Bilgisayar Bilimleri ve Mühendisliği Dergisi

Cilt-17
Sayı-2
2024

Makine Öğrenmesi Temelli Obezite Durum Tahmini

Machine Learning Based Obesity Status Prediction

Ercan ÖLÇER
Kocaeli Sağlık ve Teknoloji Üniversitesi, Doğa Bilimleri ve Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü
Kocaeli - Türkiye
ercan.olcer@kocaelisaglik.edu.tr
ORCID: 0000-0003-3786-6230

Öz

Obezite ciddi bir halk sağlığı sorunudur ve Dünya üzerinde gittikçe artış göstermektedir. Biyolojik, fizyolojik, psikolojik ve çevresel faktörlerden etkilenen karmaşık bir konudur. Yaşam kalitesini olumsuz etkileyen bir hastalık olarak kabul edilmektedir. Yüksek tansiyon, koroner arter hastalığı, kalp krizi, uyku apnesi, nefes alma zorluğu, eklem ağrısı ve osteoartrit eklem hastalıklarının oluşmasına neden olabilir. Ayrıca çeşitli kanser türlerinin görülme riski obez bireylerde daha yüksektir. Yüksek tansiyon, yüksek kan şekeri, yüksek trigliserid seviyeleri ve düşük HDL kolesterol seviyeleri gibi faktörlerin bir araya gelmesiyle oluşan metabolik sendrom riskini de artırır. Çalışma, makine öğrenimi sınıflandırıcıları kullanarak obezite tahmini için risk faktörlerini belirlenmesini amaçlamaktadır. Makine öğrenimi yöntemleri özellikle büyük veri kümelerinin analiz edilmesi ve bu verilerden obezitenin ana belirleyici değişkenlerini saptanmasını kolaylaştırır. Bu yöntemlerin uygulanması ile risk faktörlerinin öncelikle belirlenerek takibinin kolaylaşmasını sağlayabilir. Makine öğrenimi, obeziteyle ilgili sonuçların anlaşılması ve tahmin edilmesi için umut verici bir yol sunmaktadır. Araştırmacılar, büyük veri kümelerinden ve karmaşık algoritmalardan yararlanarak obezitenin temel belirleyicilerini ve risk faktörlerini belirleyebilir ve bu da önleme ve müdahale stratejilerine bilgi sağlayabilir. Geliştirilmiş algoritmalar, gelişmiş tahmin başarımını ortaya koyar ve çeşitli veri kaynaklarının entegrasyonu, obezite tahmin modellerini daha da geliştirebilir. Bu bilgiler, küresel obezite problemine yönelik müdahalelerin geliştirilmesine rehberlik edebilir. Bu yöntemlerin uygulanması ile elde edilecek sonuçların kullanımı ile hastalık tanılamada uzmanlara yardımcı olacaktır ve karar vermelerinde destek sağlayacaktır. Çalışmada "kaggle" ortamında temin edilen ve içeriğinde on yedi parametreyi barındıran veri seti kullanılarak gerçekleştirilen makine öğrenmesi yöntemleri kullanılmıştır. Elde edilen sınıflandırma sonuçları hastalık riski olabilecek hastaların verilerinin diyabet risklerini belirlemede kullanılabilir. Bu veri kümesi üzerinde on bir farklı makine öğrenme algoritması kullanılmıştır. Bu makalede bu yöntemler karşılaştırılarak tahminde en başarılı yöntemler belirlenmiştir. Örneklemelerde veri seti içinden eğitim ve test seti oluşturulmuştur. Algoritmaların başarımı çeşitli metriklerle karşılaştırılmıştır. Ayrıca en başarılı birkaç algoritma değişkenlerin bazılarının ince ayar yapılarak başarımı arttırılmıştır. Uygulanan sınıflandırıcı algoritmalarından en başarılı başarım, Gradient Boost ve XGBoost kullanan modeller olmuştur. Bu modeller test verileri üzerinde %97 doğruluk değerini elde etmiştir. Literatür taramasında bu çalışmada elde edilen sonucun en iyi sonuç olduğu görülmektedir. Kısıtlı özelliklerle obezite konusunun çalışıldığı makalelere göre farklı özellikleri de dikkate alan bir veri kümesi olması ve on bir farklı modelleme ile obezite sınıflandırmalarının makine öğrenmesinin yapılması açısından da anlamlı bir çalışma olduğu düşünülmektedir.

***Anahtar sözcükler*:** Makine Öğrenmesi, Obezite, Tanı

Abstract

Obesity is a serious public health problem and is increasing worldwide. It is a complex issue affected by biological, physiological, psychological and environmental factors. It is recognized as a disease that negatively affects quality of life. It can cause high blood pressure, coronary artery disease, heart attack, sleep apnea, breathing difficulties, joint pain and osteoarthritis joint diseases. In addition, the risk of various types of cancer is higher in obese individuals. It also increases the risk of metabolic syndrome, a combination of factors such as high blood pressure, high blood sugar, high triglyceride levels and low HDL cholesterol levels. The study aims to

Gönderme, düzeltme ve kabul tarihi: 17.04.2024 - 21.07.2024 – 09.08.2024
Makale türü: Araştırma

identify risk factors for obesity prediction using machine learning classifiers. Machine learning methods facilitate the analysis of large datasets and the identification of key determinants of obesity from these data. By applying these methods, risk factors can be identified first and follow-up can be facilitated. Machine learning offers a promising way to understand and predict obesity-related outcomes. By leveraging large datasets and complex algorithms, researchers can identify key determinants and risk factors of obesity, which can inform prevention and intervention strategies. Improved algorithms reveal improved prediction performance, and the integration of various data sources can further enhance obesity prediction models. This information can guide the development of interventions to address the global obesity problem. Using the results obtained by applying these methods will help experts in disease diagnosis and support decision-making. It is considered to be a meaningful study in terms of being a dataset that takes into account different features compared to the articles where obesity is studied with limited features and machine learning of obesity classifications with eleven different modeling.

Keywords: Machine Learning, Obesity, Diagnosis

1. Giriş

Obezite, sağlığı olumsuz etkilere neden olabilecek derecede yüksek oranda vücuttaki yağın birikmesi olarak tanımlanabilir. Kişi ağırlığının boyunun karesine bölünmesiyle vücut kitle endeksi elde edilir. Bu endeks "BMI" olarak tanımlanır. BMI istatistiklerde aşırı kilo ve obezitenin bir göstergesi olarak tanımlanır [1]. Yetişkinlerde 30 kg/m2 aşıldığında kişi obez olarak ifade edilir. Bazı Asya ülkelerinde daha düşük değerler öngörülmektedir [2].

Obezite, hayatı zorlaştıran önemli bir nedendir. Tip 2 diyabet, kardiyovasküler hastalıklar, uyku apnesi ve kanser tipleri gibi hastalıkların gelişmesine sebep olmaktadır [3]. Obezite oranları Dünya'da 1975 yılından beri üç kat arttığı ifade edilmektedir. 2022 yılında dünyadaki her 8 kişiden 1'i obez olduğu kaydedilmiştir. Dünya çapında yetişkin obezitesi 1990'dan bu yana iki kattan fazla, ergen obezitesi ise dört kat artmıştır. 2,5 milyar yetişkin (18 yaş ve üzeri) fazla kilolu olduğu ifade edilmiştir. Bunlardan 890 milyonu obeziteyle yaşadığı raporlanmıştır. 2022 yılında 18 yaş ve üzeri yetişkinlerin %43'ü aşırı kilolu, %16'sı ise obezdir. 5 yaş altı 37 milyon çocuk aşırı kilolu olduğu, 5-19 yaş arası 390 milyondan fazla çocuk ve ergen aşırı kilolu ve bunların 160 milyonu obez olduğu ifade edilmiştir [4].

Dünya genelinde obezitenin, önümüzdeki yıllarda artmaya devam edeceği ifade edilmektedir. Dünya Obezite Federasyonu'nun yayımladığı rapora göre, 2030 yılına kadar obez insan sayısının 1 milyar artması beklenmektedir. Bu durum, sağlık sistemleri ve bireyler için önemli bir sorun oluşturmaktadır. Avrupa ülkelerinde yetişkinlerin yüzde 30'unun obez olacağı öngörülmektedir. Aynı zamanda 2010 yılına göre 2030 yılındaki obez insan sayısının dünya genelinde 2 katına çıkması beklenmektedir [5].

Son zamanlarda makine öğrenme yöntem ve modelleri, sağlıkta, özellikle hastalıkların sınıflandırılmasında yaygın olarak kullanılmaya başlanmıştır. Sağlık hizmetlerinin yaygınlaşması ve giyilebilir teknolojiler aracılığı ile artık yüksek miktarda sağlık verisi toplanabilmektedir. Ancak yüksek veri yoğunluğu nedeniyle önemli miktarda verinin değerlendirilebilmesi için makine öğrenmesi, derin öğrenme gibi yeni yöntem ve akıllı sistemlere ihtiyaç duyulmaktadır. Hastalıkların değerlendirilmesinde makine öğrenme algoritmaları kullanılabilmektedir. Obezitenin hem erken tespitinde hem de izlenmesinde bu algoritmalar kullanılabilir. Ayrıca makine öğrenmesi yöntemleri insan nüfusun obezite tahmin etme yöntemi, başka hastalıkların da önemli sebeplerinden biri olarak kabul edilen obezitenin önüne geçilmesi için uygulanabilir.

Bu makalede "Kaggle" obezite veri kümesi kullanılarak on bir farklı makine öğrenmesi algoritması uygulanmıştır. Test verilerinin obezite sınıflarının tahmin edilmesinde kullanılması ile doğruluk, kesinlik, duyarlılık ve F1 ölçüsü başarım ölçütleri hesaplamaları ile algoritmaların başarımları hesaplanmıştır. Başarım karşılaştırma tablosunda yer alan en iyi üç skor üzerinde hiper parametre iyileştirmesi yapılarak daha yüksek başarım değerine ulaşılmıştır. İlerleyen kısımlarda ise, önceki yıllarda literatürde geçen benzer çalışmaların başarım karşılaştırma yapılmıştır.

Kısıtlı özelliklerle obezite konusunun çalışıldığı makalelere göre farklı özellikleri de dikkate alan bir veri kümesi olması ve on bir farklı modelleme ile obezite sınıflandırmalarının makine öğrenmesinin yapılması açısından da anlamlı bir çalışma olduğu düşünülmektedir.

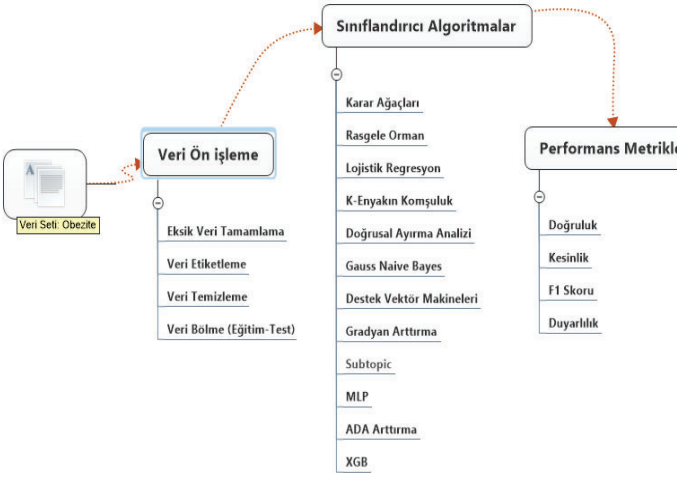
2. Veri Kümesi ve Yöntem

2.1 Önerilen Model

Şekil-1'de bu makaleye konu olan çalışmada izlenen yol özetlenmiştir. "csv" Formatında Intel i7-1260p işlemcili 64GB RAM'li lokal makineye indirilen veri kümesi Python dilinde yazılan bir yazılım aracılığı ile öncelikle veri ön işleme yöntemleri uygulanarak elden geçirilmiştir. Veri ön işleme aşamasında verilerin tekrar eden satırların temizlenmesi ve eksik verilerin tamamlanması işlemlerine tabi tutulmuştur. Ayrıca verilerin sınıflandırıcı algoritmalar tarafından işlenebilmesi için gerekli etiketleme işlemleri gerçekleştirilmiştir. Sonraki aşamada ise makine öğrenmesi süreçlerinde kullanılmak üzere veri eğitim ve test verileri olarak (0,8/0,2) oranı ile ikiye ayrılmıştır. Sınıflandırıcı algoritmalarına uygulanarak başarımları doğruluk, tutturma (precision), bulma (recall) ve F1-ölçüsü (F1Score) sınıflarında karşılaştırmaya tabi tutulmuştur.

2.2. Veri Kümesi

Kaggle web sitesinden elde edilen veri kümesi [6], bireylerin demografik özellikleri, fiziksel özellikleri ve yaşam tarzı alışkanlıkları hakkında kapsamlı bilgi sağlayarak obezite durum analizini ve tahminini kolaylaştırmayı amaçlamaktadır. Yaş, cinsiyet, boy, kilo, fiziksel aktivite düzeyi ve obezite kategorisi gibi temel değişkenleri içerir ve obezite sonuçlarını etkileyen faktörlere ilişkin değerli bilgiler sunar.

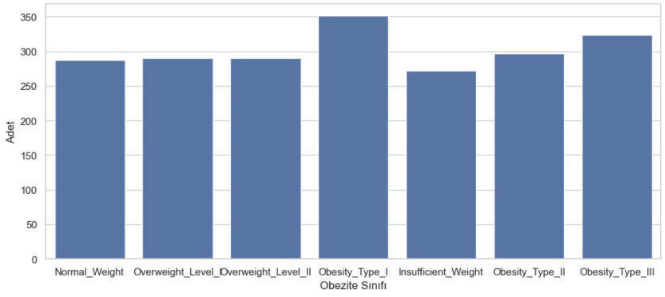


Şekil-1. Önerilen Model

Çizelge-1. Obezite Sınıfları

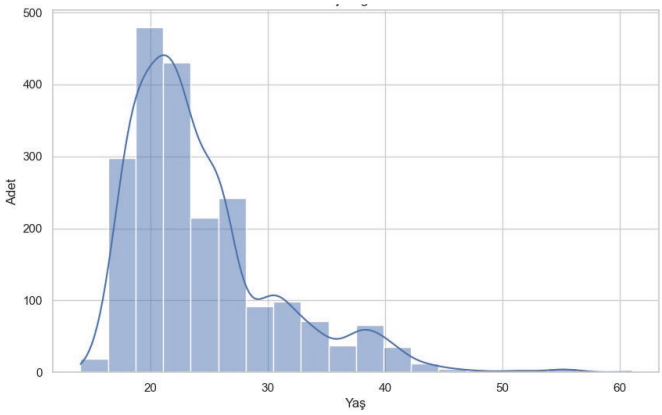
Obezite ("NObeyesdad")
Yetersiz kilo (insufficient wight)
Normal kilo (normal wight)
Aşırı kilolu Düzey 1 (overweight wight level 1)
Aşırı kilolu Düzey 2 (overweight wight level 2)
Obezite Tip 1 (obesity type 1)
Obezite Tip 2 (obesity type 2)
Obezite Tip 3 (obesity type 3)

Kullanılan veri kümesi farklı yaş ve cinsiyetten 17 özelliği içeren 2111 kaydı barındırmaktadır. Veri kümesinde obezite durumunu gösteren ve bağımlı değişken olarak tanımlanan "NObeyesdad" isimli veri alanında "yetersiz kilo, normal kilo, aşırı kilolu düzey I, fazla kilolu düzey II, obezite tip I, obezite tip II ve obezite tip III" olarak tanımlanmış ve Çizelge-1'de gösterilmiştir. Aynı şekilde bağımsız değişkenlerin listesi de Çizelge-2'de açıklamaları ile birlikte görülebilir.



Şekil-2. Obezite Sınıf Dağılımı

Obezite sınıflarının dağılımı Şekil-2'de yer almaktadır. Genel olarak eşit bir dağılım olduğu görülmektedir.

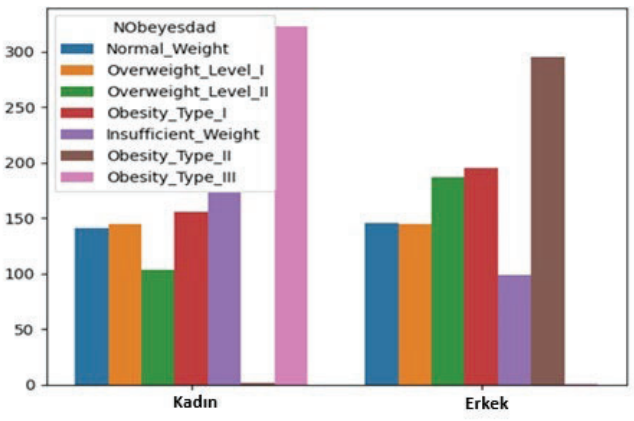


Şekil-3. Yaş Dağılımı

Katılımcıların yaş dağılımları Şekil-3'te görüldüğü gibidir. Ağırlıklı olarak katılımcılar 18-22 yaşları arasında yer almaktadır.

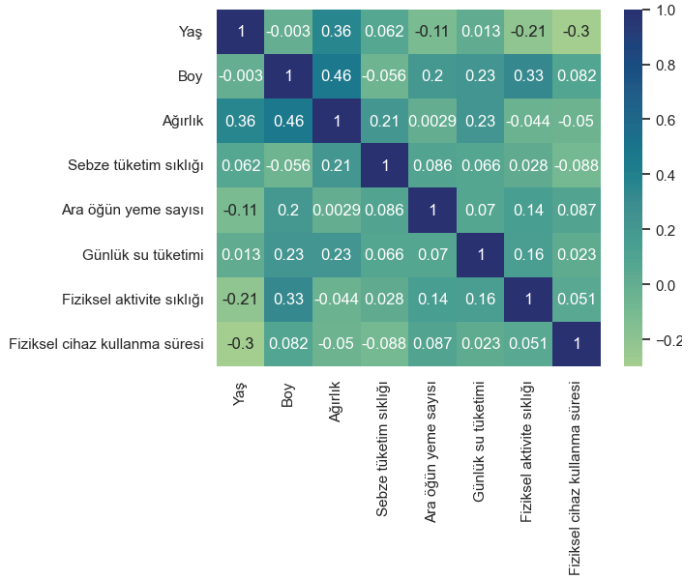
Çizelge-2. Veri Kümesi

Değişken	Açıklama
Gender	Cinsiyet
Age	Yaş
Height	Boy
Weight	Kilo
Family history overweight	Ailede fazla kilo öyküsü
FAVC	Yüksek kalorili gıda tüketimi
FCVC	Sebze tüketim sıklığı
NCP	Ara öğün yeme sayısı
CAED	Öğünler arası yeme sıklığı
CH20	Günlük su tüketimi
CALC	Alkol tüketimi
SMOKE	Sigara tüketimi
SCC	Kalori tüketimi
FAF	Fiziksel aktivite sıklığı
TUE	Fiziksel cihaz kullanma süresi
MTRANS	Kullanılan ulaşım türü



Şekil-4. Obezite Sınıfları ve Cinsiyet Dağılımı

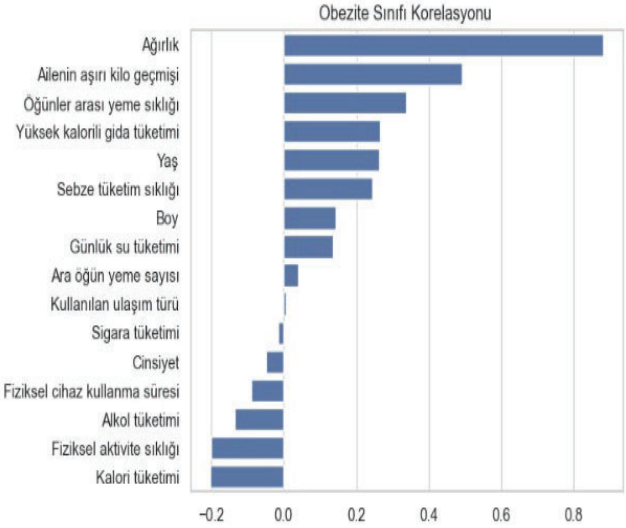
Obezite ile cinsiyet arasındaki dağılım Şekil-4'te görülmektedir. Buna göre obezite 3 tipi kadınlarda, obezite 2 tipi erkeklerde yüksek sayıda olduğu anlaşılmaktadır.



Çizelge-3. Isı Haritası

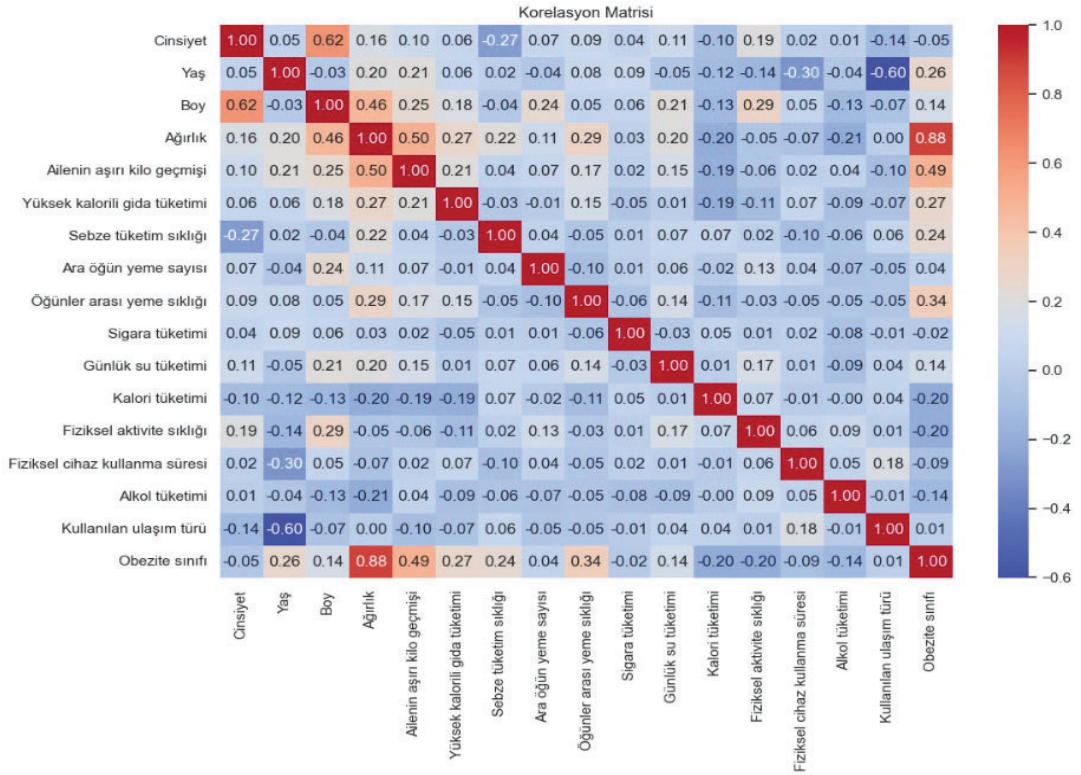
Çizelge-3'te yer alan ısı haritasında boy ve ağırlık arasında 0,46 oranında yüksek bir ilişki olduğu anlaşılmaktadır. Ardından yaş ve ağırlık arasında 0,36, boy ile fiziksel aktivite sıklığı değeri (FAF) 0.33 oranı ile doğrusal ilişkili olduğu tespit edilmektedir.

Çizelge-4'te yer alan korelasyon matrisinde bağımlı ve bağımsız değişkenlerin korelasyonları takip edilebilir. Buna göre obezite sınıflarını gösteren ve "NObeyesdad" isimli bağımlı değişkenin ağırlık ve ailede fazla kilo öyküsü ile yüksek ilişkili olduğu gözlemlenmiştir.



Şekil 5. Bağımlı Değişken("Nobeyesdad") ile ilişki

Çizelge-4.Korelasyon Matrisi



Şekil-5'te bağımlı değişken olan obezite sınıflarının (NOBayesdad) diğer bağımsız değişkenlerle olan ilişkisi grafik olarak sıralanmaktadır. Ağırlık değişkeni en yüksek korelasyona sahipken kalori tüketimi en az ve ters korelasyona sahiptir.

2.4.Veri Kümesinin Eğitim ve Sınama Veri Kümesine Bölünmesi ve Ölçünlü Biçime Getirilmesi

Makine öğrenmesi modellemelerinin yapılabilmesi ve öğrenme sürecinin gerçekleştirilebilmesi için veri kümesi rasgele seçim yapılarak %80 eğitim seti-%20 sınama kümesi olmak üzere ikiye bölünmüştür. Ayrıca verilerin

değerlendirilmesinde ölçünleme yöntemi kullanılarak veriler minimum ve maksimum değerlerine göre oranlanıp ölçünlü verilere dönüştürülmüştür. Bunun için numpy kütüphanesi kullanılmış ve oranlama gerçekleştirilerek verilerin [0-1] aralığında dağılımı gerçekleştirilmiştir.

2.5. Model Değerlendirme

Makine öğrenmesi sürecinde on bir farklı sınıflandırma algoritması kullanılmıştır. Modellerin karşılaştırılması ve sınıflandırmaların değerlendirilmesi için eğitim keskinlik ölçüsü, doğruluk ölçüsü, karışıklık matrisi sınıflandırma raporu ölçüleri kullanılmıştır. Karışıklık matrisi, sınıflandırma modelinin başarımını özetleyen bir çizelgedir ve modelin farklı sınıflar genelinde hedef değişkeni ne kadar iyi tahmin ettiğine dair kapsamlı bir genel bakış sağlar. Doğruluk ile modelimizin doğru sınıflandırma yüzdesini verir. Ancak tek başına kullanılmaz. Tutturma ile tüm olumlu tahminlerden hangilerinin olumlu olduğuna dair bilgi verir. Bulma ile tüm gerçek değerlerden kaçının pozitif olmasını doğru olarak tahmin edildiği gösterilir. F1 ölçüsü ise bize tutturma ve bulma değerlerinin harmonik ortalamasını göstermektedir

Doğruluk = (TP+TN)/(TP+FP+FN+TN)

Tutturma = TP/(TP+FP)

Bulma = TP/(TP+FN)

F1 Değeri = 2 * kesinlik *duyarlılık / (kesinlik + duyarlılık).

Obezite sınıflarının her biri için sınıflandırma raporu üretilerek en başarılı olan model belirlenmiştir.

2.6. Modeller

Bu çalışmada on bir farklı sınıflandırma algoritması kullanılmıştır. Bunlar sırasıyla Decission Tree (DTC), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Gaussian NB (GNB), SVC, Gradient Boosting (GBC), MLP, Ada Boost (ADA), Extreme Gradient Boost (XGB)'dir.

Karar Ağaçları Sınıflandırma Algoritması (Decission Tree-DTC), denetimli makine öğrenimi uygulamalarında kullanılan ağaç yapısına sahip bir algoritmadır. Karar ağaçları, düğümler, dallar ve yapraklardan oluşur [7]. Basit yapısı sebebiyle karar ağaçları sıkça tercih edilmektedir.

Rasgele Orman Sınıflandırma Algoritması (Random Forest - RF), denetimli makine öğreniminde kullanılan algoritmalardan biri olup verileri sınıflandırmak veya regresyon yapmada kullanılır. Bu algoritma, veri kümesindeki özellikleri kullanarak ağaç yapısını oluşturur ve sınıflandırma veya regresyon görevlerini gerçekleştirir [8].

Lojistik Regresyon Sınıflandırma Algoritması (Logistic Regression-LR) bir lojistik modelin değişkenlerinin tahmin edilmesidir [9]. İkili lojistik regresyonda, bir gösterge değişkeni tarafından kodlanan tek bir ikili bağımlı değişken vardır ve "0" ve "1" olarak etiketlenir, bağımsız değişkenlerin her biri ikili veya sürekli değişken olabilir. Etiketleme, log-oranları olasılığa dönüştüren fonksiyon lojistik fonksiyondur [10].

K-En Yakın Komşuluk (K-Nearest Neighbors - KNN), 1951'de Evelyn Fix ve Joseph Hodges tarafından geliştirilen [11] ve Thomas Cover tarafından ise genişletilen parametrik olmayan denetimli bir öğrenme algoritmasıdır [12]. Sınıflandırma ve regresyon için kullanılır. Her iki durumda da girdi, bir veri kümesindeki en yakın k eğitim örneğinden oluşur. Çıktı, k-NN'nin sınıflandırma veya regresyon için kullanılıp kullanılmamasına bağlıdır.

Doğrusal Ayırma Analizi (Linear Discriminant Analysis - LDA), özniteliklerin doğrusal birleşimini bularak veriyi sınıflara ayırmaya yarayan bir sınıflandırma algoritmasıdır [13]. Doğrusal ayırma analizi, değişkenlerin, veriyi en iyi açıklayan doğrusal birleşiminin incelenmesi bakımından temel bileşen analizi (TBA) ve faktör analizi ile yakından ilişkilidir [14]. 1936 yılında R. A. Fischer tarafından geliştirilen bir sınıflama yöntemidir. Basit olmasına rağmen kompleks problemlerde iyi sonuçlar üreten bir modeldir.

Gauss Naive Bayes Sınıflandırıcı Algoritması (Gaussian NB - GNB), hedef sınıfa göre özelliklerin koşullu olarak bağımsız olduğunu varsayan bir sınıflandırıcıdır. Bu sınıflandırıcılar en basit Bayes ağ modelleri arasındadır. İstatistik literatüründe bu sınıflandırıcı modeli, basit Bayes ve bağımsız Bayes gibi çeşitli isimler altında bilinmektedir [15]. 2006 yılında diğer sınıflandırma algoritmalarıyla yapılan kapsamlı bir karşılaştırma, Bayes sınıflandırmasının diğer yaklaşımlardan daha iyi başarım gösterdiğini ifade etmiştir [16]. Naive Bayes'in bir avantajı, sınıflandırma için gerekli parametreleri tahmin etmek için yalnızca küçük miktarda eğitim verisine ihtiyaç duymasıdır. Gaussian NB'de ise her bir sınıfla ilişkili sürekli değerlerin Gauss dağılımına göre olmasıdır.

Destek Vektör Makineleri Sınıflandırma Algoritması (Suport Vector Machine - SVM), Makine öğreniminde vektör ağlarını da destekleyen destek vektör makineleri, sınıflandırma ve regresyon analizi için verileri analiz eden ilişkili öğrenme algoritmalarına sahip ve denetlenen bir modeldir. AT&T Bell Laboratuvarlarında Vladimir Vapnik ve arkadaşları geliştirilmiştir [17].

Gradyan Artırma Sınıflandırma Modeli (Gradient Boosting - GBC), denetimli öğrenme modelleri için yüksek başarım sağlayan algoritmalardan biridir. Tipik olarak basit karar ağaçları olan, veriler hakkında çok az varsayımda bulunan modeller gibi zayıf tahmin modellerin birleşimi şeklinde tahmin modeli önerir [18]. Bir karar ağacı zayıf öğrenen olduğunda, ortaya çıkan algoritmaya gradyan destekli ağaçlar adı verilir; genellikle rastgele ormandan daha iyi başarım gösterir [19].

Çok Katmanlı Algılayıcı Algoritması(Multi Layer Perceptron - MLP), doğrusal olarak ayrılamayan verileri ayırt edebilmesiyle dikkat çeken, en az üç katman halinde organize edilmiş, doğrusal olmayan bir tür aktivasyon işlevine sahip tamamen bağlı nöronlardan oluşan modern ileri beslemeli yapay sinir ağının adıdır [20].

Adaptif Arttırma Algoritması (Adaptive Boosting - ADA) Yoav Freund ve Robert Schapire tarafından 1995 yılında formüle edilen istatistiksel bir sınıflandırma algoritmasıdır [21].

Gradyan Arttırma Algoritması (Extreme Gradient Boost – XGB) Derin Makine Öğrenimi Topluluğu (DMLC) grubunun bir parçası olarak Tianqi Chen tarafından geliştirilmiştir. XGBoost, özellikle büyük veri setleri ve karmaşık veri yapıları üzerinde çalışırken yüksek başarım ve hız sunar. XGBoost, karar ağaçlarına göre daha yüksek doğruluk sağlaması nedeniyle tercih edilmektedir.

3. Deneysel Sonuçlar ve Tartışma

Yapılan modelleme çalışmasından elde edilen doğruluk değerleri aşağıdaki tabloda görülmektedir:

Çizelge 5. Modellerin Doğruluk Değerleri

Model Adı	Eğitim ve Sınama Süresi (ms)	Doğruluk
DTC	19	0.53
RF	446	0.96
LR	25	0.50
KNN	331	0.89
LDA	23	0.90
GNB	19	0.64
SVC	792	0.57
GBC	2798	0.94
MLP	3218	0.63
ADA	184	0.27
XGB	561	0.96

Çizelge-5'e göre RF ve XGB algoritmaları 0,96 ile en yüksek doğruluk değerini alırken GBC algoritması 0,94 değeri ile üçüncü en yüksek değer olarak sıralanmaktadır. Ayrıca eğitim ve test verileri üzerinde geçen süreler yine çizelgede görülmektedir. Doğruluğu yüksek olan modellere ait süreler 0,5 saniyenin altında gerçekleşmiştir.

RF algoritmasına göre elde edilen sınıflandırma raporu Çizelge-6'da verilmiştir: Test verileri tahmin değerleri %96 doğruluk değerini almıştır. Tabloda ayrıca her obezite sınıfı için elde edilen diğer başarım verileri görülmektedir.

Çizelge-6. RF Algoritması Göre Başarım Değerleri

Obezite Sınıfı	tutturma	bulma	f1-değeri	destek
Yetersiz kilo	0.90	0.97	0.93	62
Normal kilo	1.00	0.96	0.98	56
Aşırı kilolu Düzey 1	0.93	0.89	0.91	56
Aşırı kilolu Düzey 2	0.98	0.96	0.97	50
Obezite Tip 1	0.99	0.97	0.98	78
Obezite Tip 2	0.97	0.98	0.97	58
Obezite Tip 3	1.00	1.00	1.00	63
accuracy			0.96	423
macro avg	0.96	0.96	0.96	423
weighted avg	0.97	0.96	0.96	423

XGB algoritmasına göre elde edilen sınıflandırma raporu aşağıdaki Çizelge-7'de verilmiştir: Çizelgeye göre, obezite sınıflarına ait test verilerinin tahmin değerleri %96 doğruluğa ulaşmıştır. Tabloda ayrıca her obezite sınıfı için elde ettiği diğer başarım verileri görülmektedir.

Çizelge-7. XGB Algoritmasına Göre Başarım Değerleri

Obezite Sınıfı	tutturma	bulma	f1-değeri	destek
Yetersiz kilo	0.95	0.89	0.92	62
Normal kilo	0.93	1.00	0.97	56
Aşırı kilolu Düzey 1	0.91	0.95	0.93	56
Aşırı kilolu Düzey 2	1.00	0.98	0.99	50
Obezite Tip 1	0.97	0.96	0.97	78
Obezite Tip 2	0.97	0.97	0.97	58
Obezite Tip 3	1.00	1.00	1.00	63
accuracy			0.96	423
macro avg	0.96	0.96	0.96	423
weighted avg	0.96	0.96	0.96	423

GBC algoritmasına göre elde edilen sınıflandırma raporu aşağıdaki Çizelge-8'de verilmiştir: Çizelgeye göre, obezite sınıflarına ait test verileri için doğruluk değeri %96'ya ulaşmıştır. Tabloda ayrıca her obezite sınıfı için elde ettiği diğer başarım verileri görülmektedir.

Çizelge-8. GBC Algoritmasına Göre Başarım Değerleri

Obezite Sınıfı	tutturma	bulma	f1-değeri	destek
Yetersiz kilo	0.91	0.85	0.88	62
Normal kilo	0.93	0.96	0.95	56
Aşırı kilolu Düzey 1	0.88	0.91	0.89	56
Aşırı kilolu Düzey 2	0.94	0.96	0.95	50
Obezite Tip 1	0.97	0.92	0.95	78
Obezite Tip 2	0.93	0.98	0.96	58
Obezite Tip 3	1.00	1.00	1.00	63
accuracy			0.94	423
macro avg	0.94	0.94	0.94	423
weighted avg	0.94	0.94	0.94	423

Ancak doğruluk değerlerinin arttırılması amacıyla algoritma giriş değerlerinde (Hiper Parametre) ince ayar yapılarak başarım iyileştirme çalışması yapılmıştır. Makine öğrenimi modellerinde hiper parametreler bulunur. Hiper parametreler, bir makine öğrenimi modelinin belirli bir görev veya veri kümesi için özelleştirilmesine olanak tanıyan seçim veya yapılandırma noktalarıdır.

En iyi parametreleri belirlemek için her tekrarda farklı rasgelelik ile Katmanlı K-Katlama yöntemi uygulanmıştır. Katlama sayısı 5 tekrar sayısı 2 olarak seçilmiştir. Daha sonra verilen hiper parametre rasgele arama algoritması ile taranarak en iyi hiper parametreleri tespit edilmiştir.

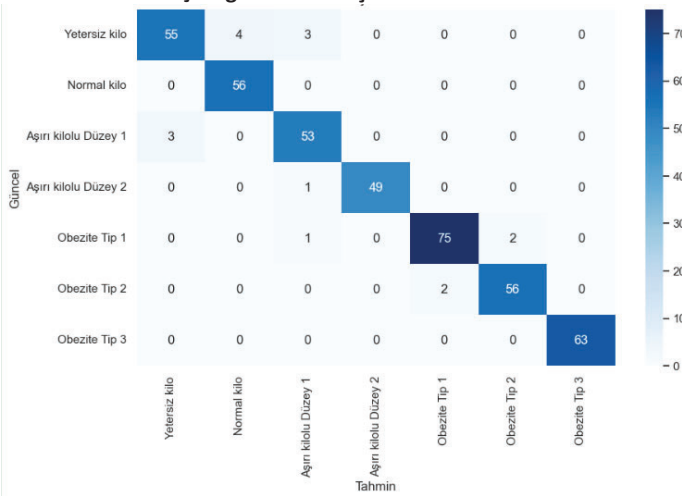
Elde edilen yeni doğruluk değerleri ve ilgili parametreleri aşağıdaki Çizelge-9'da yer almaktadır:

Çizelge-9. En İyi Hiper Parametre Değerleri

Model	Random Forest	Gradient Boost	XG Boost
N Estimator	1000	500	100
Max features	Log2	3	6
Learning rate	0.1	0.1	0.1
Önceki Doğruluk	0.96	0.94	0.96
İyileştirilmiş Doğruluk	0.96	0.97	0.97
Geçen Süre (sn)	29	1466	152

Çizelgeye göre en iyi hiper parametre değerleri hesaplanarak iyileştirilmiş başarım doğruluk değerlerine ulaşılmıştır. Gradient Boost ve XGBoost test verileri üzerinde %97 doğruluk oranını yakalamıştır. Buna göre hesaplama süreleri Çizelge-9'a eklenmiştir. En iyi başarım modeline göre (XGBoost) Çizelge-10. Karmaşıklık Matrisi Çizelge-10'da karmaşıklık matrisi gösterilmektedir. Güncel (gerçek) değerlere göre tahmin değerlerin durumu bu tablodan görülebilir. Küçük hata ile yüksek sayıda sınıf tahmini yapıldığı anlaşılmaktadır.

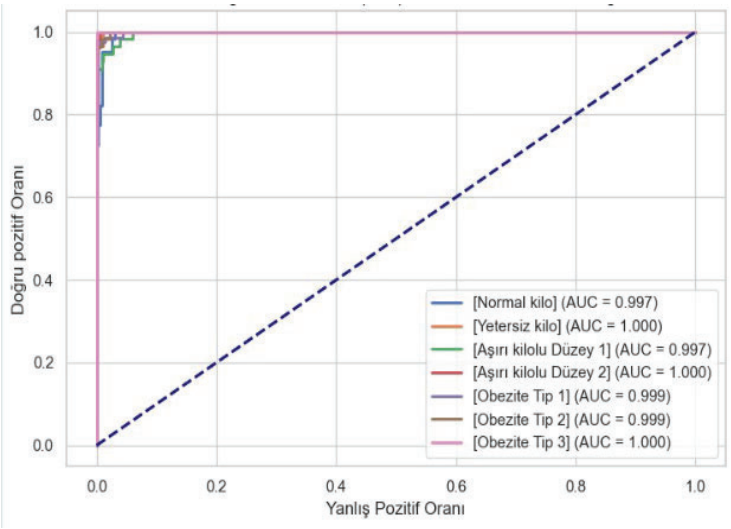
Çizelge-10. Karmaşıklık Matrisi



Aynı şekilde kullanılan modele ait ROC eğrisi Şekil-6'da verilmiştir. Bu eğride çoklu sınıflandırma eğrilerinin hızlı bir şekilde yükselerek "1" değerine yakınsadığı ve yanlış pozitif oranın tüm sınıflar içinde oldukça küçük olduğu görülmektedir.

Çizelge 11'de literatür taramalarında yer alan ve obezite verilerinin makine öğrenmesi modelleri ile bu makalede yer alan model karşılaştırmaları görülmektedir. Tabloda 2111 kayıt ve 17 özellik olarak gösterilen veri kümeleri makaleler incelendiğinde farklı kaynaklardan temin edilse de bu çalışmada kullanılan veri kümesi ile aynı kaynaktan olduğu anlaşılmaktadır.

Tabloya baktığımızda, Turan [22], RF ve KNN modellerini deneyerek %94'lük doğruluk oranına erişse de veri kümesinde korelasyonu düşük değişkenlerin çıkarılması ve kullanılan optimizasyon parametrelerin yeterli düzeyde olmaması bu çalışmaya göre doğruluğun nispeten düşük kalmasına neden olmuştur. Yine, Cuhadar ve ark. [23], yaptığı çalışma yüksek doğruluk oranı elde etse bile günlük kişisel alışkanlıkları kullanan veri kümesinden farklı olarak kişisel kan değerlerine dayalı olması nedeniyle obezite tahmin etme konusuna farklı bir yaklaşım olarak değerlendirilebilir.



Şekil-6. XGBoost ROC Eğrisi

Farklı modellerde elde edilen başarım verilerine göre bu çalışmada Gradient Boost ve XGBoost sınıflandırıcı kullanan iki modelin daha iyi sonuçlar ürettiği saptanmıştır.

4.Sonuç

Bu çalışmada, Kaggle web sitesinden açık kaynak veri kümesi olarak erişilen obezite veri kümesi kullanılmıştır. Veri kümesi çeşitli sınıflandırma algoritmalarıyla eğitim ve test süreçlerinden geçirilmiştir. Veri kümesi rasgele seçilerek ikiye bölünmüş ve elde edilen iki gruptan biri eğitim seti olarak makine öğrenmesinde kullanılmıştır. Diğer grup algoritmanın doğruluğunu görebilmek için test veri kümesi olarak kullanılmıştır. Test veri kümesi çapraz doğrulama modeli ile doğrulanmıştır. On bir farklı sınıflandırma algoritması kullanılarak elde edilen modellerde eğitim ve testler gerçekleştirilmiş ve elde edilen başarım verileri bir tabloda sunularak karşılaştırılmıştır. Tabloda verilen ve başarımı en yüksek üç model tekrar ele alınarak en iyi doğruluk değerini sağlayan parametreler saptanmıştır. İyileştirme sonrası iki model ile elde edilen başarım sonucu literatür taramasında elde edilen diğer başarım sonuçları ile karşılaştırılmıştır. Sonuç olarak çalışmada oluşturulan ve Gradient Boost ve XGBoost sınıflandırıcı kullanan iki modelin test verileri üzerinde %97 doğruluk oranını yakalayarak en yüksek skorları elde ettiği görülmüştür. Bu çalışmayla aynı zamanda birçok çalışmada kısıtlı özelliklerle obezite ilişkisinin çalışıldığı makalelere göre farklı özellikleri de dikkate alan bir veri kümesi olması nedeniyle anlamlı olduğu düşünülmektedir. Farklı özellikler ile obezite sınıf ilişkisine bakılmıştır. Aynı zamanda bu çalışma on bir farklı modelleme ile obezite sınıflandırmalarının makine öğrenmesinin yapılması açısından da anlamlı olduğu düşünülmektedir.

Çizelge 11.Kaynak Araştırma Çalışmaları ile Başarım Karşılaştırması

Çalışma Adı	Veri kümesi	Model	Doğruluk
Five Machine Learning Supervised Algorithms for The Analysis and the Prediction of Obesity [24]	2111 kayıt 17 özellik	RF	91
Estimation of Obesity Levels Based on Decision Trees [25]	2111 kayıt 17 özellik	XGBoost	86
Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique [26]	2111 kayıt 17 özellik	DL	93
OBESEYE: Interpretable Diet Recommender for Obesity Management using Machine Learning and Explainable AI [27]	19-95 yaş arası 146 hasta verileri	LightGBM	86
Using machine learning to predict obesity in high school students [28]	Tennessee eyaletinde 2015 anket verileri	KNN	89
Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People [29]	İngiltere Milenyum Kohort Çalışması verileri	MLP	90
Machine learning approaches for the prediction of obesity using publicly available genetic profiles [30]	164 adet genetik profil veri	SVM	90
Optimize Edilmiş Denetimli Öğrenme Algoritmaları ile Obezite Analizi ve Tahmini [22]	2111 kayıt 17 özellik	RF	94
Deep Learning-Based Prediction Of Obesity Levels Accordıng To Eating Habits And Physical Condition [31]	2111 kayıt 17 özellik	CNN	82
Predictive Analytics: The Application of J48 Algorithm on Grocery Data to Predict Obesity [32]	35 hane halkından elde edilen 8273 adet veri	J48 decision tree	89
Obesity level prediction based on data mining techniques [33]	2111 kayıt 17 özellik	MLP	95
Obesity Prediction Using Ensemble Machine Learning Approaches. [34]	Belirtilmemiş	RF	89
Detection of Obesity Stages Using Machine Learning Algorithms [35]	2111 kayıt 17 özellik	RF	96
Yapılan Çalışma	**2111 kayıt 17 özellik**	**GB, XGB**	**97**

Kaynakça

[1] Khanna D., Peltzer C., Kahar P. ve Parmar, M. S. "Body mass index (BMI): a screening tool analysis," *Cureus,* cilt 14, 2022.

[2] Tan K. C. B. ve Ark., "Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies," *The lancet,* 2004.

[3] Haslam D. W. ve James W. P. T., "Obesity," *The Lancet,* cilt 366, pp. 1197-1209, 2005.

[4] *World Health Organization, World Health Organization,* 2024.

[5] *World obesity day atlases: Obesity Atlas 2024 (no date) World Obesity Federation Global Obesity Observatory,* 2024.

[6] Palechor F. M, Manotas A. *Estimation of obesity levels UCI dataset,* Kaggle, 2021.

[7] Quinlan J. R., "Induction of Decision Trees," *Machine Learning,* cilt 1, pp. 81-106, 1986.

[8] Breiman L., "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[9] Tolles J. ve Meurer W. J., "Logistic Regression: Relating Patient Characteristics to Outcomes," *JAMA,* cilt 316, p. 533, August 2016.

[10] Hosmer D. W., Lemeshow S. ve Sturdivant R. X., "Applied Logistic Regression: Hosmer/Applied Logistic Regression," 2005.

[11] Silverman B. W. ve Jones M. C., "E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)," *International Statistical Review / Revue Internationale de Statistique,* cilt 57, p. 233, December 1989.

[12] Cover T. M. ve Hart P. E., "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory,* cilt 13, pp. 21-27, 1967.

[13] Doğan M., Orman A., Örkcü M. ve Örkcü H., "Çok gruplu sınıflandırma problemlerine regresyon analizi ve matematiksel programlama tabanlı yeni bir yaklaşım," *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi,* June 2019.

[14] Martinez A. M. ve Kak A. C., "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* cilt 23, p. 228–233, 2001.

[15] Hand D. J. ve Yu K., "Idiot's Bayes: Not So Stupid after All?," *International Statistical Review / Revue Internationale de Statistique,* cilt 69, p. 385, December 2001.

[16] Caruana R. ve Niculescu-Mizil A., "An empirical comparison of supervised learning algorithms," %1 içinde *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA, 2006.

[17] Cortes C. ve Vapnik V., "Support-vector networks," *Machine Learning,* cilt 20, p. 273–297, September 1995.

[18] Madeh Piryonesi S. ve El-Diraby T. E., "Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling," *Journal of Infrastructure Systems,* cilt 27, June 2021.

[19] Piryonesi S. M. ve El-Diraby T. E., "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index," *Journal of Infrastructure Systems,* cilt 26, p. 04019036, 2020.

[20] Cybenko G., "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems,* cilt 2, p. 303–314, December 1989.

[21] Freund Y. ve Schapire R. E., "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences,* cilt 55, pp. 119-139, 1997.

[22] Turan T., "Optimize Edilmiş Denetimli Öğrenme Algoritmaları ile Obezite Analizi ve Tahmini," *Mehmet Akif Ersoy Üniversitesi Fen Bilimleri Enstitüsü Dergisi,* cilt 14, p. 301–312, 2023.

[23] Cuhadar S. N., Karaduman G., Uyanik A. ve Durmaz H., "Performance Analysis Of Machine Learning-Based Models For Early Diagnosis Of Obesity Using Blood Test Parameters," *International Journal of Engineering Science and Application,* cilt 7, p. 117–128, 2023.

[24] Kabongo J., Luzolo M., CLEM'S ve D. R. Congo, "Five Machine Learning Supervised Algorithms for The Analysis and the Prediction of Obesity".

[25] Cui T., Chen Y., Wang J., Deng H. ve Huang Y., "Estimation of Obesity Levels Based on Decision Trees," %1 içinde *2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM)*, 2021.

[26] Yagin F. H., Gülü M., Gormez Y., Castañeda-Babarro A., Colak C., Greco G., Fischetti F. ve Cataldi S., "Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique," *Applied Sciences,* cilt 13, p. 3875, March 2023.

[27] Roy M., Das S. ve Protity A. T., *OBESEYE: Interpretable Diet Recommender for Obesity Management using Machine Learning and Explainable AI,* 2023.

[28] Zheng Z. ve Ruggiero K., "Using machine learning to predict obesity in high school students," %1 içinde *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017.

[29] Singh B. ve Tawfik H., "Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People," %1 içinde *Computational Science – ICCS 2020*, Cham, 2020.

[30] Montanez C. A. C., Fergus P., Hussain A., Al-Jumeily D., Abdulaimma B., Hind J. ve Radi N., "Machine learning approaches for the prediction of obesity using publicly available genetic profiles," %1 içinde *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017.

[31] Kıvrak M., "Deep Learnıng-Based Predıctıon Of Obesıty Levels Accordıng To Eatıng Habıts And Physıcal Condıtıon," *The Journal of Cognitive Systems,* cilt 6, p. 24–27, June 2021.

[32] Daud N., Mohd Noor N. L., Aljunid S. A., Noordin N. ve Fahmi Teng N. I. M., "Predictive Analytics: The Application of J48 Algorithm on Grocery Data to Predict Obesity," %1 içinde *2018 IEEE Conference on Big Data and Analytics (ICBDA)*, 2018.

[33] Alqahtani A., Albuainin F., Alrayes R. ve Ark., "Obesity level prediction based on data mining techniques," *International journal of computer science and network security: IJCSNS,* cilt 21, p. 103–111, 2021.

[34] Jindal K., Baliyan N. ve Rana P. S., "Obesity Prediction Using Ensemble Machine Learning Approaches," 2018.

[35] Kitis S. ve Goker H., "Detection of Obesity Stages Using Machine Learning Algorithms," *Anbar Journal of Engineering Sciences,* cilt 14, p. 80–88, April 2023.

[36] Sançar B. ve Özcanarslan F., "Akademisyenlerin Obesite Farkındalıklarının Belirlenmesi: Toros Üniversitesi Örneği," *Uluslararası İktisadi ve İdari Bilimler Dergisi,* cilt 7, December 2021.

[37] Alebna P. L., Mehta A., Yehya A., daSilva-deAbreu A., Lavie C. J. ve Carbone S., "Update on obesity, the obesity paradox, and obesity management in heart failure," *Progress in Cardiovascular Diseases,* cilt 82, pp. 34-42, 2024.