data with 16 features from UCI machine learning repository, and features were classified with Extreme Gradient Augmentation (XGBoost) and Decision Tree algorithms. When the results were examined, a 1% increase in accuracy rates was observed after feature selection for both algorithms. Since the system works according to the purpose, it is expected that a study will be conducted to help experts with the optimal number of features for predicting the degree of obesity.

**Key words:** Obesity, Artificial Intelligence, Machine Learning, Feature Selection, Health.

## 1. Giriş

Obezite, dünya genelinde beraberinde getirdiği etmenler ile insanların yaşam kalitesini negatif anlamda etkileyen bir sağlık sorunudur. Özellikle gelişmiş ülkelerde her yaştan ve cinsiyetten bireyleri önemli ölçüde etkileyen obezite, genç ve orta yaşlı bireylerde daha sık rastlanan bir hastalıktır [1]. Obezitenin etiyolojik yapısı tam olarak bilinmemekle birlikte gerçekleştirilen araştırmalara göre vücut kitle indeksinin 30 $kg/m^2$ üzerinde olması obezitenin temelini oluşturmaktadır [2]. Obezite oluşumunu etkileyen birçok faktör bulunmaktadır. Bunların başında yüksek kalori alımı, hareketsiz yaşam, sağlıksız yiyecek-içecek tüketimi gibi somut faktörlerin yanı sıra psikolojik faktörler de obeziteyi tetiklemektedir. Tüm bu faktörlerin yol açtığı aşırı kilo alımıyla birlikte obezite riskinin artması, obezite tanısının ve kontrolünün erken aşamada belirlenmesi ihtiyacını doğurmaktadır

* Sorumlu yazar: cdanaci@firat.edu.tr. Yazarların ORCID Numarası: [1,2] 0000-0003-2414-1310, [3] 0000-0002-5204-0501, [4] 0000-0001-6472-8306

[3]. Teknoloji bu tür ihtiyaçların giderilmesi için günümüzde en sık başvurulan yöntemlerden biridir. Teknolojinin gösterdiği hızlı gelişim insan yaşamına yön vermekle beraber birçok farklı alanı içinde barındırarak problemlere çözüm üretmektedir. Teknoloji alt dallarından biri olan yapay zekâ en sık karşımıza çıkan kavramlardan biridir.

Yapay zekâ insana ait öğrenme yapısından esinlenerek, beynin eldeki bilgiden bir probleme yönelik çözüm üretme, karar verebilme yeteneğini kullanan sistemler olarak tanımlanabilir [4]. Sahip olduğu makine öğrenmesi, derin öğrenme, uzman sistemler, doğal dil işleme gibi alt dallar ile birçok farklı alanda faaliyet gösteren yapay zekânın en etkin kullanıldığı alanlardan biri sağlıktır [5]. Sağlık alanında teşhis ve tedavi aşamalarında uzman kararına başvurulmakla beraber, uzman kararını destekleyecek, iş gücü açısından maksimum fayda sağlayacak, zaman kavramını minimuma indirgeyecek bir destek sistemi yapay zekâ teknolojileri ile sağlanmaktadır [6,7]. Obezite açısından incelendiğinde ise yapay zekâ obezite teşhisi, kilo kontrolü, ilaç kullanımı vb. alanlarda uzmanlara yardımcı olmaktadır. Literatür incelendiğinde yapay zekâ yöntemleri ile obezite teşhisi, obezite seviyesi belirleme gibi birçok alanda gerçekleştirilen çalışmalar bulunmaktadır. Bu çalışmalardan bazıları şu şekildedir:

Cervantes ve Palacio makine öğrenimi yöntemleriyle obezite seviyelerini belirlemek için Kolombiya, Meksika ve Peru ülkelerinden 18-25 yaş arası, 81 erkek ve 97 kadın toplamda 178 öğrenciyi çalışmaya dâhil etmişlerdir. Algoritmaları verilere uygulayabilmek adına WEKA aracını kullanarak Karar Ağacı algoritması ve Destek Vektör Makineleri (DVM) ile sınıflandırma işlemini gerçekleştirmişlerdir. Performans değerlendirme aşamasında Kesinlik, Gerçek Pozitif Oranı, Yanlış Pozitif Oranı ve Roc Alanı metriklerini kullanarak algoritma performanslarını değerlendirmişlerdir. Son adımda her sınıflandırma algoritmasının sonucunu, kümeleme algoritması ile birleştirerek K-Means tekniğine dayalı bir akıllı sistem elde etmişlerdir. K-Means yöntemi ile obezite seviyelerini dört kümeye ayırarak küme bazlı Karar Ağaçlarını kullanarak yeniden sınıflandırma işlemi gerçekleştirmişlerdir. Bu çalışma sonucunda oluşturdukları K-Means + Karar Ağaçları yapısı ile %99.5 Roc Alanı değerini elde etmişlerdir. Elde ettikleri akıllı sistemin sonuçlarına paralel olarak bu sistemin farklı hastalıkların da incelenmesinde öncü olabileceğini belirtmişlerdir [8].

Ferdowsy ve arkadaşları obezite riskini tahmin edebilmek için makine öğrenimi temelli bir yaklaşım önermişlerdir. Çalışmaya farklı yaşlardan 1100'den fazla hasta verilerini dâhil etmişlerdir. Toplamda 9 farklı makine öğrenimi algoritması (K-En Yakın Komşu (KNN), Rastgele Orman, Lojistik Regresyon (LR), Çok Katmanlı Algılayıcı (ÇKA), DVM, Naive Bayes, AdaBoost, Karar Ağaçları ve Gradyan Artırma) ile tahmin işlemini gerçekleştirerek algoritmaların performanslarını karşılıklı olarak inceleyip en iyi modele karar vermişlerdir. Çalışma sonucunda en iyi performansı %97.07 doğruluk oranıyla Lojistik Regresyon (LR) ile elde ederken çözmeye çalıştıkları problem için en düşük başarıma sahip algoritmanın %64.08 ile Gradyan Artırma algoritması olduğunu belirtmişlerdir [9].

Cui ve arkadaşları UCI makine öğrenimi deposundan aldıkları obetize düzeylerini içeren açık erişim veri setini kullanarak makine öğrenimi yöntemleri ile obezite tahmini gerçekleştirmişlerdir. İlk adımda veri içerisinde birden fazla obetize seviyesi bulunduğu için bu seviyeleri iki grupta birleştirerek 1 ve 0 etiketlerini atamışlardır. Daha sonra özellikler arasındaki korelasyonu belirleyebilmek için özelliklere ait ısı haritasını inceleyerek bazı özellikleri veri içerisinden çıkarmışlardır. Elde edilen son veriyi LR, DVM, KNN, Karar Ağaçları, XGBoost, LightGBM sınıflandırıcılarına vererek algoritma performanslarını değerlendirmişlerdir. Algoritma performanslarını karşılıklı olarak incelemiş ve XGBoost algoritması ile %85.99 oranında doğruluk elde ederek en iyi sonuca ulaştıklarını ifade etmişlerdir [10].

Estren ve arkadaşları farklı ülkelerde bulunan üniversite öğrencilerine yönelik gerçekleştirilen ankete dayalı olarak oluşturulan veri setini kullanarak obezite düzeylerini tahminlemeyi amaçlamışlardır. Veri üzerinde ilk olarak veri temizleme ve dönüştürme işlemlerini uyguladıktan sonra InfoGain, GainRatio, Ki-Kare ve Relief özellik seçim yöntemlerini kullanarak özellik seçme işlemini gerçekleştirmişlerdir. Seçilen özellikleri Rastgele Orman (RO), ÇKA, DVM ve Lojistik Model Ağacı (LMA) algoritmalarına vererek sınıflandırma işlemini gerçekleştirmişlerdir. Sırasıyla RO (%95.62), ÇKA (%94,41), DVM (%83.89) ve LMA (%96,65) hassasiyet oranlarını elde ederek bu problem için en iyi çözümü gösterdiği yüksek performans sayesinde LMA algoritmasının sunduğunu belirlemişlerdir [11].

Quiroz hastaların yaşam tarzına bağlı olarak obezite seviyelerini belirlemek için makine öğrenimi tabanlı bir yaklaşım önermiştir. Çalışmaya Kolombiya, Meksika ve Peru ülkelerinden toplamda 2111 hasta verisini dâhil etmiştir. Obezite seviyelerini belirlemek için Hafif Gradyan Artırma, Rastgele Orman, Karar Ağacı ve LR algoritmalarını kullanmıştır. Sonuçları karşılaştırmalı olarak değerlendirdiğinde %99 ile en yüksek doğruluk değerini Hafif Gradyan Artırma algoritması ile elde edildiğini belirlemiştir [12].

Alqahtani ve arkadaşları UCI makine öğrenimi deposundan aldıkları obezite seviyelerini içeren veri setini kullanarak makine öğrenimi yöntemleri ile obezite seviyelerini tahmin etmeyi amaçlamışlardır. Verileri NObesity kullanarak Aşırı Kilolu I-II, Yetersiz Kilo, Normal Kilo ve Obezite Tip I-III olarak etiketleyerek kategorilere

ayırmışlardır. Makine öğrenimi aşamasında Rastgele Orman ve ÇKA kullanarak sınıflandırma işlemini gerçekleştirmişlerdir. Çalışma sonucunda Rastgele Orman algoritması ile %96.7 oranında doğruluk oranına ulaşarak çalışmanın amacına ulaştığını belirtmişlerdir [13].

Pang ve arkadaşları 2 yaşına kadar olan hastalara ait Elektronik Sağlık Kaydı (ESK) verilerini kullanarak > 2 ile ≤ 7 yaş arasındaki çocukluk obezitesini tahmin etmek için yedi farklı makine öğrenimi algoritması kullanmışlardır. Çalışmaya Philadelphia Çocuk Hastanesi'nden 11.194.579 sağlık hizmeti başvurusu olan 860.510 hastanın ESK verilerini dâhil ederek Karar Ağacı, Gaussian Naive Bayes, Bernoulli Naive Bayes, LR, Yapay Sinir Ağları, DVM ve XGBoost algoritmaları ile tahminleme işlemini gerçekleştirmişlerdir. Algoritmalara ait performans değerlerini karşılaştırmalı olarak inceleyerek en iyi performansı, 0.81 AUC değeri ile ederek sunulan modelin diğer ESK çalışmaları için uyarlanabilir olabileceğini belirtmişlerdir [14].

Bu çalışma, UCI makine öğrenimi deposundan [15] elde edilen yaşları 14 ile 61 arasında değişiklik gösteren toplamda 2111 hastaya ait obezite seviyelerini içeren veri setini kullanarak makine öğrenimi yöntemleri ile obezite seviyesini belirlemeyi amaçlamaktadır. Çalışma kapsamında belirlenen amaca ek olarak optimum parametre ile daha yüksek sınıflandırma başarısı elde etmek için karara etki eden parametrelerin önem seviyeleri, KBA özellik seçme yöntemi kullanılarak araştırılmıştır. Makine öğrenimi aşamasında Karar Ağaçları ve XGBoost algoritmalarını kullanılarak özellik seçme yönteminin performansı, performans değerlendirme metrikleri ile değerlendirilmiştir. Bu çalışmanın ikinci bölümünde kullanılan veri setine ve yöntemlere, üçüncü bölümde bulgular ve tartışmaya, dördüncü bölümde ise sonuçlar başlığına yer verilmiştir.

## 2. Materyal ve Metot

### 2.1. Veri seti

Çalışmaya UCI makine öğrenimi deposundan [15] Meksika, Peru ve Kolombiya ülkelerinde bulunan farklı fiziksel durumlara ve çeşitli yeme alışkanlıklarına sahip 14-61 yaş arası bireylere ait toplamda 16 özellikten oluşan 2111 veri dâhil edilmiştir. Veri kümesi, bir web platformu üzerinden kimliği belli olmayan kullanıcıların yöneltilen soruları yanıtladığı bir anket aracılığı ile toplanmıştır. Anket verileri üzerinde veri dengeleme işlemi gerçekleştirilerek nihai veri seti elde dilmiştir [16]. Elde edilen verilere ait tanımlayıcı bilgiler Tablo 1'de verilmiştir.

**Tablo 1.** Veriye ait tanımlayıcı bilgiler

| Parametre Kategorisi | Parametre | Parametre Açıklaması |
|---|---|---|
| **Fiziksel Özellikler** | Gender | Katılımcı Cinsiyeti |
| | Age | Katılımcı Yaşı |
| | Height | Katılımcı Boyu |
| | Weight | Katılımcı Kilosu |
| | Family_history_overweight | Katılımcı Aile Fazla Kilo Öyküsü |
| **Yeme-İçme ve Zararlı Alışkanlıklar** | FAVC | Yüksek Kalorili Gıda Tüketimi |
| | FCVC | Sebze Tüketim Sıklığı |
| | NCP | Ara Öğün Yeme Sayısı |
| | CAEC | Öğünler Arası Yeme Sıklığı |
| | CH20 | Günlük Su Tüketimi |
| | CALC | Alkol Tüketimi |
| | SMOKE | Sigara Tüketimi |
| **Fiziksel Aktiviteler** | SCC | Kalori Tüketimi |
| | FAF | Fiziksel Aktivite Sıklığı |
| | TUE | Teknolojik Cihaz Kullanma Süresi |
| | MTRANS | Kullanılan Ulaşım Türü |

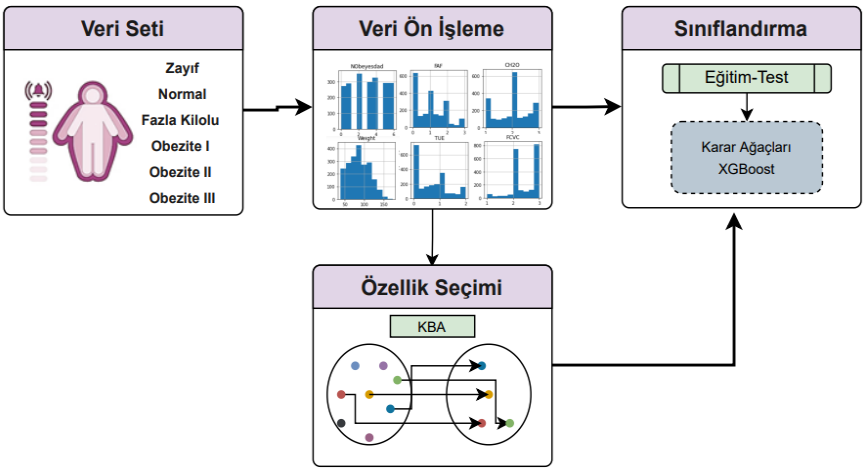
Tablo 1'de bulunan boy ve kilo değişkenlerine ait değerler kullanılarak vücut kitle endeksi hesaplanmış ve veri etiketleme işlemi gerçekleştirilmiştir. Veri etiketleri ve etikete göre vücut kitle endeksleri incelendiğinde;

- Zayıf etiketi için Vücut Kitle Endeksi ≤ 18.5
- Normal etiketi için 18.5 ≤ Vücut Kitle Endeksi ≤ 24.9

- Fazla Kilolu etiketi için 25 ≤ Vücut Kitle Endeksi ≤ 29.9
- Obezite I etiketi için 30 ≤ Vücut Kitle Endeksi ≤ 34.9
- Obezite II etiketi için 35 ≤ Vücut Kitle Endeksi ≤ 39.9
- Obezite III etiketi için Vücut Kitle Endeksi ≥ 40 olarak belirlenmiştir [16].

## 2.2. Metot

Çalışma, makine öğrenimi yöntemlerini kullanarak UCI makine öğrenimi deposundan elde edilen ve anket verilerinden oluşan veri kümesi ile obezite seviyelerini tahmin etmeyi amaçlamıştır. İlk adımda veriye ait çeşitli tanımlayıcı istatistikler incelenerek gerekli ön işlem adımları belirlenmiş ve ön işleme adımından sonra elde edilen kullanılmaya hazır veriler sınıflandırıcılara verilerek algoritma performansları incelenmiştir. Elde edilen anlamlı sonuçlar doğrultusunda sınıflandırma algoritmalarının performansını optimum parametre sayısı ile iyileştirerek hesaplama, zaman ve iş gücü maliyetlerini minimuma indirgemek hedeflenmiştir. Hedef doğrultusunda KBA özellik seçim yöntemi kullanılarak 16 özellikten oluşan veri kümesi 5 özelliğe indirgenerek optimum parametre sayısı belirlenmiştir. Belirlenen yeni parametreler ile sınıflandırma işlemi yeniden gerçekleştirilerek özellik seçim yöntemlerinin performansları değerlendirilmiştir. Şekil 1'de gerçekleştirilen çalışmaya ait süreç tasarımı verilmiştir.




**Şekil 1.** Çalışma süreç tasarımı

### 2.2.1. Veri ön işleme

Çalışma kapsamında UCI makine öğrenimi deposundan [15] elde edilen verileri sınıflandırma aşamasında hazır duruma getirebilmek için veri ön işleme adımları uygulanmıştır. İlk olarak veri içerisinde boş veri olup olmadığı analiz edildiğinde, boş verilerin varlığı gözlemlenmemiştir. Bir sonraki adımda veriye ait betimleyici istatistikler ve dağılım grafikleri incelenerek veri üzerinde uygulanabilecek gerekli işlemler belirlenmiştir. İncelemeler sonucunda weight parametresinin diğer özelliklere kıyasla sapma değerinin yüksek olduğu ve bu duruma paralel olarak normale yakın bir dağılıma sahip olmadığı belirlenmiştir. Mevcut parametre istatistikleri ve dağılımı göz önünde bulundurularak weight parametresi için normal dağılımdan uzak olan verileri, normal dağılım formuna yaklaştırabilmek için kullanılan box-cox dönüşümü uygulanmıştır [17]. Ön işlem adımları tamamlandıktan sonra kullanılmaya hazır hale getirilen veri için sırasıyla "sınıflandırma" ve "özellik seçimi sonrası sınıflandırma" adımları değerlendirilmiştir.

### 2.2.2. Özellik seçimi

Özellik seçimi çok özellikli veri setlerinde, en iyi özelliklerin seçilerek optimum parametre sayısının elde edilmesi olarak tanımlanabilir. Özellik seçiminin temel amacı model performansını artırmak, zaman ve iş maliyetlerinden tasarruf sağlamaktır [18]. Özellik seçimi içerisinde birden çok yöntem barındırmaktadır. Bu

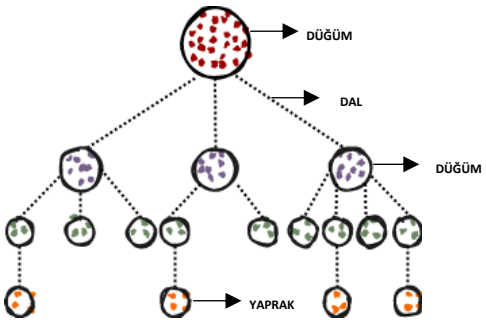
yöntemlerden biri makine öğrenimi uygulamalarında sıklıkla karşılaşılan denetimli bir mekanizmaya sahip olan KBA yöntemidir [19]. KBA özellik seçim yönteminde özellik seçimi gerçekleştirilirken veri içerisinden herhangi bir bilgi kaybı gerçekleşmez. KBA yöntemi özellik etkisini belirleyebilmek için mesafe ölçüm metriklerini ele alır. Bu yöntem ile pozitif ağırlıklar elde edilirken ağırlıklar özelliklerin belirlenmesindeki en önemli unsurdur. Daha düşük ağırlıklar daha az önemli özellikleri temsil ederken, yüksek ağırlıklar önem seviyesi yüksek özellikleri temsil etmektedir [20]. Bu çalışmada KBA özellik seçim yöntemi kullanılarak her özellik için bir ağırlık değeri atanmış ve sonrasında bu ağırlıklara paralel olarak özellikler önem seviyesine göre sıralanmıştır. Sonraki adımda sınıflandırma algoritmalarına minimum özellik sayısı ile başlanarak verilen özelliklerin en iyi performansı 5 özellik ile sağladığı gözlemlenerek optimum parametre sayısı 5 olarak belirlenmiştir.

### 2.2.3. Sınıflandırma

Sınıflandırma makine öğreniminin temel alt dallarından biridir. Sınıflandırma işlemini gerçekleştirebilmek için birçok algoritma bulunmaktadır. Bu çalışmada XGBoost ve Karar Ağacı algoritmaları kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. İlk olarak veriler %70 eğitim ve %30 test verisi olarak kullanılmak için ayrıştırıldıktan sonra, eğitim verileri ile model eğitimi sağlanmıştır. Elde edilen anlamlı sonuç doğrultusunda test verileri ile model performansı çeşitli metrikler kullanılarak değerlendirilmiştir.

XGBoost, denetimli öğrenme modelleri için yüksek performans sağlayan algoritmalardan biridir. Sınıflandırma ve regresyon problemleri için uyarlanabilir olan bu yöntem sağladığı yüksek çalışma hızı ve kontrollü mekanizması sayesinde sıklıkla tercih edilmektedir [21]. XGBoost yöntemi, aşırı uyum sorununun önüne geçerek model doğruluğunu artırmaya yönelik tasarlanmış bir algoritmadır. Uygun parametre ayarı XGBoost sınıflandırıcısının performansını etkileyen en önemli etmenlerden biridir [22]. Bu çalışmada her algoritma için uygun parametreler otomatikleştirilmiş hiper parametre optimizasyon yöntemlerinden biri olan OPTUNA yöntemi kullanılarak belirlenmiştir. XGBoost algoritması için belirlenen hiper parametreler öğrenme katsayısı 0.38, maksimum derinlik 72, ağaç sayısı 45, alt örneklem 1 olarak belirlenmiştir.

Karar ağaçları, makine öğrenimi uygulamalarında sıklıkla başvurulan ağaç tabanlı ardışık bir model yapısına sahip olan algoritmadır. Karar ağaçları ana düğümler, dallar ve yapraklardan oluşmaktadır. Düğümler karar ağaçlarında özelliklerin temsilini sağlarken, dallar sorulara ait cevapları, yapraklar ise sınıf etiketlerini temsil etmektedir [23]. Anlaşılabilir ve basit yapısı sebebiyle karar ağaçları farklı uygulama alanlarında sıklıkla tercih edilmektedir. Çalışma kapsamında karar ağacı hiper parametreleri sırasıyla; maksimum derinlik 7, öğrenme katsayısı 0.5 ve minimum örnek bölünme sayısı 3 olarak belirlenmiştir. Şekil 2'de karar ağacına ait örnek bir yapı verilmiştir.




**Şekil 2.** Karar ağacı örnek yapı

Çalışmada sınıflandırma algoritmalarının performanslarını değerlendirmek için doğruluk, duyarlılık, kesinlik ve F1-skor metrikleri kullanılmıştır. Performans değerlendirme metrikleri hesaplanırken karmaşıklık matrisi baz alınmıştır. Karmaşıklık matrisine göre;

Doğru Pozitif (DP): Obezite olup Obezite olarak tanımlanmış hasta sayısı.

Yanlış Pozitif (YP): Obezite olup Obezite değil olarak tanımlanmış hasta sayısı.

Doğru Negatif (DN): Obezite değil olup Obezite değil olarak tanımlanmış hasta sayısı.

Yanlış Negatif (YN): Obezite değil olup Obezite olarak tanımlanmış hasta sayısını ifade etmektedir.

Yukarıda verilen tanımlamalara göre performans değerlendirme metriklerine ait hesaplama formülleri aşağıda verilmiştir [17].

$$Doğruluk = (DN + DP)/(DN + DP + YP + YN) \quad (1)$$

$$Duyarlılık = DP/(DP + YN) \quad (2)$$

$$Kesinlik = DP/(DP + YP) \quad (3)$$

$$F1 - Skor = (2 \times Duyarlılık \times Kesinlik)/(Duyarlılık + Kesinlik) \quad (4)$$

## 3. Bulgular ve Tartışma

Çalışma, UCI veri deposundan elde edilen ve açık kaynak olarak erişim sağlanabilen obezite hastalarına ait verileri kullanarak makine öğrenimi yöntemleri ile optimum özellik kullanarak obezite seviyelerini yüksek doğruluk ile tahmin etme yaklaşımına dayanarak gerçekleştirilmiştir. Toplamda 16 özellikten oluşan veri seti çalışmaya dâhil edilerek özellik seçim işlemi ile özellik sayısı 5'e indirgenmiştir. Özellik seçim işlemi ve sonrası Karar Ağacı ve XGBoost algoritmalarının performansları karşılıklı olarak incelenerek doğruluk, duyarlılık, kesinlik ve F1-Skor metrikleri ile değerlendirilmiştir. Çalışmaya ait sınıflandırıcıların özellik seçim işleminden önceki performans metrikleri Tablo 2 ve Tablo 3'te verilmiştir.

**Tablo 2.** Özellik seçim işlemi öncesi sınıflandırma performansları

| | **Kesinlik (%)** | | **Duyarlılık (%)** | | **F1-Skor (%)** | |
|---|---|---|---|---|---|---|
| **Sınıflar** | **XGBoost** | **Karar Ağacı** | **XGBoost** | **Karar Ağacı** | **XGBoost** | **Karar Ağacı** |
| **Insufficient Weight** | 93 | 95 | 97 | 95 | 95 | 95 |
| **Normal Weight** | 96 | 86 | 91 | 74 | 93 | 80 |
| **Obesity Type_I** | 97 | 90 | 95 | 91 | 96 | 90 |
| **Obesity Type_II** | 98 | 97 | 99 | 97 | 98 | 97 |
| **Obesity Type_III** | 100 | 100 | 99 | 100 | 99 | 100 |
| **Overweight Level_I** | 92 | 69 | 97 | 86 | 94 | 77 |
| **Overweight Level_II** | 97 | 88 | 96 | 81 | 97 | 84 |

**Tablo 3.** Özellik seçim işlemi öncesi ortalama sınıflandırma performansları

| **Algoritmalar** | **Ortalama Kesinlik (%)** | **Ortalama Duyarlılık (%)** | **Ortalama F1-Skor (%)** | **Ortalama Doğruluk (%)** |
|---|---|---|---|---|
| **XGBoost** | 96 | 96 | 96 | 96 |
| **Karar Ağacı** | 90 | 89 | 89 | 89 |

Tablo 2 incelendiğinde elde edilen anlamlı sonuçlar doğrultusunda çalışmanın ilk adımı olan sınıflandırma işleminin başarılı bir şekilde gerçekleştiği görülmektedir. Her obezite seviye sınıfı için ayrı değerlendirilme gerçekleştirildiğinde en iyi tahmin edilme oranına sahip olan sınıfın Obesity_Type_III sınıfı olduğu, diğer sınıflara kıyasla daha düşük oranla tahmin edilen sınıfın ise Overweight_Level_I sınıfı olduğu belirlenmiştir. Veri seti incelendiğinde bu farkın sınıf etiketlerinin veride bulunma sayısından kaynaklandığı belirlenmiştir. Tablo 3'te verilen ortalama doğruluk metriği incelendiğinde XGBoost algoritmasının %96 oran ile %89 doğruluk oranına sahip Karar Ağacı algoritmasından daha yüksek performans sergilediği gözlemlenmiştir. Sınıflandırma işleminden sonra KBA ile özellik seçimi gerçekleştirilerek çalışmaya pozitif performans yönünden en çok etki eden özellikler, sırasıyla önem seviyelerine göre "Height", "Weight", "Gender", "Age" ve "CALC" olarak belirlenmiştir. KBA yöntemi ile seçilen 5 özellik tekrar sınıflandırma işlemine tabi tutularak algoritma performansları karşılıklı olarak incelenmiştir. Özellik seçim işlemi sonrası algoritma performansları Tablo 4 ve Tablo 5'te verilmiştir.

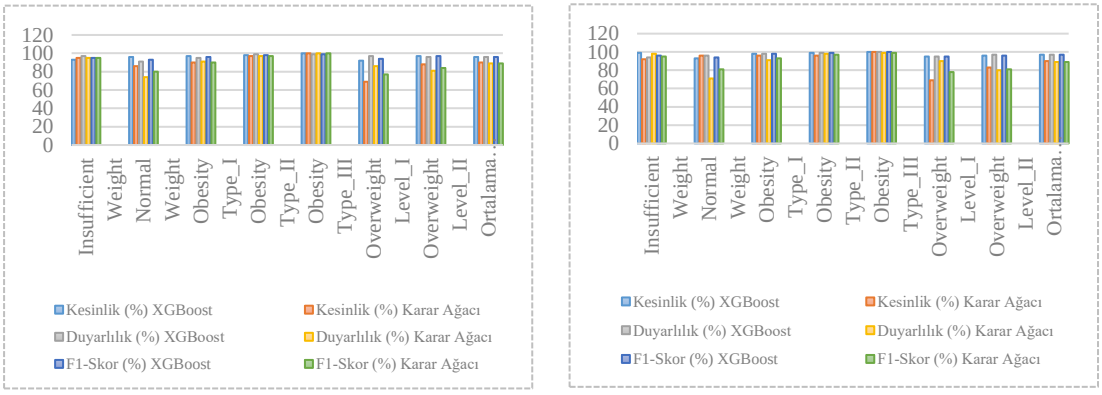
**Tablo 4.** Özellik seçim işlemi sonrası sınıflandırma performansları

| | Kesinlik (%) | | Duyarlılık (%) | | F1-Skor (%) | |
|---|---|---|---|---|---|---|
| **Sınıflar** | **XGBoost** | **Karar Ağacı** | **XGBoost** | **Karar Ağacı** | **XGBoost** | **Karar Ağacı** |
| **Insufficient Weight** | 99 | 92 | 94 | 98 | 96 | 95 |
| **Normal Weight** | 93 | 96 | 96 | 71 | 94 | 81 |
| **Obesity Type_I** | 98 | 96 | 98 | 91 | 98 | 93 |
| **Obesity Type_II** | 99 | 96 | 99 | 98 | 99 | 97 |
| **Obesity Type_III** | 100 | 100 | 100 | 99 | 100 | 99 |
| **Overweight Level_I** | 95 | 69 | 95 | 90 | 95 | 78 |
| **Overweight Level_II** | 96 | 83 | 97 | 80 | 96 | 81 |

**Tablo 5.** Özellik seçim işlemi öncesi ortalama sınıflandırma performansları

| **Algoritmalar** | **Ortalama Kesinlik (%)** | **Ortalama Duyarlılık (%)** | **Ortalama F1-Skor (%)** | **Ortalama Doğruluk (%)** |
|---|---|---|---|---|
| **XGBoost** | 97 | 97 | 97 | 98 |
| **Karar Ağacı** | 90 | 89 | 89 | 90 |

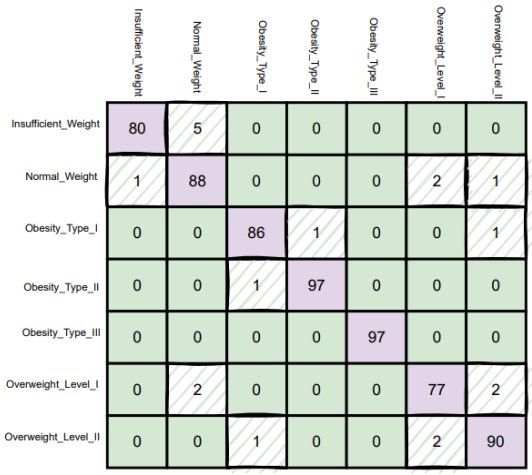
Tablo 4 incelendiğinde özellik seçim işlemi sonrası performans metriklerinin oranlarında artış olduğu görülmektedir. Her iki algoritma için Tablo 5'te verilen ortalama doğruluk oranlarına bakıldığında XGBoost algoritması için %98, Karar Ağacı algoritması için %90 değerlerine ulaşıldığı belirlenmiştir. Özellik seçim işlemi öncesi bu oranlar her iki algoritma içinde minimum %1 oranında daha düşük olmakla beraber optimum özellik sayısı ile daha yüksek doğruluk elde edilebileceği kanıtlanmıştır. Özellik seçim işlemi öncesi ve sonrası algoritma performanslarına ait grafikler anlaşılabilirliği artırmak amacıyla Şekil 3'te verilmiştir.




(a) (b)

**Şekil 3.** (a) özellik seçimi öncesi algoritma performansları (b) özellik seçimi sonrası algoritma performansları

Şekil 3 ve tablolar birlikte incelendiğinde deneysel sonuçlar sonunda oluşan performans değerlerinin çalışma amacını destekler nitelikte olduğu görülmektedir. Özellik seçim işlemi sonrası en yüksek performansı gösteren XGBoost algoritması için elde edilen karmaşıklık matrisi Şekil 4'te verilmiştir.




**Şekil 4.** XGBoost Algoritması için Karmaşıklık Matrisi

Önerilen çalışmada, UCI açık erişim makine öğrenimi deposundan elde edilen veriler kullanılarak özellik seçimi tabanlı makine öğrenimi modelleri ile doğru bir şekilde obezite seviyelerini sınıflandırmak amaçlanmıştır. Kullanılan sınıflandırma algoritmaları özellik seçim işleminden önce ve sonra olmak üzere performans metrikleri ile değerlendirilmiştir. Sonuçlar incelendiğinde optimum özellik sayısı ile daha yüksek performans hedefine ulaşıldığı görülmüştür. Uzmanlara karar verme adımında destek olacak bu sistemin daha az özellik kullanımı sayesinde iş gücü, çalışma ve hesaplama maliyeti açısından tasarruf sağlayacağı öngörülmektedir. Çalışma giriş bölümünde verilen literatür özetinde, gerçekleştirilen obezite çalışmaları ile karşılıklı olarak incelendiğinde bu alana yönelik yapay zekâ tabanlı birden fazla çalışma bulunmaktadır. Bu çalışmalardan bazıları Tablo 6'da verilmiştir.

**Tablo 6.** Literatürde Gerçekleştirilen Çalışmalar

| Referans | Yıl | Veri Seti | Algoritmalar | Özellik Seçimi | En İyi Sonuçlar (Doğruluk) |
|---|---|---|---|---|---|
| Cui ve diğerleri [7] | 2021 | UCI Makine Öğrenimi Deposu-2111 Obezite Verisi | LR, XGBoost, KNN, Karar Ağaçları, LightGBM | Isı Haritası | XGBoost- %85.99 |
| Quiroz ve diğerleri [9] | 2022 | UCI Makine Öğrenimi Deposu-2111 Obezite Verisi | Hafif Gradyan Artırma, Rastgele Orman, Karar Ağacı, LR | - | Hafif Gradyan Artırma- %99 |
| Alqahtani ve diğerleri [10] | 2021 | UCI Makine Öğrenimi Deposu-2111 Obezite Verisi | Rastgele Orman, Çok Katmanlı Algılayıcılar | - | Rastgele Orman- %96.7 |
| Çelik ve diğerleri [24] | 2021 | UCI Makine Öğrenimi Deposu-2111 Obezite Verisi | Karar Ağacı, Yapay Sinir Ağları, DVM | Geriye Dönük Özellik Eleme | Cubic SVM- %97.8 |
| Zheng ve diğerleri [25] | 2017 | Gençlik Riskli Davranış İzleme Sistemi Açık Erişim Veri Seti | Karar Ağacı, Yapay Sinir Ağları, KNN | - | KNN- %88.82 |
| **Önerilen Çalışma** | 2022 | UCI Makine Öğrenimi Deposu-2111 Obezite Verisi | XGBoost, Karar Ağacı | KBA | Tüm Veri- XGBoost- %96<br>Seçilen Özellikler- XGBoost- %98 |

Tablo 6'da verilen çalışmalara göre özellik seçimi gerçekleştirilerek yapılan çalışmaların doğruluk oranlarının yüksek seviyede olduğu görülmektedir. Fakat gerçekleştirilen mevcut çalışma ile karşılaştırmalı olarak incelendiğinde çalışmada kullanılan özellik seçim yönteminin daha önce aynı veri seti üzerinde hiç uygulanmadığı belirlenmiştir. Kullanılan özellik seçim yöntemi ile elde edilen verilerin algoritma performanslarına olumlu yönde

etki ettiği ve doğruluk oranının diğer çalışmalara kıyasla daha yüksek olması gerçekleştirilen çalışmayı değerli kılmaktadır.

## 4. Sonuçlar

Obezite çocuk, genç ve erişkin bireylerde günümüzde en sık rastlanan sağlık sorunlarından biridir. Obezite bir sağlık sorunu olmakla beraber, yol açtığı diyabet, kalp rahatsızlıkları vb. hastalıklarda bulunmaktadır. Bu aşamada obezitenin erken seviyede önlenmesi yaşam kalitesinin artması yönünde olumlu etki yaratmaktadır. Obezite her birey için farklı seviyelerde seyredilebilen bir hastalık olması sebebiyle obezite seviyelerinin belirlenerek kişilere özel tedavilerin uygulanması önem arz etmektedir.

Bu çalışmada obezite hastalarına ait açık erişim verileri kullanarak obezite seviyelerini özellik seçimi tabanlı makine öğrenimi algoritmalarıyla sınıflandırıp optimum özellik sayısı ile yüksek performans elde etmek hedeflenmiştir. Hedefe yönelik olarak özellik seçim aşamasında KBA yöntemi kullanılarak algoritma performansına en çok etki eden beş parametre "Height", "Weight", "Gender", "Age" ve "CALC" olarak belirlenmiştir. Elde edilen özellikler sınıflandırma algoritmalarına verilerek özellik seçim işlemi öncesi XGBoost algoritması için %96, Karar Ağacı algoritması için %89 doğruluk değerlerine ulaşılırken, özellik seçim işlemi sonrası XGBoost algoritması için %98, Karar Ağacı algoritması için %90 doğruluk değerleri elde edilerek başlangıç hedefine doğru bir şekilde ulaşıldığı gözlemlenmiştir. Literatürde obezite seviyelerinin özellik seçim tabanlı makine öğrenimi modelleri ile sınıflandırılmasına yönelik çalışmaların az sayıda olması çalışmayı özgün kılmakla beraber optimum sayıda özellik ile hastalık tahmini için gerçekleştirilebilecek diğer çalışmalara ışık tutması beklenmektedir.

## Kaynaklar


[1] Sipahi, B. B. (2021). Türkiye'de obezite üzerine sosyoekonomik faktörlerin etkisi ve gelir eşitsizliği. Ankara Üniversitesi SBF Dergisi, 76(2), 547-573.

[2] Parmaksız, H. (2007). Yetişkin obezlerde fiziksel aktivite seviyesinin belirlenmesi (Doctoral dissertation, DEÜ Sağlık Bilimleri Enstitüsü).

[3] Quiroz, J. P. S. (2022). Estimation of obesity levels based on dietary habits and condition physical using computational intelligence. Informatics in Medicine Unlocked, 29, 100901.

[4] PİRİM, A. G. H. (2006). Yapay zekâ. Yaşar Üniversitesi E-Dergisi, 1(1), 81-93.

[5] Arıkan, M., Yapay zeka nedir? Yapay Zekâ Uygulama Alanları Nelerdir?. https://www.mediaclick.com.tr/tr/blog/yapay-zeka-nedir. Erişim Tarihi: 25.12.2022

[6] Büyükgöze, S., & Dereli, E. (2019). Dijital sağlık uygulamalarında yapay zekâ. VI. Uluslararası Bilimsel ve Mesleki Çalışmalar Kongresi-Fen ve Sağlık, 7(10).

[7] İŞLER, B., & KILIÇ, M. (2021). EĞİTİMDE YAPAY ZEKÂ KULLANIMI VE GELİŞİMİ. Yeni Medya Elektronik Dergisi, 5(1), 1-11.

[8] Cervantes, R. C., & Palacio, U. M. (2020). Estimation of obesity levels based on computational intelligence. Informatics in Medicine Unlocked, 21, 100472.

[9] Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., & Habib, M. T. (2021). A machine learning approach for obesity risk prediction. Current Research in Behavioral Sciences, 2, 100053.

[10] Cui, T., Chen, Y., Wang, J., Deng, H., & Huang, Y. (2021, May). Estimation of Obesity Levels Based on Decision Trees. In 2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM) (pp. 160-165). IEEE.

[11] Molina Estren, D., De la Hoz Manotas, A. K., & Mendoza Palechor, F. (2021). Classification and features selection method for obesity level prediction.

[12] Quiroz, J. P. S. (2022). Estimation of obesity levels based on dietary habits and condition physical using computational intelligence. Informatics in Medicine Unlocked, 29, 100901.

[13] Alqahtani, A., Albuainin, F., Alrayes, R., muhanna, N. A., Alyahyan, E., & Aldahasi, E. (2021). Obesity Level Prediction Based on Data Mining Techniques. International Journal of Computer Science and Network Security, 21(3), 103–111. doi: https://doi.org/10.22937/IJCSNS.2021.21.3.14

[14] Pang, X., Forrest, C. B., Lê-Scherban, F., & Masino, A. J. (2021). Prediction of early childhood obesity with machine learning and electronic health record data. International journal of medical informatics, 150, 104454.

[15] https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition. . Erişim Tarihi: 25.12.2022

[16] Palechor, F. M., & de la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in brief, 25, 104344.

[17] Danacı, Ç. (2022). Covid-19 Tanısında Biyokimya Parametre Baskınlığının Makine Öğrenimi Yöntemleri Kullanılarak Belirlenmesi, Yüksek Lisans Tezi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü

[18] Budak, H. (2018). Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22(Özel), 21. doi: https://doi.org/10.19113/sdufbed.01653

[19] Koc, M., Sut, S. K., Serhatlioglu, I., Baygin, M., & Tuncer, T. (2022). Automatic prostate cancer detection model based on ensemble VGGNet feature generation and NCA feature selection using magnetic resonance images. Multimedia Tools and Applications, 81(5), 7125-7144.

[20] Tuncer, T., Dogan, S., Pławiak, P., & Acharya, U. R. (2019). Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ECG signals. Knowledge-Based Systems, 186, 104923.

[21] Hayri, A. B. A. R. (2020). Xgboost Ve Mars Yöntemleriyle Altin Fiyatlarinin Kestirimi. Ekev Akademi Dergisi, (83), 427-446.

[22] Osman, A. I. A., Ahmed, A. N., Chow, M. F., Huang, Y. F., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, 12(2), 1545-1556.

[23] Bavaş, E., Karar Ağaçları (Decision Trees) ile Veri Sınıflandırma. http://erdoganb.com/2017/07/karar-agaclari-decision-trees-ile-veri-siniflandirma/ Erişim Tarihi: 25.12.2022

[24] Y. Celik, S. Guney and B. Dengiz, "Obesity Level Estimation based on Machine Learning Methods and Artificial Neural Networks," 2021 44th International Conference on Telecommunications and Signal Processing (TSP), 2021, pp. 329-332, doi: 10.1109/TSP52935.2021.9522628.

[25] Z. Zheng ve K. Ruggiero, "Lise öğrencilerinde obeziteyi tahmin etmek için makine öğrenimini kullanmak", 2017 IEEE Uluslararası Biyoinformatik ve Biyotıp Konferansı (BIBM) , 2017, s. 2132-2138, doi: 10.1109/BIBM.2017.8217988.