MAKÜFEBED

ISSN Online: 1309-2243
http://dergipark.org.tr/makufebed
https://doi.org/10.29048/makufebed.1372323

Mehmet Akif Ersoy Üniversitesi Fen Bilimleri Enstitüsü Dergisi 14(2): 301-312 (2023)
The Journal of Graduate School of Natural and Applied Sciences of Mehmet Akif Ersoy University 14(2): 301-312 (2023)

Araştırma Makalesi / *Research Paper*

Optimize Edilmiş Denetimli Öğrenme Algoritmaları ile Obezite Analizi ve Tahmini

Tülay TURAN

Burdur Mehmet Akif Ersoy Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Burdur

Geliş Tarihi (Received): 06.10.2023, Kabul Tarihi (Accepted): 07.11.2023
Sorumlu Yazar (Corresponding author): tulayturan@mehmetakif.edu.tr*
+90 248 2134552 *+90 248 2134598*

ÖZ

Obezite dünya genelinde gerçekleşen ölümlerin en önemli beşinci nedeni olarak karşımız çıkan bir sağlık sorunudur. Dünya Sağlık Örgütü (DSÖ) 2022 yılında yayınladığı raporda obezitenin birçok hastalığın temelini oluşturduğunu ve gerekli önlemler ve politikalar izlenerek durdurulabileceğini vurgulamıştır. Bu nedenle makine öğrenmesi algoritmaları ile obezite analizi ve tahmin uygulamaları önemlidir. Bu çalışmada, UCI makine öğrenmesi veri havuzundan alınan veriler kullanılarak, denetimli öğrenme algoritmalarından K-En Yakın Komşu (KNN) algoritması ve Rastgele Orman (RF) algoritması ile tahmin modelleri geliştirilmiştir. Bu modeller farklı istatistiksel değerlendirme kriterleri kullanılarak karşılaştırılmıştır. Değerlendirme sonucunda hiper parametre optimizasyonu gerçekleştirilen RF modeli %94 ortalama doğruluk (accuracy) sonucu ile en iyi tahmin sonucunu elde etmiştir. Çalışma obezite prevalansını etkileyen faktörleri analiz etmesi, görselleştirmesi ve yüksek bir başarı oranı ile seviyelerini tahmin etmesiyle önemlidir.

Anahtar Kelimeler: Obezite Analizi, Denetimli Makine Öğrenmesi, Hiper parametre Optimizasyonu

Obesity Analysis and Prediction with Optimized Supervised Learning Algorithms

ABSTRACT

Obesity is a health problem that is the fifth most important cause of death worldwide. In the report published in 2022, the World Health Organization (WHO) emphasized that obesity forms the basis of many diseases and can be stopped by following the necessary measures and policies. Therefore, obesity analysis and prediction applications with machine learning algorithms are important. In this study, prediction models were developed with K-Nearest Neighbor (KNN) algorithm and Random Forest (RF) algorithm, which are supervised learning algorithms, using data from the UCI machine learning data repository. These models were compared using different statistical evaluation criteria. As a result of the evaluation, the RF model with hyperparameter optimization achieved the best prediction result with an average accuracy of 94%. The study is important because it analyzes and visualizes the factors affecting the prevalence of obesity and predicts its levels with a high success rate.

Keywords: Obesity Analysis, Supervised Machine Learning, Hyperparameter Optimization

GİRİŞ

Obezite, vücut yağının fazlalığı olarak tanımlanan, dünya çapında ciddi bir halk sağlığı sorunudur. Çok sayıda çalışma, obezitenin bireylerin genetik yapılarından, toplumsal, kültürel sağlıksız beslenme alışkanlıklarından kaynaklanan, karmaşık bir sağlık sorunu olduğunu göstermiştir (Caballero, 2005; Ogden ve ark., 2006). Dünya Sağlık Örgütü (DSÖ) tarafından ölüm ve sakatlığın önemli bir belirleyicisi olarak tanımlanan obezite; yüksek tansiyon, beslenme riskleri ve tütünden sonra bulaşıcı olmayan hastalıklar açısından dördüncü en yaygın risk faktörü olarak dikkat çekmektedir (Koo ve ark., 2023; Vizmanos ve ark., 2023). En son veriler aşırı kilo ve obezitenin dünya çapında her yıl 1,3 milyondan fazla ölüme yol açtığını göstermektedir (Bansal ve ark., 2023; Brero ve ark., 2023; Wanjau ve ark., 2023).

DSÖ tarafından paylaşılan Avrupa Bölgesel Obezite Raporu'nda; yetişkinlerin neredeyse üçte ikisinin aşırı kilolu veya obez olduğu ve bu düzeylerin artmaya devam ettiği belirtilmiştir Raporda obezitenin kardiyovasküler hastalıklar (CVD), 13 kanser türü, tip 2 diyabet (T2DM), obstrüktif uyku apnesi (OSA) ve kronik solunum hastalıkları dahil olmak üzere birçok bulaşıcı olmayan hastalık (BOH) için artan riskle bağlantılı olduğuna dikkat çekilmiştir. Rapor, önümüzdeki yıllarda bölgedeki bazı ülkelerde obezitenin kanser açısından ana risk faktörü olan sigarayı geride bırakacağını öngörmektedir. Raporda ayrıca obezitenin sadece bir risk faktörü değil, özel olarak tedavi edilmesi ve yönetilmesi gereken bir durum olduğunun da altı çizilmiştir (WHO,2022).

DSÖ 2030 yılında dünyadaki ölümlerin %30'unun yaşam tarzı hastalıklarıyla başlayacağını öngörmektedir. Bu sürecin ilgili risk faktörlerinin uygun şekilde tanımlanması, ele alınması ve davranışsal katılım politikalarıyla durdurulabileceğini öngörmektedir (WHO,2022). Bu nedenle obezitenin mümkün olduğu kadar erken tespit edilmesi çok önemlidir. Yapay zekanın bir alt kümesi olan makine öğrenmesi ile obezite teşhisi, obezite seviyelerinin sınıflandırılması, obeziteye neden olan risk faktörleri arasındaki bağlantıların tanımlanması, bu bağlamda dikkat çeken önemli bir konu başlığıdır.

Literatür taramalarında online paylaşılan obezite verileri kullanılarak, çeşitli makine öğrenmesi uygulamalarının geliştirildiği görülmüştür. Cui ve arkadaşları (2021) UCI Makine Öğrenimi Havuzunda bulunan bir veri kümesini kullanarak obezite seviyelerini sınıflandırmışlardır. Çalışmalarında Logistic Regression, SVM (Support Vektor Machine), KNN (K Nearest Neighbor), DT (Decision Tree) ve RF (Random Forest) makine öğrenme yöntemlerini kullanmışlardır. Sonuç olarak RF modelinin %85,58 ile en iyi sonucu elde ettiğini belirtmişlerdir. Clem's ve arkadaşları (2022) çalışmalarında, keşifsel veri analizi ve veri ön işleme sonrasında k-katlı çapraz doğrulama ile modellerinin eğittiklerini. Eğitim sonucunda SVM modelinin en iyi doğruluk sonucunu elde ettiğini belirtmişlerdir. Jeon ve arkadaşları (2023) yetişkinler (19-79 yaş) için obezitenin yaşa ve cinsiyete özgü risk faktörlerini araştırmışlardır. Çalışma bulgularında, farklı makine öğrenimi algoritmaları altında obezite risk faktörlerinin yaşa ve cinsiyete duyarlı olduğunu gösterdiğini belirtmişlerdir. Çalışmalarının sonucunda 19-39 yaş grubu için %70'in üzerinde doğruluk oranı, 60-79 yaş grubu için yaklaşık %65 doğruluk oranı elde etmişlerdir. Yaygın ve arkadaşları (2023) eğitimli sinir ağı modeli kullanarak obezite düzeyini tahmin etmeyi amaçlamışlardır. Obeziteyle ilişkili en kritik faktörleri belirlemek için ki-kare, F-Classify ve karşılıklı bilgi sınıflandırma algoritmaları kullandıklarını belirtmişlerdir. Roy ve Protity (2023) bir bireyin sağlıklı olması için ihtiyaç duyduğu besin miktarını tahmin etmek amacıyla makine öğrenimine dayalı yeni bir sistem önermişlerdir. Geliştirdikleri LightGBM modeli ile düşük kök ortalama kare hatası (RMSE) ve %79,27 doğruluk değeri elde ettikleri görülmüştür. Mondal ve arkadaşları (2023) çalışmalarında BMI ölçümleri ve cinsiyet gibi temel bilgileri kullanarak obeziteyi tahmin etmektedir. Modellerinin, bir çocuğun beş yaşındaki obezite kategorisini (normal, fazla kilolu veya obez) üç uygulama senaryosu için sırasıyla %89, %77 ve %89 doğrulukla tahmin edebildiğini belirtmişlerdir. Danacı ve arkadaşları (2023) Komşuluk Bileşen Analizi (KBA) yöntemi ile özellik seçimi yapılan veriler üzerinde XGBoost ve DT algoritmaları ile obezite düzeylerini sınıflandırmışlardır. Çalışmalarında özellik seçimi sonrasında her iki model sonucunda %1 artış gözlemlediklerini belirtmişlerdir. Dugan ve arkadaşları (2015) CHICA veri setini kullanarak, iki yaşından sonra çocuklukta obeziteyi tahmin etmeyi amaçlamışlardır. Çalışmalarında eğittikleri ID3 modelinin %85 doğruluk ve %89 hassasiyetle en iyi genel performansı el ettiğini belirtmişlerdir. Ferdowsy ve arkadaşları (2021) çalışmalarında yüksek, orta ve düşük obezite seviyelerini sınıflandırmışlardır. Lojistik Regresyon Algoritmasının diğer sınıflandırıcılara göre %97,09 ile en yüksek doğruluğa ulaştığını belirtmişlerdir.

Literatürde gerçekleştirilen çalışmaların genel olarak beden kitle indeksinin hesaplanmasıyla sınırlı kaldığı, aile obezite geçmişi, fiziksel aktiviteler, beslenme geçmişi gibi obezite ile ilişkili önemli faktörlerin göz ardı edildiği görülmüştür. Bu bağlamda çalışmada obeziteyi etkileyen faktörler dikkate alınarak tahmin modelleri geliştirilmiştir. Çalışma ile

- Çağımızın en dikkat çeken hasatlıklarından biri olan obezitenin önlenmesi için gerçekleştirilen yapay zeka çalışmalarına katkı sağlaması,
- Obezite üzerinde en önemli etkiye sahip prediktif faktörlerin grafikler ile sunulması,
- Obezite seviyesinin otomatik tahmini için makine öğrenmesi modellerinin geliştirilmesi ve hiper parametre optimizasyonu ile en iyi sonuçların elde edilmesi,
- Makine öğrenimine dayalı bu tahmin sayesinde bireylerin sağlıkları konusunda daha dikkatli ve bilinçli olması amaçlanmaktadır.

MATERYAL VE YÖNTEM

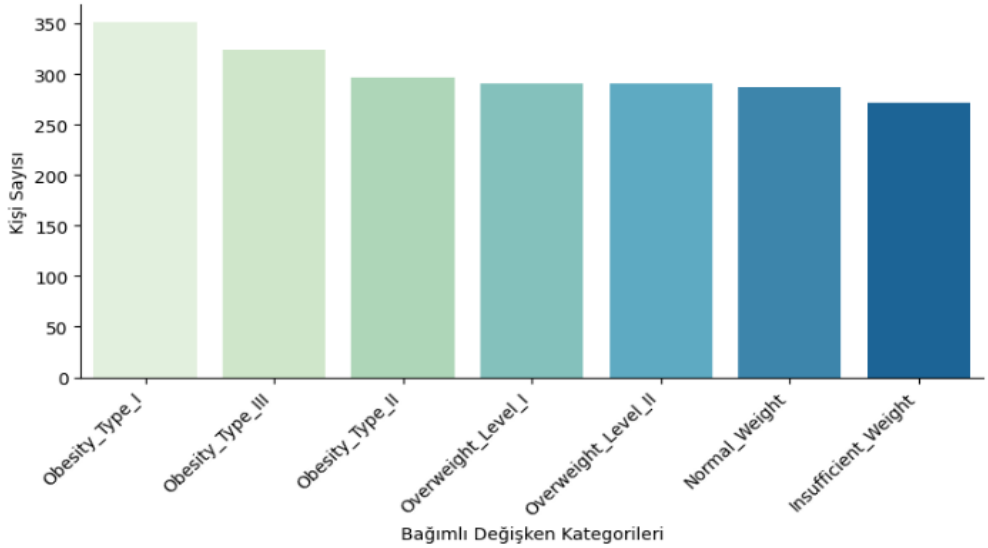
Veri Seti

Çalışmada kullanılan veri seti UCI makine öğrenmesi veri havuzundan elde edilmiştir. Veri seti Meksika, Peru ve Kolombiya ülkelerindeki bireylerin yeme alışkanlıklarına ve fiziksel durumlarına göre obezite düzeylerinin tahmin edilmesine yönelik bilgileri içermektedir. Veriler 17 nitelik ve 2111 kayıttan oluşmaktadır (UCI, 2023). Tablo 1'de veri seti içerisinde yer alan ve çalışmada bağımsız değişken olarak kullanılan alanların isimleri ve açıklamaları gösterilmektedir.

Tablo 1. Veri seti bağımsız değişken bilgileri ve açıklamaları

Bağımsız Değişkenler	Değişken Açıklaması	Özellik Kümesi
Cinsiyet	Cinsiyet türünün belirtildiği kategorik değer	Kişisel Özellikler
Yaş	Bireylerin yaşını gösteren sayısal değer	
Boy	Bireylerin boyunu metre cinsinden gösteren sayısal değer	
Kilo	Bireylerin kilosunu kilogram cinsinden gösteren sayısal değer	
Aile obezite geçmişi	Aile obezite geçmişinin belirtildiği kategorik değer	
Yüksek kalorili yiyeceklerin sık tüketilmesi (FAVC)	Yüksek kalorili yiyeceklerin sık tüketilmesinin belirtildiği kategorik değer	Beslenme Alışkanlıkları
Sebze tüketim sıklığı (FCVC)	Sebze tüketim sıklığı (1 = hiçbir zaman, 2 = bazen, 3 = her zaman)	
Ana öğün sayısı (NCP)	Ana öğün sayısı (1 = 1 ile 2 arası, 2 = üç, 3 = üçten fazla, 4 = cevap yok)	
Öğün aralarında besin tüketimi (CAEC)	Öğün aralarında besin tüketimi (1 = hayır, 2 = bazen, 3 = sık sık, 4 = her zaman)	
Sigara (SMOKE)	Sigara içip/ içmediğinin belirtildiği kategorik değişken)	
Günlük su tüketimi (CH2O)	Günlük su tüketimi miktarı (1 = bir litreden az, 2 = 1 ile 2 litre arası, 3 = 2 litreden fazla)	
Alkol tüketimi (CALC)	Alkol tüketim miktarı (1= hayır, 2 = bazen, 3 = sık sık, 4 = her zaman)	
Kalori tüketimi izleme (SCC)	Bireylerin kalori alıp almadığını izlemesini kontrol eden kategorik değer	Fiziksel Aktiviteler
Fiziksel aktivite sıklığı (FAF)	Fiziksel aktivite sıklığı (1=hiçbir zaman, 2=haftada bir veya iki kez, 3=haftada iki veya üç kez, 4=haftada dört veya beş kez)	
Teknoloji cihazlarını kullanma süresi (TUE)	Bireylerin teknoloji cihazlarını kullanma süresi (0 = yok, 1 = bir saatten az, 2 = bir ile üç saat arası, 3 = üç saatten fazla)	
Kullanılan ulaşım aracı (MTRANS)	Bireylerin kullandığı ulaşım araçlarını gösteren kategorik değişken (otomobil, motosiklet, bisiklet, toplu taşıma, yürüyüş)	

Veri setinde yer alan her bir kayıt Yetersiz Kilo, Normal Kilo, Fazla Kilo Seviye I, Fazla Kilo Seviye II, Obezite Tip I, Obezite Tip II ve Obezite Tip III olmak üzere yedi obezite kategorisi olarak etiketlenmiştir. Şekil 1'de veri seti içerisinde yer alan ve çalışmada bağımlı değişken olarak kullanılan "NObeyesdad" alanın her biri kategorisine ait veri setinde yer alan kişi sayısı gösterilmektedir.



Şekil 1. Bağımlı değişken kategorileri ve kişi sayıları

Model Değerlendirme Metrikleri

Değerlendirme metrikleri makine öğrenimi görevlerine bağlıdır. Sınıflandırma ve regresyon görevleri için farklı metrikler vardır. Sınıflandırma, giriş verileri verilen sınıf etiketlerinin tahmin edilmesiyle ilgilidir ve denetimli öğrenme yöntemlerindendir. Sınıflandırma performansını ölçmenin birçok yolu vardır. Accuracy (doğruluk), confusion matrix (karışıklık matrisi), log-loss (log kaybı) ve AUC-ROC en popüler ölçümlerden bazılarıdır (Lever,2016). Precision (hassasiyet) ve recall (geri çağırma), sınıflandırma problemlerinde yaygın olarak kullanılan ölçümlerdir.

Confusion Matrix bir sınıflandırıcının, farklı sınıf etiketlerini ne ölçüde sınıflayabileceğini gösteren analiz aracıdır (Susmaga,2004). Tablo 2'de gösterildiği gibi Confusion matrix sınıflandırma sonuçlarını alır ve bunları dört kategoriye ayırır.

Tablo 2. Confusion matrix kategorileri

Gerçek Pozitif (TP)	Hem gerçek hem de tahmin edilen değerler 1 olduğunda elde edilen değer.
Gerçek Negatif (TN)	Hem gerçek hem de tahmin edilen değerler 0 olduğunda elde edilen değer.
Yanlış Pozitif (FP)	Gerçek değer 0 ancak tahmin edilen değer 1 olduğunda elde edilen değer.
Yanlış Negatif (FN)	Gerçek değer 1 ancak tahmin edilen değer 0 olduğunda elde edilen değer.

Confusion Matrix'de yer alan değerler kullanılarak, modellere ait accuracy, preccion ve recall hesaplamaları yapılabilmektedir. Accuracy, sınıflandırıcının ne sıklıkta doğru tahminde bulunduğunu ölçer. Doğru tahmin

sayısının toplam tahmin sayısına oranı olarak tanımlayabiliriz. Denklem 1'de accuracy hesaplama formülü verilmiştir.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Preccion, gerçek pozitiflerin model tarafından tahmin edilen tüm pozitiflere oranıdır. Çarpık ve dengesiz veri kümesi için kullanışlıdır. Model ne kadar çok Yanlış pozitif tahmin ederse, kesinlik de o kadar düşük olur. Denklem 2'de preccion hesaplama formülü verilmiştir.

$$Preccion = \frac{TP}{TP + FP} \quad (2)$$

Recall, gerçek pozitiflerin veri kümenizdeki tüm pozitiflere oranıdır. Modelin pozitif örnekleri tespit etme yeteneğini ölçer. Model ne kadar çok yanlış negatif tahmin ederse hatırlama oranı da o kadar düşük olur. Denklem 3'te recall hesaplama formülü verilmiştir.

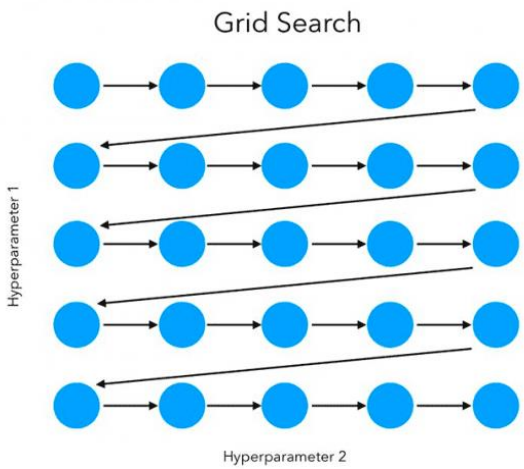
$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Hiper parametre Optimizasyonu

Makine Öğrenimi modeli, verilerden öğrenilmesi gereken bir dizi parametrenin yer aldığı matematiksel bir model olarak tanımlanabilir. Bir modeli mevcut verilerle eğiterek model parametrelerine uyum sağlayabiliriz. Bu parametrelerden farklı olarak model mimarisini tanımlayan parametreler de vardır ve hiper parametre olarak adlandırılır. Hiper parametreler gerçek eğitim süreci başlamadan önce ayarlanabilen değerlerdir. Bu parametreler ile modelin karmaşıklığı veya ne kadar hızlı öğrenmesi gerektiği gibi önemli özellikler belirlenebilir ve model performansı arttırılabilir.

Makine öğrenme modellerinde varsayılan olarak verilen hiper parametreler değerleri en iyi performansın sağlanacağını garanti etmemektedir (Young ve ark.,2015). Bu nedenle, hiper parametre değerlerinin ayarlanması (tune edilmesi) modelin performansını büyük ölçüde etkileyebilmektedir. Modellerin fazla miktarda hiper parametreye sahip olması bu değerlerin elle ayarlanmasını neredeyse imkânsız hale getirmektedir.

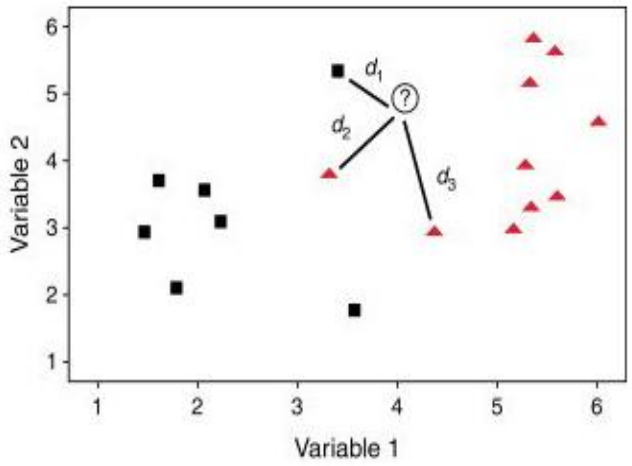
GridSearch, otomatik hiper parametre ayarlama için en sık kullanılan optimizasyon yöntemi olarak literatürde görülmektedir. GridSearch önceden tanımlanmış bir kümedeki olası her hiper parametre kombinasyonu için bir modelin eğitilmesini içeren bir ayarlama yöntemidir (Bashir,2016). Algoritma modelleri tüm olası kombinasyonlar için eğitilir ve doğruluk, F1 puanı gibi belirli bir ölçüm ile değerlendirir. Son aşamada kurulan modellerde, en iyi model performansıyla sonuçlanan hiper parametrelerin kombinasyonu optimal set olarak seçilir ve model bu parametre değerlerine göre eğitilerek en iyi tahmin sonucu elde edilir. Şekil 2'de GridSearch kullanılarak hiper parametre ayarlarının yapılması gösterilmektedir.



Şekil 2. GridSearch kullanılarak hiper parametre ayarlarının yapılması (URL-1, 20223)

K-en Yakın Komşu Algoritması (KNN)

KNN, benzer veri noktalarının benzer etiketlere veya değerlere sahip olma eğiliminde olduğu fikrine dayanan bir algoritmadır. Eğitim aşamasında tüm eğitim veri setini referans olarak saklar. Daha sonra seçilen bir mesafe ölçüsünü kullanarak giriş veri noktası ile tüm eğitim örnekleri arasındaki mesafeyi hesaplayarak tahminlerde bulunur. Şekil 3'te KNN algoritmasında değişkenler arasında uzaklık ölçümü ve sınıflandırılması gösterilmektedir.



Şekil 3. KNN algoritmasında uzaklık ölçümü ve sınıflandırma (URL-2, 2023)

Uzaklık ölçüsünün hesaplanması için euclidean, manhattan, minkowski, mahalanobis, tangential, cosine ölçüm yöntemleri kullanılabilir (Moosavian ve ark.,2013). Euclidean değeri, dikkate alınan iki noktayı birleştiren düz çizginin uzunluğu olarak tanımlanabilir. Bu ölçüm, bir nesnenin iki durumu arasında yapılan net yer değiştirmeyi hesaplamamıza yardımcı olur. Denklem 4'te Euclidean uzaklık hesabı verilmiştir.

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4)$$

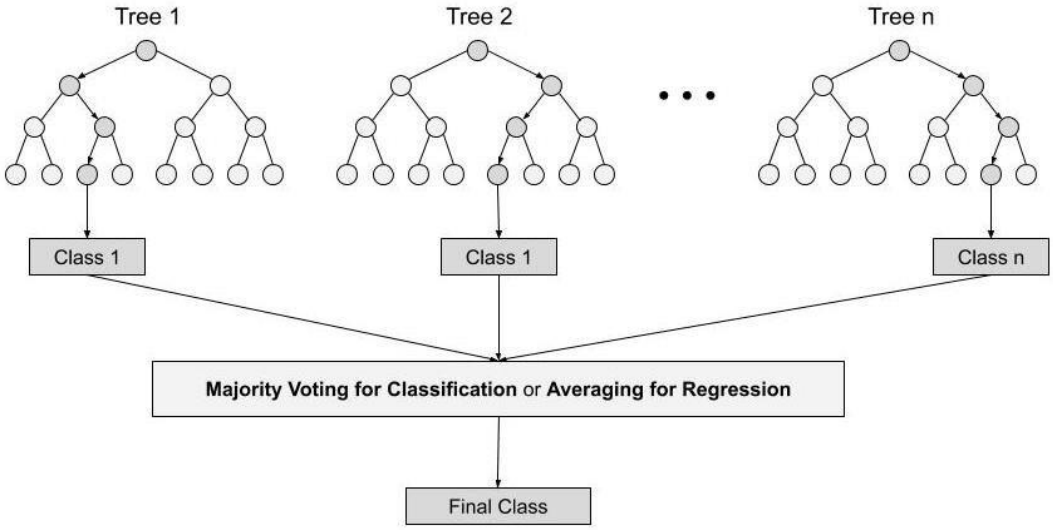
Manhattan ölçümü nesnenin kat ettiği toplam mesafeyi ölçmek için kullanılır. Bu metrik, n boyutlu noktaların koordinatları arasındaki mutlak farkın toplanmasıyla hesaplanır. Denklem 5'te Manhattan uzaklık hesabı verilmiştir.

$$\sum_{i=1}^{n}|x_i - y_i| \quad (5)$$

Denklemlerde yer alan n değeri "boyut sayısı", x değeri "veri kümesinden veri noktası" ve y değeri "yeni veri noktası (tahmin edilecek)" olarak tanımlanabilir.

Rastgele Orman Algoritması (RF)

RF, belirli bir veri kümesinin çeşitli alt kümelerindeki bir dizi karar ağacını içeren ve bu veri kümesinin tahmin doğruluğunu artırmak için ortalamayı alan bir sınıflandırıcıdır. RF, tek bir karar ağacına güvenmek yerine her ağaçtan tahmini alır ve tahminlerin çoğunluk oylarına dayanarak nihai çıktıyı tahmin eden algoritmadır. Şekil 4'te RF genel yapısı görülmektedir.



Şekil 4. RF genel yapısı

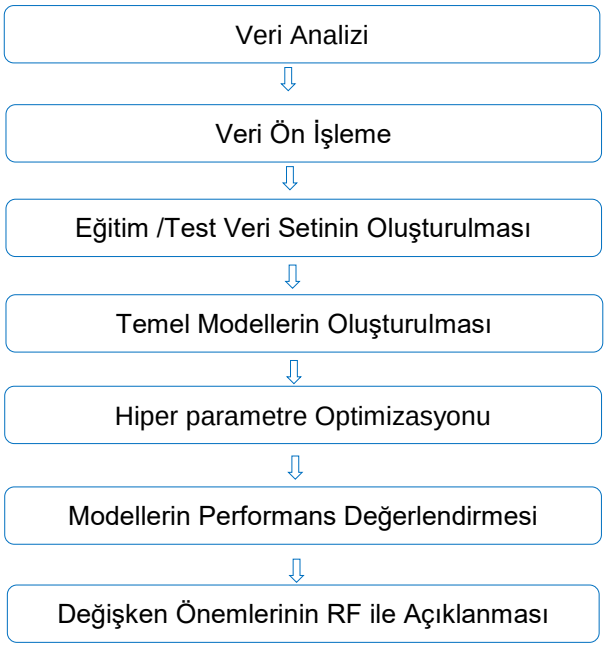
RF sınıflandırıcısı, nitelik seçim ölçüsü olarak Gini İndeksini kullanır. Belirli bir T eğitim seti için, rastgele bir durumun Ci sınıfına ait olup olmadığı hesaplamak için Gini indeksi Denklem 6'da gösterildiği gibi hesaplanır. Denklemde f (Ci, T) / |T| seçilen durumun Ci sınıfına ait olma olasılığıdır.

$$\sum\sum_{j\neq i}(f(C_i,T)/|T|)(f(C_j,T)/|T|) \quad (6)$$

Her seferinde bir ağaç, özelliklerin bir kombinasyonunu kullanarak yeni eğitim verileri üzerinde maksimum derinliğe kadar büyütülür. Tamamen büyümüş olan bu ağaçlar budanmaz. Bu, rastgele orman sınıflandırıcısının Quinlan (2014) tarafından önerilene benzer diğer karar ağacı yöntemlerine göre en büyük avantajlarından biridir. Çalışmalar, öznitelik seçim ölçütlerinin değil, budama yöntemlerinin seçiminin ağaç tabanlı sınıflandırıcıların performansını etkilediğini ileri sürmektedir (Mingers,1989; Pal ve ark.,2003). Breiman (2001) ağaç sayısı arttıkça genelleme hatasının her zaman ağacı budamadan bile yakınsadığını ve Büyük Sayılar Kuvvetli Yasası (Feller,1991) nedeniyle aşırı uyumun bir sorun olmadığını öne sürmektedir.

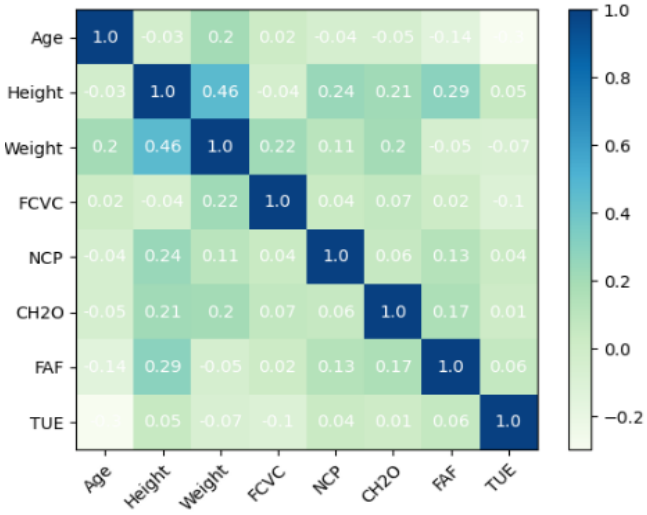
BULGULAR VE TARTIŞMA

Çalışmada obezite ile ilişkili önemli faktörleri tespit etmek ve obezite seviyelerinin tahmini için denetimli öğrenme modeli sınıflandırma tekniklerinden K-En Yakın Komşu algoritması ve Rastgele Orman algoritması ile modeller geliştirilmiştir. Çalışma sırasında gerçekleştirilen işlem adımları Şekil 5'te verilmiştir.



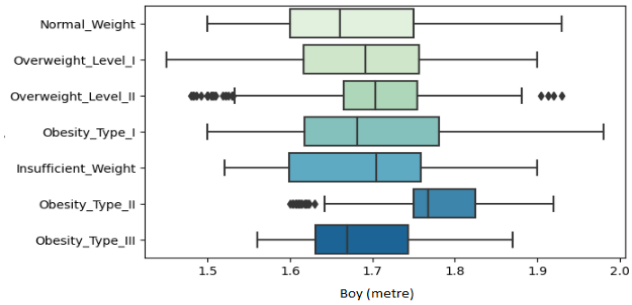
Şekil 5. Çalışma sırasında gerçekleştirilen işlem adımları

Veri kümesini analiz etmek için öncelikle korelasyon hesabı gerçekleştirilmiştir. Korelasyon hesabı ilişkili veya ilişkisiz değişkenleri verimli bir şekilde tanımasına olanak tanıyan bir hesaplamadır. Korelasyon katsayısı −1 ile 1 arasında olup, bu değer ne kadar 1'e yakınsa değişkenler arasındaki ilişki o kadar güçlü anlamına gelmektedir. Bu yöntemle yalnızca niceliksel verilerin ilişkisi belirlenebilir. Bu nedenle niceliksel değişkenlerden gelen sayıları niteliksel, kategorik değişkenlerin sayılarıyla karşılaştıramayız. Şekil 6'da veri setinde yer alan niceliksel değişkenlerin birbiri ile olan ilişkisi ısı grafiği ile gösterilmektedir.

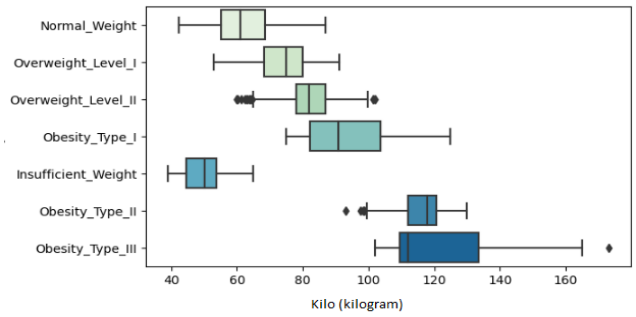


Şekil 6. Nicel değişkenlerin ısı grafiği

Grafiğe göre boy ve kilo nicel değişkenlerinin 0,46 korelasyon değeri ile en yüksek ilişki seviyesine sahip olan iki değişken oldukları görülmektedir. En yüksek korelasyon değerine sahip bu değişkenlerin obezite seviyesi tahmininin ayrılmaz bir parçası olduğu görülmüştür. Şekil 7'de boy değişkeni, Şekil 8'de kilo değişkeni için veri setinde yer alan katılımcı dağılımı gösterilmektedir.

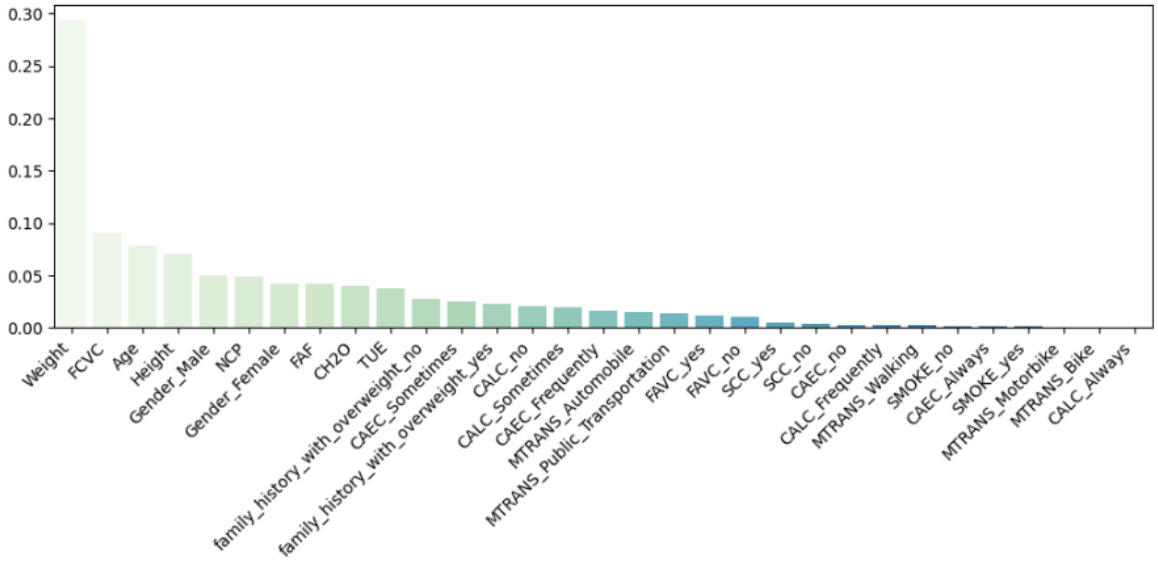


Şekil 7. Boy değişkeninin obezite seviyelerine göre kişi sayısı



Şekil 8. Kilo değişkeninin obezite seviyelerine göre kişi sayısı

Değişken özellik önemi, bağımlı değişkeni tahmin etmede bağımsız değişkenlerin ne kadar yararlı olduklarını gösteren, bir puan atama yöntemidir. Çalışmada bağımsız değişkenlerin önem dereceleri Rastgele Orman Önem Hesaplaması ile ortaya koyulmuştur. Şekil 9'da sınıflandırma görevi için bağımsız değişkenlerin önemini seviyeleri gösterilmektedir. Şekle göre weight, FCVC, age, height değişkenleri sonuca en çok etki eden değişkenler olarak görülürken, smoke, MTRANS ve CALC değişkenleri sonuca etkisi en az olan değişkenler olarak görülmektedir. Çalışmada modelinizin karmaşıklığını azaltmak için daha az etkili değişkenler kaldırılmıştır. Bu durum aşırı uyumun önlenmesine ve model sonuçlarının iyileştirilmesine yardımcı olmuştur.



Şekil 9. Rastgele Orman ile değişken önem hesaplaması

Veri ön işleme, ham verinin anlaşılır bir formata dönüştürülmesi işlemidir. Ham verilerle çalışamayacağımız için veri madenciliğinde de önemli bir adımdır. Çalışmada ilk olarak veri setinde yer alan object ve text veri türüne sahip değişkenler kategorik veri türüne dönüştürülmüştür. Daha sonra float veri türündeki değişkenler en yakın tam sayı veri türüne dönüştürülmüştür.

Çalışmada makine öğrenimi modellerini eğitmeden önce Holdout yöntemi ile veriler %70 eğitim, %30 test verisi olmak üzere iki kümeye bölünmüştür. Buna göre veri setinde yer alan 2111 kayıttan 1478 tanesi eğitim seti, 633'ü test seti olarak belirlenmiştir. Daha sonra K-En Yakın Komşu algoritması ve Rastgele Orman algoritması için temel modeller kurulmuştur. KNN temel modeli ortalama %87 doğruluk değeri ile obezite seviyelerini tahmin etmiştir. Temel modelinin her obezite sınıfı için elde ettiği precision, recall, f1-score performans değerleri Tablo 3'de gösterilmektedir.

Tablo 3. KNN temel modeli tahmin performans değerleri

Obezite Sınıfı	Precission	Recall	F1-Score	Accuracy
Yetersiz Kilo	0.80	0.98	0.88	0.87
Normal Kilo	0.87	0.51	0.64	
Obezite Tip I	0.87	0.90	0.89	
Obezite Tip II	0.97	0.98	0.89	
Obezite Tip III	0.99	1.00	0.99	
Fazla Kilo Seviye I	0.82	0.94	0.87	
Fazla Kilo Seviye II	0.78	0.71	0.74	

RF temel modeli ortalama %93 doğruluk değeri ile KNN modele göre daha yüksek bir tahmin başarısında bulunmuştur. RF temel modelinin her obezite sınıfı için elde ettiği precision, recall, f1-score performans değerleri Tablo 4'de gösterilmektedir.

Tablo 4. RF temel modeli tahmin performans değerleri

Obezite Sınıfı	Precission	Recall	F1-Score	Accuracy
Yetersiz Kilo	0.99	0.95	0.97	0.93
Normal Kilo	0.80	0.90	0.85	
Obezite Tip I	0.98	0.95	0.96	
Obezite Tip II	0.99	1.00	0.99	
Obezite Tip III	1.00	1.00	1.00	
Fazla Kilo Seviye I	0.89	0.81	0.85	
Fazla Kilo Seviye II	0.87	0.93	0.90	

Temel modellerden elde edilen tahmin değerlerinin iyileştirilmesi için, hiper parametre optimizasyonu gerçekleştirilmiştir. Optimizasyon için literatürde en sık kullanılan yöntemlerden olan scikit–learn kütüphanesindeki GridSearch nesnesi kullanılmıştır. GridSearchCv ile modelde denenmesi istenen hiper parametre değerleri için ayrı ayrı modeller kurulmuştur. KNN algoritmasında n-neighbours (komşu sayısı), metric (mesafe ölçüm yöntemi) ve weight (ağırlık) hiper parametreleri için en başarılı sonucu veren hiper parametre değerleri belirlenmiştir. RF algoritmasında criterion (bölünme yöntemi), max-depth (maksimum derinlik) ve n_estimators (ağaç sayısı) hiper parametreleri için en başarılı sonucu veren hiper parametre değerleri belirlenmiştir. Modeller için en iyi doğruluk sonucunun elde edilmesini sağlayan hiper parametre değerleri Tablo 5'te gösterilmektedir.

Tablo 5. Modellere ait hiperparametre değerleri

Modeller	Hiperparametre	Değer
K-en Yakın Komşu Modeli	n_neighbors	4
	p(metric)	manhattan
	weights	distance
Rastgele Orman Modeli	criterion	entropy
	max_depth	90
	n_estimators	160

Hiperparametre optimizasyonundan sonra nihai modeller kurulmuş ve modellerin sınıflandırma performansları precision, recall, f1-score ve accuracy değerlendirme ölçütleri ile karşılaştırılmıştır. KNN optimize edilmiş modeli ortalama %91 doğruluk değeri ile obezite seviyelerini tahmin etmiştir. Optimize edilmiş modelinin her obezite sınıfı için elde ettiği precision, recall, f1-score performans değerleri Tablo 6'da gösterilmektedir.

Tablo 6. KNN optimize edilmiş model tahmin performans değerleri

Obezite Sınıfı	Precission	Recall	F1-Score	Accuracy
Yetersiz Kilo	0.87	1.00	0.93	0.91
Normal Kilo	0.85	0.60	0.70	
Obezite Tip I	0.92	0.95	0.93	
Obezite Tip II	0.98	1.00	0.99	
Obezite Tip III	1.00	1.00	1.00	
Fazla Kilo Seviye I	0.87	0.89	0.88	
Fazla Kilo Seviye II	0.86	0.87	0.87	

RF optimize edilmiş modeli ortalama %94 doğruluk değeri ile KNN modele göre daha yüksek bir tahmin başarısında bulunmuştur. RF temel modelinin her obezite sınıfı için elde ettiği precision, recall, f1-score performans değerleri Tablo 7'de gösterilmektedir.

Tablo 7. RF optimize edilmiş model tahmin performans değerleri

Obezite Sınıfı	Precission	Recall	F1-Score	Accuracy
Yetersiz Kilo	0.99	0.93	0.96	0.94
Normal Kilo	0.81	0.91	0.86	
Obezite Tip I	0.98	0.95	0.96	
Obezite Tip II	0.99	1.00	0.99	
Obezite Tip III	1.00	1.00	1.00	
Fazla Kilo Seviye I	0.87	0.84	0.86	
Fazla Kilo Seviye II	0.90	0.92	0.91	

Tablo hiper parametre optimizasyonu ile elde edilen doğruluk değerlerinin daha iyi olduğunu göstermektedir.

SONUÇ

Obezite küresel nüfusun büyük bir bölümünü etkileyen ciddi bir halk sağlığı sorunudur. Doğru analiz edildiği takdirde obezite oranlarının kontrol altına alınabileceği görülmüştür. Bu çalışmada Meksika, Peru ve Kolombiya ülkelerinden bireylerin beslenme alışkanlıklarına ve fiziksel durumlarına göre obezite düzeylerinin tahmin edilmesine yönelik modeller geliştirilmiştir. Çalışma ile insanların neden fazla kilolu olduğunu belirlememize ve kişinin fazla kilolu olup olmayacağını tahmin etmemize yardımcı olabilecek bir model oluşturmak amaçlanmıştır. Bunun için ilk olarak veri analizi gerçekleştirilmiş ve veriler arasındaki bağlantılar incelenmiştir. Bu inceleme sonucunda nicel değişkenlerden boy ve kilo bağımsız değişkenlerinin obezite seviyesini belirlenmesinde en etkili değişkenler olduğu görülmüştür. Daha sonra Rastgele Ormanlar Önem Hesaplaması yapılmış ve veri setindeki en önemli 13 bağımsız değişkenler belirlenmiştir. Bağımsız değişkenlerin belirlenmesinden sonra denetimli öğrenme algoritmalarından KNN ve RF ile önce temel modeller daha sonra optime edilmiş modeller kurulmuştur. Çalışmada geliştirilen RF modeli, optimizasyon işleminden sonra %94'lük doğruluk değeri ile en iyi performans gösteren model olmuştur. Geliştirilen model obezite seviyelerinden “yetersiz kilo” ve “aşırı kilo-1” tahminlerini %100’e yakın bir tahminleme başarısı ile de dikkat çekmektedir. Geliştirilen çalışma ve literatürde yer alan çalışmaların modelleri, tahmin başarı oranları ve kullandıkları yöntemler Tablo 8’de karşılaştırılmıştır. Buna göre çalışmanın elde ettiği başarı oranı ve kullandığı yöntemleri ile ön plana çıktığı görülmektedir.

Tablo 8. Literatüde yer alan çalışmaların ve gerçekleştirilen çalışmanın karşılaştırılması

Çalışma Adı	Model	Accuracy
Five Machine Learning Supervised Algorithms for The Analysis and the Prediction of Obesity (Clem’s ve ark., 2022)	RF	%91
Estimation of Obesity Levels Based on Decision Trees (Cui ve ark., 2021)	XGBoost	%85,99
Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique (Yagin ve ark., 2023)	DL	%93.06
OBESEYE: Interpretable Diet Recommender for Obesity Management using Machine Learning and Explainable AI (Roy ve ark., 2023)	LightGBM	%86.02
Using machine learning to predict obesity in high school students (Zheng ve ark., 2017)	KNN	%88,82
Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People(Singh ve ark., 2020)	MLP	%90
Machine learning approaches for the prediction of obesity using publicly available genetic profiles(Montañez ve ark., 2017)	SVM	%90,5
Gerçekleştirilen Çalışma	RF	%94

Gerçekleştirilen çalışma ile bireylerin obeziteden mustarip olup olmayacakları tahmin edebilir ve uzmanlar tarafından onlara bazı önerilerde bulunulmasında yol gösterebilir. Gelecek çalışmalarda veri seti üzerinde farklı yapay zeka algoritmaları ile modeller kurulup, uygulamalar geliştirilebilir ve obezite hastalığının engellenmesi için yapılan çalışmalara katkı sağlanabilir.

KAYNAKLAR

Bansal, S., Jin, Y. (2023). Heterogeneous Effects of Obesity on Life Expectancy: A Global Perspective. *Annual Review of Resource Economics*, 15; DOI 10.1146/annurev-resource-022823-033521.

Bashir, M. B., Abd Latiff, M. S. B., Coulibaly, Y.,Yousif, A. (2016). A survey of grid-based searching techniques for

large scale distributed data. *Journal of Network and Computer Applications*, 60, 170-179.

Breiman, L. (2001). Random forests. *Machine learning*, 45: 5-32.

Brero, M., Meyer, C. L., Jackson-Morris, A., Spencer, G., Ludwig-Borycz, E., Wu, D., Nugent, R. (2023). Investment case for the prevention and reduction of childhood and adolescent overweight and obesity in Mexico. *Obesity Reviews*, 24(9); DOI 10.1111/obr.13595.

Caballero, B. (2005). A nutrition paradox underweight and obesity in developing countries. *New England Journal of Medicine*, 352(15): 1514-1516.

Clem's, M. L., Manıamfu, P., Louıson, D. K. Five Machine Learning Supervised Algorithms for The Analysis and the Prediction of Obesity. *International Journal of Innovative Science and Research Technology,* 7(12):1956-1964.

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.

Danacı, Ç., Avcı, D. Arslan Tuncer, S. (2023). Komşuluk Bileşen Analizi Tabanlı Makine Öğrenimi Yöntemleri ile Obezite Seviyelerinin Tahmini. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 35 (2): 433-442.

Dugan, T. M., Mukhopadhyay, S., Carroll, A., Downs, S. (2015). Machine learning techniques for prediction of early childhood obesity. *Applied clinical informatics*, 6(3):506-520.

Feller, W. (1991). *An introduction to probability theory and its applications,* John Wiley & Sons.Ltd, England.

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; DOI 10.1016/j.crbeha.2021.100053.

Jeon, J., Lee, S., Oh, C. (2023). Age-specific risk factors for the prediction of obesity using a machine learning approach. *Frontiers in Public Health*, 10; DOI 10.3389/fpubh.2022.998782.

Koo, H. C., Tan, L. K., Lim, G. P., Kee, C. C., Omar, M. A. (2023). Obesity and Its Association with Undiagnosed Diabetes Mellitus, High Blood Pressure and Hypercholesterolemia in the Malaysian Adult Population: A National Cross-Sectional Study Using NHMS Data. *International Journal of Environmental Research and Public Health*, *20*(4); DOI 10.3390/ijerph20043058.

Lever, J. (2016). Classification evaluation: It is important to understand both what a classification metric expresses and what it hides. *Nature methods*, 13(8): 603-605.

Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine learning*, 4: 227-243.

Mondal, P. K., Foysal, K. H., Norman, B. A., Gittner, L. S. (2023). Predicting Childhood Obesity Based on Single and Multiple Well-Child Visit Data Using Machine Learning Classifiers. *Sensors*, 23(2): 759; DOI 10.3390/s23020759.

Montañez, C. A. C., Fergus, P., Hussain, A., Al-Jumeily, D., Abdulaimma, B., Hind, J., Radi, N. (2017). Machine learning approaches for the prediction of obesity using publicly available genetic profiles. 2017 International Joint Conference on Neural Networks (IJCNN), May 14-19,2017, Online,IEEE, 2743-2750.

Moosavian, A., Ahmadi, H., Tabatabaeefar, A., Khazaee, M. (2013). Comparison of two classifiers; K-nearest neighbor and artificial neural network, for fault diagnosis on a main engine journal-bearing. *Shock and Vibration*, 20(2), 263-272.

Ogden, C. L., Carroll, M. D., Curtin, L. R., McDowell, M. A., Tabak, C. J., Flegal, K. M. (2006). Prevalence of overweight and obesity in the United States, 1999-2004.

Pal, M., Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote sensing of environment*, 86(4): 554-565.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Roy, M., Das, S., & Protity, A. T. (2023). OBESEYE: Interpretable Diet Recommender for Obesity Management using Machine Learning and Explainable AI. *arXiv preprint*. 4(6); DOI https://doi.org/10.48550/arXiv.2308.02796.

Singh, B., Tawfik, H. (2020). Machine learning approach for the early prediction of the risk of overweight and obesity in young people. In *Computational Science–ICCS 2020: 20th International Conference, June 3–5, 2020, Amsterdam, The Netherlands,* 523-535.

Susmaga, R. (2004). Confusion matrix visualization. In Intelligent Information Processing and Web Mining, May 17–20,2004, Berlin, Heidelberg.107-1116.

UCI (2023). https://archive.ics.uci.edu/ (Erişim Tarihi: 10.06.2023)

URL-1 (2023). https://www.researchgate.net/publication/278050782_The_use_of_the_k_nearest_neighbor_method_to_classify_the_representative_elements

URL-2 (2023). https://pyimagesearch.com/2021/05/24/grid-search-hyperparameter-tuning-with-scikit-learn-gridsearchcv/

Vizmanos, B., Cascales, A. I., Rodríguez-Martín, M., Salmerón, D., Morales, E., Aragón-Alonso, A., Garaulet, M. (2023). Lifestyle mediators of associations among siestas, obesity, and metabolic health. *Obesity*, 31(5): 1227-1239.

Wanjau, M. N., Kivuti-Bitok, L. W., Aminde, L. N., & Veerman, J. L. (2023). The health and economic impact and cost effectiveness of interventions for the prevention and control of overweight and obesity in Kenya: a stakeholder engaged modelling study. *Cost Effectiveness and Resource Allocation*, 21(1): 1-21.

WHO (2022). European regional obesity report 2022. https://www.who.int/europe/publications/i/item/9789289057738 (Erişim Tarihi: 15.07.2023)

Yagin FH, Gülü M, Gormez Y, Castañeda-Babarro A, Colak C, Greco G, Fischetti F, Cataldi S.(2023) Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique. *Applied Sciences*. 13(6); DOI 10.3390/app13063875

Young, S. R., Rose, D. C., Karnowski, T. P., Lim, S. H., Patton, R. M. (2015). Optimizing deep learning hyper-parameters through an evolutionary algorithm. MLHPC '15: Proceedings of the Workshop on Machine Learning in High-

Performance Computing Environments, Nov 15,2015, New York,United States,1-15.

Zheng, Z., Ruggiero, K. (2017). Using machine learning to predict obesity in high school students. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Nov 13-16,2017, Online,IEEE. 2132-2138.