



Kahramanmaraş Sütçü İmam University

Journal of Engineering Sciences



Geliş Tarihi : 04.04.2024
Kabul Tarihi : 19.07.2024

Received Date : 04.04.2024
Accepted Date : 19.07.2024

DİYABET RİSK DURUMUNUN BELİRLENMESİNDE SINIFLANDIRMA ALGORİTMALARININ PERFORMANSLARININ KAPSAMLI BİR ŞEKİLDE KARŞILAŞTIRILMASI

A COMPREHENSIVE COMPARISON OF THE PERFORMANCE OF CLASSIFICATION ALGORITHMS IN DETERMINING DIABETES RISK STATUS

Rukiye UZUN ARSLAN^{1*} (ORCID: 0000-0002-2082-8695)
İrem ŞENYER YAPICI² (ORCID: 0000-0003-0655-340X)
Okan ERKAYMAZ³ (ORCID: 0000-0002-1996-8623)

¹ Zonguldak Bülent Ecevit Üniversitesi, Elektrik-Elektronik Mühendisliği Bölümü, Zonguldak, Türkiye

² Zonguldak Bülent Ecevit Üniversitesi, Bilgisayar Mühendisliği Bölümü, Zonguldak, Türkiye

³ Milli Savunma Üniversitesi, Deniz Harp Okulu, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye

*Sorumlu Yazar / Corresponding Author: Rukiye UZUN ARSLAN, rukiyeuzun67@gmail.com

ÖZET

Diyabet, dünya çapında prevalansı her geçen gün artan metabolik bir halk sağlığı sorunudur. Bu hastalık kontrol altına alınmaması durumunda birçok doku ve organ üzerinde geri dönüşümü olmayan hasarlara neden olabilmektedir. Bu nedenle, diyabet hastalığının erken teşhisi ve etkili bir şekilde yönetilmesi, hastaların yaşam kalitesinin artırılması ve potansiyel sağlık risklerinin azaltılması açısından kritik bir öneme sahiptir. Sağlık sektöründe, hastalıkların teşhisinde yaygın olarak kullanılan makine öğrenimi (MÖ) tabanlı karar destek sistemleri (KDS) önem arz etmektedir. Bu çalışmada, diyabet teşhisi için önerilen MÖ tabanlı bir KDS sunulmuştur. Çalışma kapsamında, veri seti rasgele 80:20 oranında beş kez bölünmüş ve beş farklı MÖ algoritmasıyla (k-en yakın komşu, ridge, aşırı gradyan artırma, ekstra ağaç ve gradyan artırma) performansları değerlendirilmiştir. Bunun için veri setindeki özellikler RO algoritmasıyla değerlendirilerek Ki-kare testine dayalı SelectKBest yöntemiyle en anlamlı özellikler belirlenmiştir. Ayrıca önerilen sistem performansı üzerinde yeniden örnekleme tekniklerinin (sentetik azınlık aşırı örnekleme tekniği, Near Miss) etkileri analiz edilmiştir. Yapılan analizler sonucunda, Near Miss yeniden örnekleme tekniğinin veri setine uygulanmasıyla gradyan artırma algoritmasının en iyi performansı sergilediği tespit edilmiştir. Bu durumda, test verileriyle yapılan analizlere göre F-skor, kesinlik, doğruluk ve duyarlılık değerleri sırasıyla %99.44, %98.89, %99.45 ve %100 olarak hesaplanmıştır.

Anahtar Kelimeler: Diyabet, sınıflandırma, karar destek sistemi, örnekleme teknikleri.

ABSTRACT

Diabetes is a metabolic public health problem with an increasing prevalence worldwide. If untreated, it can cause irreversible effects on many tissues and organs. Therefore, early diagnosis and effective management of diabetes is critical to improve patients' quality of life and reduce potential health risks. In the healthcare industry, machine learning (ML) based decision support systems (DSS) are widely used for disease diagnosis. In this study, a proposed ML-based CDS for diabetes diagnosis is presented. Within the scope of the study, the dataset is randomly split five times in a ratio of 80:20 and the performances of five different ML algorithms (k-nearest neighbor, ridge, extreme gradient boosting, extra tree and gradient boosting) are evaluated. For this purpose, the features in the dataset are evaluated with the RO algorithm and the most significant features are determined by the SelectKBest method based on the Chi-square test. In addition, the effects of resampling techniques (synthetic minority oversampling technique, Near Miss) on the performance of the proposed system were analyzed. As a result of the

ToCite: UZUN ARSLAN, R., ŞENYER YAPICI, İ. & ERKAYMAZ, O., (2024). DİYABET RİSK DURUMUNUN BELİRLENMESİNDE SINIFLANDIRMA ALGORİTMALARININ PERFORMANSLARININ KAPSAMLI BİR ŞEKİLDE KARŞILAŞTIRILMASI. *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi*, 27(4), 1320-1333.

analysis, it was found that the gradient boosting algorithm performed best when the Near Miss resampling technique was applied to the dataset. In this case, the F-score, precision, accuracy and sensitivity values were calculated as 99.44%, 98.89%, 99.45% and 100%, respectively, based on the analysis with the test data.

Keywords: Diabetes, classification, decision support system, sampling techniques.

GİRİŞ

Diyabet dünya çapında prevalansı giderek artan bir halk sağlığı sorunu olup, pankreasın insülin hormonunu yeterince üretememesi veya vücudun üretilen insüline etkili bir şekilde yanıt verememesi sonucu ortaya çıkan kronik bir metabolik hastalıktır. Diyabet, gözlerden böbreklere, kalpten kan damarlarına kadar birçok farklı dokuda kronik hasarlara ve işlevsel bozukluklara neden olabilmektedir (Krasteva vd., 2011). Dünya Sağlık Örgütü'ne göre, beraberinde getirdiği komplikasyonlarla zamanla ölümcül sonuçlar doğurabilen diyabet dünya genelinde önde gelen ölüm nedenleri arasında yer almaktadır (WHO, 2024). Uluslararası Diyabet Federasyonu diyabet hastalığının 2021 yılında 6,7 milyon insanın ölümüne sebep olduğunu açıklamıştır. Bu rakamlar her geçen gün artmakta olup, hastalığın tedavi edilmemesi halinde 2045 yılına kadar diyabetli hasta sayısının 693 milyonu aşacağı tahmin edilmektedir (IDFA, 2022). Diyabetin bu kadar hızlı bir şekilde artması hastalığın önlenmesi, erken teşhisin konulması ve etkili bir şekilde tedavi edilmesi açısından önem taşımaktadır.

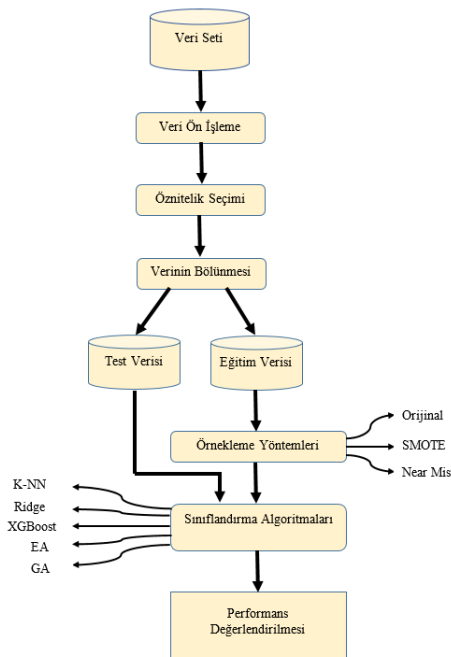
Günümüzde kullanılan teşhis yöntemlerinin zaman alıcı ve bazı durumlarda yeterli olmaması, araştırmacıları yeni arayışlara yönlendirmiştir. Sağlık sektöründe yeni teknolojilerin kullanılmasıyla birlikte yapay zekâ tabanlı yöntemler, diyabet teşhisi ve yönetiminde giderek daha büyük bir paya sahip olmaya başlamıştır. Bu bağlamda literatürde diyabet teşhisinde makine öğrenmesi (MÖ) temelli çok sayıda çalışma bulunmaktadır. Örneğin, Das vd. (2018), bireylerin diyabetik durumunun erken evrede teşhis edilebilmesi için MÖ temelli bir model önermişlerdir. Önerilen modelde çalışma süresi ve tahmin doğruluğu açısından kullanılan sınıflandırma algoritmalarından Navie Bayes (NB)'in J48'e kıyasla daha yüksek performans sergilediği tespit edilmiştir. Akyol ve Şen (2018) diyabet teşhisinde öznitelik seçimi ve sınıflandırma tekniklerinin etkilerini irdelemişlerdir. Bunun için öncelikle PIMA yerlileri veri seti üzerinde özellik seçimi veya ağırlıklılandırma yöntemleri kullanılarak veri setindeki baskın özellikler tespit edilmiş ve üç farklı MÖ algoritmasının performansları karşılaştırılmıştır. Yapılan analizlerle, kararlılık seçimi yöntemi ve Adaboost algoritmasıyla en yüksek başarımın elde edilebileceği gösterilmiştir. Alehegen vd. (2019) iki farklı veri seti kullanarak diyabet teşhisinde rasgele orman (RO), k en yakın komşu (KNN), NB ve J48 algoritmalarının tahminlerini birleştiren bir topluluk öğrenme algoritmasının başarımını irdelemişlerdir. Jakka ve Vakula Rani (2019) diyabet hastalığının erken evrede teşhis edilebilmesi için PIMA yerlileri veri seti üzerinde KNN, RO, lojistik regresyon (LR), karar ağacı (KA) ve destek vektör makinaları (DVM) algoritmalarının başarımlarını değerlendirmişlerdir. Farklı performans metriklerine göre yapılan değerlendirmeler sonucunda LR algoritmasının %77.6'lık bir oranla diyabeti doğru sınıflandırdığı belirlenmiştir. Daghistani ve Alshammari (2020), Suudi Arabistan'daki bir sağlık kuruluşundan elde ettikleri veri setini kullanarak LR ve RO algoritmalarının diyabet teşhisindeki performanslarını değerlendirmişlerdir. Yapılan analizlerde RO algoritmasının en yüksek sınıflandırma doğruluğuna (%88) sahip olduğu belirlenmiştir. Turhan vd. (2020) dengesiz bir veri setinde diyabet hastalığı teşhisi için üç farklı örnekleme tekniği kullanarak topluluk öğrenmesi yöntemlerinin performanslarını analiz etmişlerdir. Analizler sonucunda, sınıf dengesizliği olan veri setlerinde yeniden örnekleme tekniklerinin sınıflandırma algoritmalarına uygulanmasıyla daha yüksek başarımların elde edilebileceğini göstermişlerdir. Shuja vd. (2020), Kaşmir'deki bir laboratuvarından alınan dengesiz bir diyabet veri seti üzerinde sentetik azınlık aşırı örnekleme tekniği (SMOTE) uygulayarak, beş farklı MÖ algoritmasının diyabet teşhisindeki performanslarını karşılaştırmışlardır. Yapılan analizler sonucunda en yüksek başarımlar SMOTE tekniğiyle birlikte kullanılan KA algoritmasıyla elde edilmiştir. Mohammed vd. (2020), gerçek bir diyabet veri seti kullanarak, MÖ algoritmalarının (KNN, KA, NB, LR, DVM ve yapay sinir ağları) diyabet hastalığını sınıflandırma performanslarını irdelemişlerdir. Çalışmalarında dengesiz sınıf problemiyle karşılaştıkları için SMOTE tekniğiyle birlikte, üç farklı normalizasyon tekniğini kullanarak algoritmaların başarımlarını kıyaslamışlardır. Yapılan analizler sonucunda, SMOTE yeniden örnekleme yöntemi ve farklı normalleştirme tekniklerinin sınıflandırma algoritmalarının performansını önemli ölçüde artırdığı ortaya konulmuştur. Mesquita vd. (2021), PIMA yerlileri veri seti üzerinde on farklı MÖ ile altı farklı aşırı örnekleme algoritmalarının diyabet teşhisi üzerindeki performansını irdelemişlerdir. Yapılan benzetim çalışmalarında, tüm kombinasyonlar arasında en iyi sonucun DVM-SMOTE yöntemiyle birlikte çalışan Adaboost algoritmasıyla elde edildiği tespit edilmiştir. Özlüer Başer vd. (2021) bireylerin diyabetik durumunun tespitinde altı farklı MÖ algoritmasının başarımını k-katmanlı çapraz doğrulama yöntemine göre analiz etmişlerdir. Yapılan analizlerde en yüksek başarımların %84.78 ile RO algoritmasıyla ulaşılmıştır. Harman (2021) dengesiz bir diyabet veri seti üzerinde SMOTE tekniği uygulayarak

DVM ve NB algoritmalarının başarımını irdelemiştir. Analizlerde en yüksek sınıflandırma başarımı %88 ile DVM’de elde edilmiştir. Özkan vd. (2022) bireylerin diyabetik durumunun tespitinde iki farklı yaklaşım kullanarak sekiz farklı MÖ algoritmasının başarımlarını karşılaştırmışlardır. Çalışmada diyabet tanısında istatistiksel ve kliniksel açıdan önemli (anlamlı) bulunan özellikleri kullanan modellerin performansları 10-kat çapraz doğrulama tekniğiyle değerlendirilmiştir. Yapılan analizlerde her iki yaklaşım içinde RO algoritmasının diğer algoritmalara kıyasla daha iyi sınıflandırma performansa sahip olduğu tespit edilmiştir. Sevli (2022) dengesiz bir diyabet veri seti kullanarak altı farklı MÖ algoritmasının performanslarını on dört farklı yeniden örnekleme tekniğine göre analiz etmiştir. Örnekleme tekniklerinin sınıflandırma başarımı üzerinde olumlu etkisinin olduğunu ve en yüksek doğruluk değerinin InstanceHardnessThreshold az örnekleme tekniğinin RO (%96,296) algoritmasına uygulanması durumunda elde edildiğini tespit etmiştir. Yılmaz vd. (2023) diyabetin erken evrede tespitine yönelik MÖ tabanlı bir model önermişlerdir. Beş farklı MÖ algoritmasının başarımlarının karşılaştığı çalışmada en yüksek sınıflandırma doğruluğu %96 ile RO’da elde edilmiştir. Özoğur ve Orman (2023) bireylerin diyabetik durumunun tespitinde beş farklı MÖ algoritmasının başarımlarını çeşitli yeniden örnekleme yöntemlerine dayanarak karşılaştırmışlardır. Yapılan analizler sonucunda en yüksek sınıflandırma başarımı SMOTE-ENN ve MICE örnekleme tekniklerinin DVM ile birlikte kullanıldığında ulaşılabildiğini ortaya koymuşlardır. Korkmaz ve Kaplan (2023), diyabet hastalığının teşhisi için MÖ temelli bir model önermişlerdir. Önerilen modelde yedi farklı sınıflandırma algoritmasının başarımı 10-kat çapraz doğrulama yöntemi uygulanarak analiz edilmiştir. Yapılan analizlerde, aşırı gradyan artırma (XGBoost) ve hafif gradyan artırma (LightGBM) algoritmalarının en yüksek doğruluk değerini (%90.01) verdiği tespit edilmiştir.

Yukarıda bahsedilen çalışmalar göz önüne alındığında, literatürde açık erişimli diyabet veri setlerinin kullanıldığı görülmektedir. Bu çalışmada ise, bireylerin diyabetik risk durumunun tespiti için güncel bir halka açık veri seti kullanılarak MÖ algoritmalarının başarımları analiz edilmiştir. Algoritmaların başarımları farklı performans metriklerine göre yeniden örnekleme tekniklerinin kullanıldığı ve kullanılmadığı durumlar için karşılaştırılarak yorumlanmıştır.

MATERYAL VE METOT

Gerçekleştirilen bu çalışmada farklı MÖ algoritmaları kullanılarak bireylerin diyabetik durumunun tespit edilmesi amaçlanmıştır. Bu bağlamda, Pakistan’ın iki ayrı bölgesinden toplanan ve açık erişimli bir veri tabanında bulunan güncel bir veri seti kullanmıştır (Kaggle, 2024). Çalışmada diyabet tanısında kullanılan ölçümlerden elde edilen verileri içeren veri setinde, K-NN, Ridge, XGBoost, ekstra ağaçlar (EA) ve gradyan artırma (GA) algoritmalarının sınıflandırma başarımları farklı performans metriklerine göre analiz edilmiştir. Yanı sıra örnekleme tekniklerinin algoritmaların performansları üzerindeki etkileri ele alınmıştır. Önerilen modelin akış şeması Şekil 1’de verilmiştir.



Şekil 1. Veri Setinden Örnek Bir Kesit

Veri Ön İşleme

Bu çalışmada Kaggle veri tabanında “Pakistani Diabetes Dataset” başlığı altında verilen açık erişimli bir veri seti kullanılmıştır (Kaggle, 2024). 912 hastanın tıbbi kayıtlarını içeren veri seti, hastanın diyabet durumunu tahmin etmeye olanak tanıyan bir dizi özellik içermektedir. Veri setindeki her bir örnek, 18 adet giriş özelliği (öznitelik) ve bir teşhis sınıfıyla (çıkış) temsil edilmektedir. Teşhis sınıfının değeri, hastanın diyabetik olup olmasına bağlı olan ikili (binary) bir değere sahiptir. Teşhis sınıfının değerinin 0 olması hastanın sağlıklı (diyabetik olmayan), 1 olması ise hastanın sağlıklı (diyabetik) olduğunu ifade etmektedir. Veri seti 486 diyabetik, 426 sağlıklı hasta kaydı içermektedir. Şekil 2’de veri setinden rasgele seçilen 10 satır, 19 öznitelik ve bunlara karşılık gelen değerler gösterilmiştir.

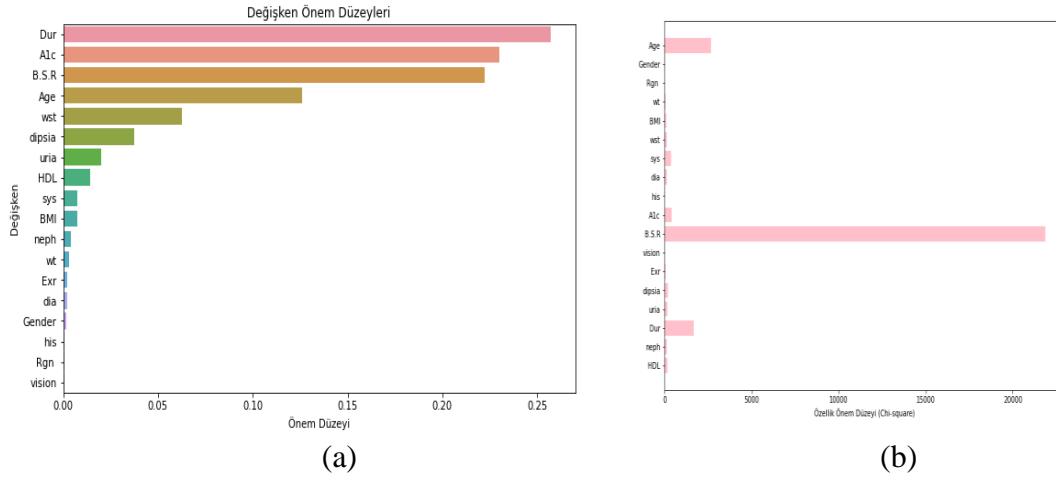
Age	Gender	Rgn	wt	BMI	wst	svs	dia	his	A1c	B.S.R	vision	Exr	dipsia	uria	Dur	neph	HDL	Outcome
24	0	0	84	26.57	34	120	80	0	5.4	123	0	30	0	0	0	0	57	0
28	0	1	58	20.63	35	130	90	0	5	92	0	20	0	0	0	0	46	0
34	0	0	70	20.92	32	110	60	1	5	98	0	35	0	0	0	0	42	0
35	1	0	67	27.9	37	137	92	0	6.8	139	1	30	1	1	1	0	39	1
35	1	0	56	26.71	38	140	90	1	6.5	97	0	20	1	0	5	0	56	1
37	1	0	59	24.6	40.5	130	78	1	9.1	253	0	20	1	0	4	0	48	1
38	1	0	82	26.5	42.5	110	80	0	11.5	238	0	0	1	0	4	1	42	1
39	1	0	83	28.3	34.5	120	110	0	5.1	174	1	0	1	1	2	1	42	1
41	1	0	52	21.66	34	124	75	1	8	201	0	0	1	1	2	1	50	1
54	1	1	81	31.63	39	150	84	0	6.7	112	1	20	0	1	8	1	48	1

Şekil 2. Veri Setinden Örnek Bir Kesit

MÖ temelli karar destek sistemlerinin geliştirilmesinde, verinin hazırlanması son derece önemlidir. Veri ön işleme süreci, veri setini modelin gereksinimlerine uygun hale getirmek için eksik verilerin tamamlanması, veri temizleme, normalleştirme ve boyut indirgeme gibi bir dizi işlemi içermektedir. Bu süreçler sayesinde, MÖ temelli modellerin güvenilir, doğru ve başarılı tahminler yapabilmesi için veri uygun hale dönüştürülmektedir. Bu bağlamda gerçekleştirilen çalışmada veri setindeki özniteliklerden önerilen modele katkısı olmayacak öznitelikler elimine edilmiştir. Bunun için öncelikle veri setindeki özniteliklerin önem düzeyini değerlendirmek için RO algoritması kullanılmıştır. Bu değerlendirme sonucunda, en anlamlı öznitelikleri tespit etmek için Ki-kare testine dayalı SelectKBest yöntemi uygulanmıştır (Şekil 3). RO algoritmasına göre belirlenen dört özellik arasından Ki-kare testine dayalı SelectKBest yöntemiyle çıkarılan üç özellik, hedef değişkenle istatistiksel olarak anlamlı bir ilişki göstermektedir. Bu özellikler, modelin hedef değişkeni tahmin etme yeteneğini geliştirmek adına seçilmiştir. RO algoritması, veri setindeki karmaşıklığı ele alarak önemli özellikleri belirlemede etkili olmuştur. Ki-kare testine dayalı SelectKBest yöntemi ise seçilen özelliklerin hedef değişkenle olan ilişkisini doğrulamak için istatistiksel bir yaklaşım sunmuştur. Bu bulgular ışığında, veri setindeki tüm öznitelikler arasında kan şekeri seviyesi değer aralığı, yaş ve zaman en anlamlı özellikler olarak tespit edilmiştir.

Örnekleme Algoritmaları

Sınıf dengesizliği, bir veri setindeki farklı sınıflara ait örneklerin dağılımında belirgin bir dengesizlik olduğu durumu ifade etmektedir. Bu durum, özellikle MÖ temelli modellerin eğitilmesi ve değerlendirilmesi sürecinde önemli bir sorun olabilmektedir. Dengesiz sınıflar, modelin az örnekliliği doğru bir şekilde öğrenememesine ve bu sınıfları yanlış sınıflandırmasına neden olabilmektedir. Bu sorunu çözmek için kullanılan teknikler arasında, örnekleme yöntemleri (örneğin, aşırı örnekleme ve azaltılmış örnekleme), sınıf ağırlıklarının dengelenmesi, sentetik veri üretimi gibi teknikler yer almaktadır. Bu teknikler, sınıf dengesizliğinin etkilerini azaltarak model performansını artırmayı amaçlamaktadır (Yavaş vd., 2020). Gerçekleştirilen bu çalışmada modellerin performansı üzerinde Near Miss az örnekleme tekniği ile SMOTE aşırı örnekleme tekniğinin etkileri irdelenmiştir.



Şekil 3. Veri Setinden Anlamlı Özelliklerin Çıkarılması a)RO Algoritmasına Göre Özelliklerin Önem Düzeyi, b) Ki-Kare Testine Dayalı SelectKBest Yöntemine Göre Çıkarılan En Anlamlı Özellikler

SMOTE yöntemi, bir veri kümesindeki azınlık sınıfının temsil edilme düzeyini artırmak için kullanılan bir tekniktir. Bu yöntemde, azınlık sınıfındaki her örnek için k en yakın komşusuna bakılarak, interpolasyon tekniğine dayalı rasgele sentetik örnekler oluşturulmaktadır. Bu yeni örnekler, sınıflandırıcının daha büyük ve daha az spesifik karar bölgeleri oluşturmasını sağlamaktadır. Böylelikle, daha dengeli bir sınıf dağılımı elde edilirken, aşırı örneklemeyle ilişkili sorunlar azaltılmış olmaktadır (Öztürk, 2022; Dal vd., 2021).

Near Miss örnekleme yöntemi, bir veri kümesinde azınlık sınıfına ait örneklerle çoğunluk sınıfına ait örnekler arasındaki uzaklıkları dikkate alarak, örneklerin seçilmesini sağlamaktadır. Bu yöntemin amacı, azınlık sınıfına ait örnekleri seçerek çoğunluk sınıfına ait örneklerle olan uzaklıklarını minimize etmektir. Bu şekilde, sınıf dengesizliği azaltılarak daha dengeli bir veri seti elde edilmesi hedeflenmektedir. Near Miss, sınıf dengesizliği problemlerini çözmek için SMOTE gibi diğer yöntemlerle birlikte kullanılabilir. Ancak, yakınlık tabanlı bir seçim kriterine dayandığı için SMOTE'den farklı bir yaklaşımı benimsemektedir (Mqadi vd., 2021).

Sınıflandırma Algoritmaları

Son yıllarda büyük boyutlu verilerin otomatik olarak analiz edilmesi gereksinimi, MÖ algoritmalarına dayalı yöntemlerin kullanımının yaygınlığının artmasına neden olmuştur. Bu bağlamda, gerçekleştirilen çalışmada literatürde diyabet teşhisinde başarılı sonuçlar verdiği bilinen KNN, Ridge, GA, EA ve XGBoost algoritmalarından yararlanılmış olup, algoritmalarının temel prensipleri aşağıda sunulmuştur.

K en yakın komşu (K Nearest Neighbour, KNN) algoritması, belirli bir örneğin sınıfını tahmin etmek için komşu veri noktalarının sınıf etiketlerine bakmaktadır. Bu algoritma, yeni bir veri noktasının sınıfını belirlemek için ona en yakın k komşuyu bulmakta ve bu komşuların sınıf etiketlerinin çoğunluğunu kullanarak tahmin yapmaktadır. KNN, basit ve esnek bir algoritma olmasına rağmen, büyük veri setlerinde ve yüksek boyutlu öznitelik uzaylarında hesaplama maliyeti artabilmektedir. Ayrıca, bu algoritmada k değerinin seçimi önemlidir; küçük k değerleri aşırı uyuma neden olabilirken, büyük k değerleri ise modelin genelleme yeteneğini azaltabilmektedir (Hacıbeyoğlu vd., 2023; Yakar vd., 2024).

Ridge algoritması, lineer regresyon yöntemine dayalı bir sınıflandırma algoritmasıdır. Temel prensibi, gürültülü veya çoklu korelasyonlu özniteliklerle başa çıkmak için lineer regresyonu düzenlemektir. Bu algoritma, modelin karmaşıklığını kontrol etmek için bir düzenleme terimi kullanmakta ve bu sayede aşırı uyumun önüne geçmektedir. Genellikle, düzenleme parametresi olan lambda (λ) değeriyle kontrol edilmektedir. λ değeri arttıkça, modelin karmaşıklığı azalmakta ve genelleme yeteneği artmaktadır. Ridge sınıflandırma algoritması, özellikle yüksek boyutlu veri setleri veya çoklu özelliklerin bulunduğu durumlarda etkili bir şekilde kullanılabilir (Turan, 2023).

Ekstra ağaçlar (Extra Trees, EA) KA tabanlı bir topluluk öğrenme algoritmasıdır. Sınıflandırma kararının çoğunluğa göre verildiği bu algortmada, tipik KA farklı bölünme noktaları rasgele seçilmektedir. Bu rasgele seçim, her bir ağacın bölünme noktalarının farklı özelliklerde ve değerlerde olmasını sağlamaktadır. Böylece

ağaçlar arasında daha fazla çeşitlilik oluşturarak, ağaçların daha yüksek genelleme yeteneğine sahip olmasını ve aşırı öğrenmeye karşı dirençli olmasını sağlamaktadır. Ayrıca, bu algoritmada performansı artırmak için topluluktaki ağaç sayısı artırılabilir. Bu nedenle de EA algoritması, veriye dayalı tahminlerde daha geniş bir çeşitlilik ve daha iyi bir performans sunmaktadır (Türkmenoğlu ve Yıldız, 2021).

Gradyan arttırma (Gradient Boosting, GA), KA gibi zayıf tahmin edicilerini birleştirerek daha güçlü bir tahmin edicinin oluşturulmasını amaçlayan bir topluluk öğrenme algoritmasıdır. Temel prensibi, ardışık öğrenme adımlarıyla önceki tahmin edicilerin hatalarını minimize etmek için yeni tahmin edicilerin eklenmesine dayanmaktadır. Böylelikle her bir adımda önceki modellerin hataları düzeltilerek önerilen modelin genel performansı artırılmaktadır (Sungur ve Bakır, 2024).

XGBoost, GA'nın daha gelişmiş bir versiyonu olarak kabul edilen ağaç tabanlı bir topluluk öğrenme algoritmasıdır. Bu algoritma hesaplama hızı, ölçeklenebilirlik ve performans açısından GA'den daha üstün bir yapıya sahiptir. Ayrıca, aşırı öğrenmeye karşı daha dirençli olması ve daha iyi genelleme yeteneğine sahip olması, onu çeşitli MÖ problemlerinde tercih edilen bir seçenek haline getirmektedir (Yakut vd., 2023).

Performans Ölçütü

Çalışmada kullanılan modelin performansının değerlendirilmesi için doğruluk, duyarlılık, kesinlik ve F- skor metrikleri kullanılmıştır. Bu metriklerin hesaplanmasında gerçek değerlerle tahmin edilen değerler arasındaki ilişkiyi sunan karmaşıklık matrisinden faydalanılmıştır. Karmaşıklık matrisinde yer alan TP, TN, FN ve FP değerleri sırasıyla Doğru Pozitif (True Positive (TP)), Doğru Negatif (True Negative (TN)), Yanlış Negatif (False Negative (FN)) ve Yanlış Pozitif (False Positive (FP)) temsil etmektedir. Metriklerin hesaplanmasında kullanılan denklemler ve tanımları Tablo 1'de verilmiştir.

Tablo 1. Performans Metriklerinin Matematiksel Denklemleri Ve Tanımları

Metrik Adı	Formülü	Tanımı
Doğruluk	$\frac{TP + TN}{TP + TN + FP + FN}$	Doğruluk, tahmin edilen örnek sayısının bütün örnek sayısına olan oranıdır. Bu değer 0 ila 1 arasında değer almakta olup, 1'e yakın olması modelin başarılı olduğunu ifade edilmektedir.
Duyarlılık	$\frac{TP}{FN + TP}$	Duyarlılık, gerçek pozitif örneklerin ne kadarının doğru bir şekilde belirlenebildiğini ölçmektedir.
Kesinlik	$\frac{TP}{FP + TP}$	Kesinlik, modelin pozitif olarak belirlediği örneklerin aslında ne kadarının doğru pozitif olduğunu ölçmektedir.
F-Skor	$\frac{2 \times \text{kesinlik} \times \text{duyarlılık}}{\text{kesinlik} + \text{duyarlılık}}$	F-skor modelin genel doğruluğunu ölçmekte olup, yanlış tahmin edilmiş örnek oranını ölçmektedir.

SONUÇLAR

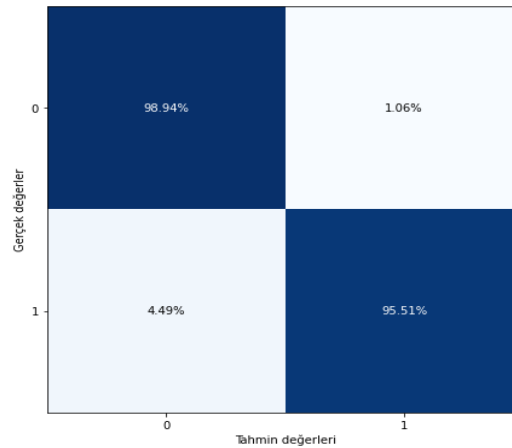
Gerçekleştirilen çalışmada daha az özellikle diyabet hastalığını tahmin edebilen MÖ tabanlı bir karar destek sistemi önerilmiştir. Önerilen sistem için, ilk olarak veri setindeki en anlamlı özellikler belirlenmiştir. Bu amaçla, RO algoritmasıyla özelliklerin önem düzeyi değerlendirilmiş ve ardından Ki-kare testine dayalı SelectKBest yöntemiyle en anlamlı özellikler seçilmiştir. Bu değerlendirme sonucunda, kan şekeri seviyesi değer aralığı, yaş ve zaman özellikleri diyabet tahmini için en belirleyici özellikler olarak belirlenmiştir. Önerilen sistemin performansını değerlendirmek için, orijinal veri seti %80 eğitim ve %20 test olarak rasgele ayrıştırılmıştır. Eğitim için ayrılan veri seti sistemin eğitiminde kullanılırken, test veri seti ise sistemin performansının objektif bir şekilde değerlendirilmesinde kullanılmıştır. Daha sonra veri seti üzerinde literatürde diyabet tahmininde sıklıkla kullanılan MÖ algoritmalarından olan K-NN, Ridge, Xgboost, GA ve EA algoritmaları ve her bir algoritma için farklı örnekleme teknikleri (Near Miss ve SMOTE) uygulanması durumunda başarımları değerlendirilmiştir. Her bir

sınıflandırıcı başarımı, örnekleme tekniklerinin uygulanıp uygulanmaması durumları için ayrı ayrı değerlendirilerek, karmaşıklık matrisleriyle birlikte sunulmuştur. Çalışma kapsamında yapılan bütün analizler, Phyton programlama dilinde gerçekleştirilmiştir. Bununla birlikte analizlerin güvenilirliğini sağlamak ve sistemin genel performansını objektif bir şekilde değerlendirmek için veri setinin rasgele bölünme işlemi beş kez tekrar edilmiştir. Her bir 80:20 bölme işlemi sonrasında elde edilen sonuçların ortalaması alınmış ve önerilen modelin performansı bu ortalama üzerinden değerlendirilmiştir. Bu doğrultuda örnekleme olmadan (orijinal) ve örnekleme tekniklerinin uygulanması durumlarında KNN algoritmasının performans metrikleri Tablo 2’de verilmiştir. F-skor, kesinlik, duyarlılık ve doğruluk değerleri, hem eğitim hem de test verileri için ayrı ayrı sunulmuştur.

Tablo 2. KNN Algoritması İçin Sınıflandırma Sonuçları

		Orijinal (%)	SMOTE (%)	Near Miss (%)
F-Skor	Eğitim	98,34	98,21	98,00
	Test	97,14	97,14	97,14
Kesinlik	Eğitim	100	100	100
	Test	98,84	98,84	98,84
Duyarlılık	Eğitim	96,73	96,47	96,08
	Test	95,51	95,51	95,51
Doğruluk	Eğitim	98,22	98,24	98,04
	Test	97,27	97,27	97,27

Tablo 2’den açıkça görüldüğü gibi, örnekleme tekniklerinin uygulanmasının KNN sınıflandırıcısının performansı üzerinde belirgin bir etkisi olmamıştır. F-skor metriği, eğitim verileri için örnekleme yapılmadığında (orijinal veri seti) %98.34, SMOTE fazla örnekleme tekniği uygulandığında %98.21 ve Near Miss az örnekleme tekniği uygulandığında %98 olarak belirlenmiştir. Test verileri üzerinden yapılan analizlerde ise F-skor değeri her üç durumda için %97 olarak hesaplanmıştır. Kesinlik metriği eğitim ve test verileri için her üç durumda da sırasıyla %100 ve %98.84 olarak elde edilmiştir. Duyarlılık metriği ise eğitim verileri için orijinal veri setinde %96.73, SMOTE uygulandığında %96.47 ve Near Miss uygulandığında %96.73 olarak hesaplanmıştır. Test verileri üzerinde ise her üç durum için duyarlılık %95.51 olarak ölçülmüştür. Doğruluk metriği açısından eğitim verileri için orijinal veri setiyle %98.22’lik, SMOTE tekniğiyle %98.24’lük ve Near Miss tekniğiyle ise %98.04’lük başarımlara ulaşılmıştır. Test verileri üzerinde ise her bir durum için %97.27’lik doğruluk oranı elde edilmiştir. Genel olarak, her bir durum için sınıflandırma performansları test verileri üzerinde benzerlik gösterirken, eğitim verileri üzerinde küçük farklılıklara neden olmuştur. Şekil 4’de test verileri için KNN algoritmasında en yüksek sınıflandırma başarımının (%97.27) elde edildiği duruma ait karmaşıklık matrisi sunulmuştur.

**Şekil 4.** Test Verileri Üzerinden En Yüksek Başarımla Elde Edilen KNN Algoritmasına Ait Karmaşıklık Matrisi

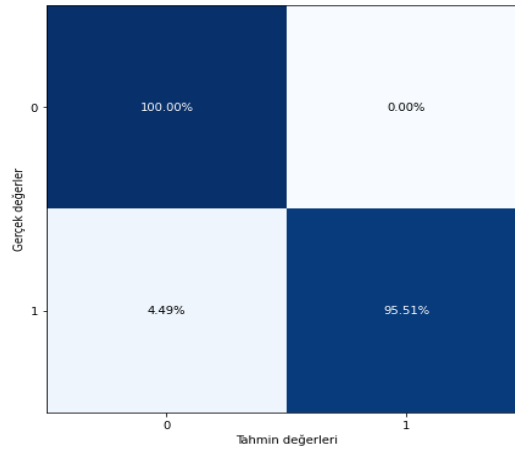
Şekil 4’ten açıkça görüldüğü gibi, KNN algoritması kullanan model, örnekleme tekniklerinin uygulanıp uygulanmamasına bakılmaksızın sağlıklı bireyleri %95.51 doğrulukla sağlıklı olarak sınıflandırmıştır. Ancak, %1.06 oranında yanlış sınıflandırma yapılmıştır. Diyabetli bireyleri ise %95.51 oranında doğru, %4.49 oranında yanlış sınıflandırmıştır. Bu sonuçlar, modelin genel performansının yüksek olduğunu ve çoğu durumda doğru sınıflandırma yaptığını göstermektedir.

Tablo 3’de orijinal veri seti ile SMOTE ve Near Miss örnekleme tekniklerinin uygulanması durumlarında Ridge algoritmasının performans metrikleri eğitim ve test verileri için ayrı ayrı verilmiştir.

Tablo 3. Ridge Algoritması İçin Sınıflandırma Sonuçları

		Orijinal (%)	SMOTE(%)	Near Miss(%)
F-Skor	Eğitim	95,09	95,08	96,15
	Test	96,05	96,59	97,70
Kesinlik	Eğitim	94,97	95,20	98,42
	Test	96,59	97,70	100
Duyarlılık	Eğitim	95,21	94,96	93,98
	Test	95,51	95,51	95,51
Doğruluk	Eğitim	94,65	95,09	96,23
	Test	96,17	96,72	97,81

Tablo 3’den açıkça görüldüğü üzere F-skor metriği eğitim verileri üzerinde örnekleme yapılmadığında %95.09, SMOTE tekniği uygulandığında %95.08 ve Near Miss tekniği uygulandığında %96.15 olarak belirlenmiştir. Test verileri içinse bu değerler sırasıyla %96.05, %96.59 ve %97.70 olarak hesaplanmıştır. Kesinlik metriği, eğitim verilerine örnekleme yapılmadığında %94.97, SMOTE tekniğinde %95.20 ve Near Miss tekniğinde %98.42 olarak elde edilmiştir. Test verileri içinse bu değerler sırasıyla %96.59, %97.70 ve %100 olarak bulunmuştur. Duyarlılık metriğine bakıldığında, eğitim verilerine örnekleme yapılmadığında %95.21, SMOTE uygulandığında %94.96, Near Miss uygulandığında ise %93.98 oranlarına ulaşılmıştır. Test verileri üzerinde ise her üç durum için duyarlılık %95.51 olarak ölçülmüştür. Doğruluk metriği açısından ise eğitim verileri üzerinden orijinal veri setiyle %94.65’lik, SMOTE tekniğiyle %95.09’luk, Near Miss tekniğiyle ise %96.23’lük başarımlara ulaşılmıştır. Test verileri içinse bu değerler sırasıyla %96.17, %96.72 ve %97.81 olarak bulunmuştur. Sonuç olarak Ridge algoritmasında test verileri üzerinden en yüksek başarımla (%97.81) Near Miss tekniğinin uygulanması durumunda gözlemlenmiştir. Bu durumda elde edilen karmaşıklık matrisi Şekil 5’de verilmiştir.



Şekil 5. Test Verileri Üzerinden En Yüksek Başarımla Elde Edilen Ridge Algoritmasına Ait Karmaşıklık Matrisi

Şekil 5’de sunulan verilere göre, Near Miss az örnekleme tekniği uygulanması durumunda Ridge algoritması kullanılan modelde, sağlıklı bireyler için %100, diyabetli bireyler için %95.5 doğruluk oranları elde edilmiştir. Ayrıca, sağlıklı bireyler için yanlış sınıflandırma yapmayan modelde, diyabetli bireylerin %4.49’u yanlışlıkla sağlıklı olarak sınıflandırılmıştır. Bu bulgular, modelin sağlıklı bireyleri sınıflandırmada çok iyi bir performans gösterdiğini ortaya koymaktadır.

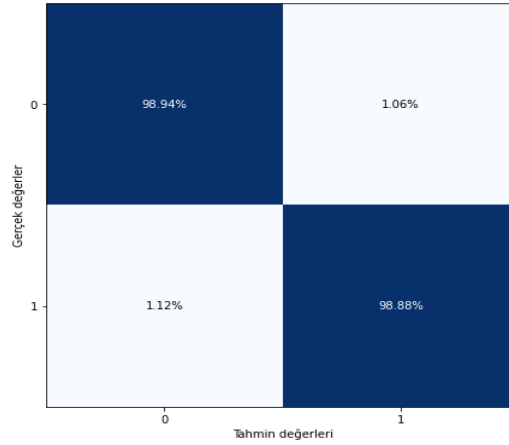
Tablo 4’de veri setine örnekleme tekniklerinin uygulanıp uygulanmama durumları için XGBoost algoritmasının sınıflandırma başarımları gösterilmiştir.

Tablo 4’den açıkça görüldüğü üzere, eğitim verileri için F-skor metriği örnekleme ve fazla örnekleme durumlarında %99.62, az örnekleme durumunda ise %99.55 olarak hesaplanmıştır. Kesinlik metriği örnekleme ve fazla örnekleme durumlarında %99.75, az örnekleme durumunda %99.70 olarak elde edilmiştir.

Tablo 4. XGBoost Algoritması İçin Sınıflandırma Sonuçları

		Orijinal (%)	SMOTE (%)	Near Miss (%)
F-Skor	Eğitim	99,62	99,62	99,55
	Test	98,88	98,88	98,88
Kesinlik	Eğitim	99,75	99,75	99,70
	Test	98,88	98,88	98,88
Duyarlılık	Eğitim	99,50	99,50	99,40
	Test	98,88	98,88	98,88
Doğruluk	Eğitim	99,59	99,62	99,55
	Test	98,91	98,91	98,91

Duyarlılık metriği ise örnekleme ve fazla örnekleme durumlarında %99.50, az örnekleme durumunda %99.40 olarak bulunmuştur. Doğruluk metriği açısından, orijinal veri setiyle %99.59, SMOTE tekniğiyle %99.62 ve Near Miss tekniğiyle %99.55 doğruluk oranlarına ulaşılmıştır. Test verileri üzerinden ise F-skor, kesinlik ve duyarlılık metrikleri örnekleme tekniğinden bağımsız olarak %98.88, doğruluk metriği ise %98.91 olarak hesaplanmıştır. Bu bulgular ışığında, XGBoost algoritmasının hem eğitim hem de test verilerinde yüksek ve tutarlı performans sergilediği, örnekleme tekniklerinin bu performansı önemli ölçüde etkilemediği tespit edilmiştir. Test verileri üzerinden XGBoost algoritmasında en yüksek sınıflandırma başarımının (%98.91) elde edildiği duruma ait karmaşıklık matrisi Şekil 6'da sunulmuştur.

**Şekil 6.** Test Verileri Üzerinden En Yüksek Başarım Elde Edilen Xgboost Algoritmasına Ait Karmaşıklık Matrisi

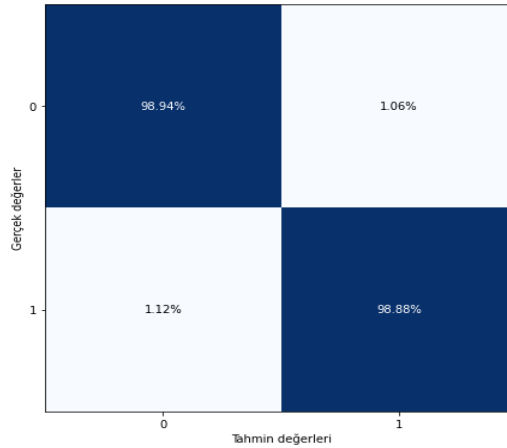
Şekil 6'dan açıkça görüldüğü gibi, örnekleme tekniğinin uygulanma durumundan bağımsız olarak XGBoost algoritması kullanan model, sağlıklı bireyleri %98.94'lük, diyabetli bireyleri %98.88'lik doğruluk oranıyla doğru bir şekilde sınıflandırmıştır. Diyabetli bireyleri yanlışlıkla sağlıklı birey olarak sınıflandırma oranı %1.12 iken, sağlıklı bireyleri yanlışlıkla diyabetli birey olarak sınıflandırma oranı ise %1.06 olarak hesaplanmıştır. Bu bulgular, modelin genel olarak yüksek performans sergileyerek çoğu durumda doğru sınıflandırma gerçekleştirdiğini ortaya koymaktadır.

Tablo 5'de veri setine örnekleme tekniklerinin uygulanıp uygulanmama durumları için EA algoritmasının sınıflandırma başarımleri gösterilmiştir.

Tablo 5. EA Algoritması İçin Sınıflandırma Sonuçları

		Orijinal (%)	SMOTE (%)	Near Miss (%)
F-Skor	Eğitim	99,75	99,75	99,70
	Test	98,88	98,88	98,88
Kesinlik	Eğitim	100	100	100
	Test	98,88	98,88	98,88
Duyarlılık	Eğitim	99,50	99,50	99,40
	Test	98,88	98,88	98,88
Doğruluk	Eğitim	99,73	99,75	99,70
	Test	98,91	98,91	98,91

Tablo 5’de verilen eğitim verilerine ait sonuçlardan, F-skor metriğinin örnekleme ve fazla örnekleme durumlarında %99.75, az örnekleme durumunda ise %99.70 olarak hesaplandığı görülmektedir. Kesinlik metriği tüm durumlarda %100 olarak sabit kalmıştır. Bununla birlikte her bir durum için duyarlılık metriği sırasıyla %99.50, %99.50 ve %99.40 elde edilirken, doğruluk metriği sırasıyla %99.73, %99.75 ve %99.70 olarak hesaplanmıştır. Test verilerine ait sonuçlar incelendiğinde ise örnekleme tekniklerinin uygulanma durumuna bakılmaksızın F-skor, kesinlik ve duyarlılık metrikleri %98.88 olarak, doğruluk metriği ise %98.91 olarak elde edilmiştir. Bu bulgular, EA algoritmasının hem eğitim hem de test verilerinde yüksek performans sergilediğini ortaya koymaktadır. Buna göre test verileri üzerinden EA algoritmasında en yüksek sınıflandırma başarımının (%98.91) elde edildiği duruma ait karmaşıklık matrisi Şekil 7’de verilmiştir.



Şekil 7. Test Verileri Üzerinden En Yüksek Başarım Elde Edilen EA Algoritmasına Ait Karmaşıklık Matrisi

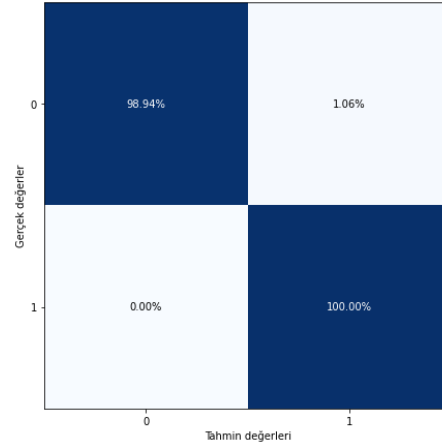
Şekil 7’den açıkça görüldüğü üzere, örnekleme tekniğinin uygulanma durumuna bakılmaksızın EA algoritması kullanan modelin sağlıklı bireyleri doğru sınıflandırma oranı %98.94, yanlış sınıflandırma oranı ise %1.06 olarak belirlenmiştir. Diyabetli bireyler için ise doğru sınıflandırılma oranı %98.88, yanlış sınıflandırma oranı ise %1.12 olarak hesaplanmıştır. Bu durum modelin genellikle doğru sınıflandırma yaptığını göstermektedir.

Tablo 6’da veri setine örnekleme tekniklerinin uygulanıp uygulanmama durumları için GA algoritmasının sınıflandırma başarım metrikleri gösterilmiştir.

Tablo 6. GA Algoritması İçin Sınıflandırma Sonuçları

		Orijinal (%)	SMOTE (%)	Near Miss (%)
F-Skor	Eğitim	99,75	99,75	99,70
	Test	98,88	98,88	99,44
Kesinlik	Eğitim	100	100	100
	Test	98,88	98,88	98,89
Duyarlılık	Eğitim	99,50	99,50	99,40
	Test	98,88	98,88	100
Doğruluk	Eğitim	99,73	99,75	99,70
	Test	98,91	98,91	99,45

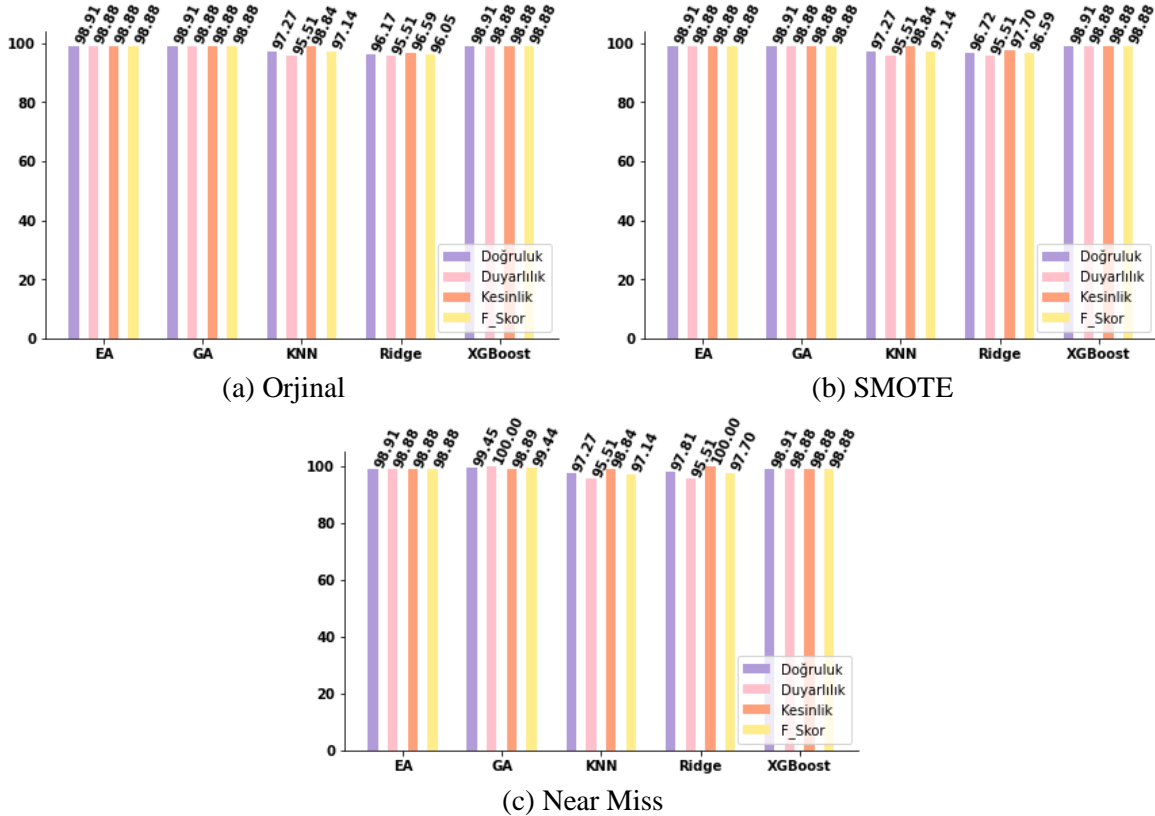
Tablo 6’daki eğitim verileri üzerinden elde edilen sonuçlar incelendiğinde, F-skor değerleri örnekleme ve fazla örnekleme durumlarında %99.75, az örnekleme durumunda ise %99.70 olarak belirlenmiştir. Kesinlik metriği her üç durumda da %100 olarak hesaplanmıştır. Duyarlılık metriği örnekleme ve fazla örnekleme durumlarında %99.50, az örnekleme durumunda %99.40 olarak bulunmuştur. Doğruluk metriği açısından ise, orijinal veri setiyle %99.73, SMOTE uygulanan veri setiyle %99.75 ve Near Miss uygulanan veri setiyle %99.70 doğruluk oranlarına ulaşılmıştır. Bununla birlikte test verileri üzerinden elde edilen sonuçlar incelendiğinde örnekleme ve fazla örnekleme durumları için F-skor, kesinlik ve duyarlılık metrikleri (%98.88) sabitken, doğruluk değeri %98.91 olarak bulunmuştur. Az örnekleme durumunda ise bu metrikler sırasıyla %99.44, %98.86, %100 ve %99.45 olarak elde edilmiştir. Bu bulgular, Near Miss yönteminin GA algoritmasının test verilerindeki performansını olumlu yönde etkilediğini göstermektedir. Şekil 8’de test verileri için GA algoritmasında en yüksek sınıflandırma başarımının (%99.45) elde edildiği duruma ait karmaşıklık matrisi sunulmuştur.



Şekil 8. Test Verileri Üzerinden En Yüksek Başarım Elde Edilen GA Algoritmasına Ait Karmaşıklık Matrisi

Şekil 8'e göre az örnekleme tekniği uygulanan GA algoritmasını kullanan modelin sağlıklı bireyleri doğru sınıflandırma oranı %98.94, yanlış sınıflandırma oranı ise %1.06 olarak belirlenmiştir. Model diyabetli bireyleri yanlış sınıflandırma yapmamıştır. Bu durum modelin diyabetli bireyler için doğru sınıflandırma yaptığını göstermektedir.

Şekil 9'da bireylerin diyabetik durumunun tespitinde kullanılan sınıflandırma algoritmalarının başarımlarının değerlendirilmesi veri setine örnekleme tekniğinin uygulanıp uygulanmaması durumları için tüm performans metrikleri açısından ayrı ayrı gösterilmiştir. Şekilden de açıkça görüldüğü gibi veri setine Near Miss az örnekleme tekniğinin uygulanması durumunda GA algoritması için en yüksek doğruluk değeri (%99.45) elde edilmiştir. Bu örnekleme yöntemi için GA algoritmasına ait duyarlılık, kesinlik ve F-skor metrikleri sırasıyla %100, %98.89 ve %99.44 olarak hesaplanmıştır.



Şekil 9. Diyabet Tahmininde Algoritmaların Başarımlarının Tüm Performans Metrikleri Açısından Yapılan Örnekleme Teknikleri Açısından Değerlendirilmesi a) Orjinal b) SMOTE c) Near Miss

Literatürde MÖ algoritmaları kullanılarak bireylerin diyabetik durumunun tespiti üzerinde çok sayıda çalışma yer almaktadır. Ancak insan vücudunun kompleks yapısı klinik uygulamalarda MÖ algoritmalarının kullanımını sınırlamaktadır (Chen & Pan, 2018). Gerçekleştirilen bu çalışmada açık erişimli güncel bir veri seti kullanılarak bireylerin diyabet risk durumlarının tespitinde MÖ algoritmalarının başarımları incelenmiştir. Çalışma kapsamında elde edilen bulguların önceki çalışmalarla kıyaslanması önem arz etmekte olup, Tablo 7’de bir karşılaştırma tablosu verilmiştir. Tablo 7’den de görüldüğü gibi önerilen algoritma ile elde edilen doğruluk değerinin diğer çalışmalara nazaran daha yüksek olduğu açıkça görülebilmektedir.

Tablo 7. Literatür Karşılaştırması

Yazarlar	Veri Seti	MÖ Algoritması	Başarım
Akyol ve Şen (2018)	Pima Yerlileri diyabet veri seti	Adaboost	%73,88
Jakka ve Vakula Rani (2019)	Pima Yerlileri diyabet veri seti	LR	%77,6
Daghistani ve Alshammari (2020)	Suudi Arabistan'daki bir sağlık kuruluşundan elde edilen veri seti	RO	%88
Shuja vd. (2020)	Kaşmir'deki bir laboratuvarından elde edilen dengesiz bir diyabet veri seti	SMOTE örnekleme yöntemi ve KA	%94,70
Mesquita vd. (2021)	Pima Yerlileri diyabet veri seti	SVM-SMOTE yöntemiyle ve Adaboost	%83,12
Özlüer Başer vd. (2021)	Amerikadaki bir hastaneden elde edilen diyabet veri seti	RO	%84,78
Harman (2021)	Pima Yerlileri Diyabet veri seti	SMOTE örnekleme yöntemi ve DVM	%88
Özkan vd. (2022)	İzmir Bozkaya Eğitim ve Araştırma Hastanesi, Endokrinoloji ve Metabolizma Hastalıkları polikliniğinden alınan veri seti	RO	%84,48
Sevli (2022)	Pima Yerlileri diyabet veri seti	InstanceHardnessThreshold örnekleme yöntemi ve RO	%96,30
Yılmaz vd. (2023)	Bangladeş Sylhet Diyabet Hastanesinde hastalara uygulanan bir anket yöntemi kullanılarak toplanan veri seti	RO	%96
Özoğur ve Orman (2023)	Pima Yerlileri diyabet veri seti	SMOTE-ENN ve MICE örnekleme yöntemleri ve DVM	%90
Korkmaz ve Kaplan (2023)	Pima Yerlileri diyabet veri seti	XGBoost ve LightGBM	%90,01
Önerilen model	Pakistan diyabet veri seti	Near Miss örnekleme yöntemi ve GB	%99,45

TARTIŞMA

Sağlık alanında giderek artan veri hacimleri ve karmaşık bilgi yapıları, MÖ algoritmalarının kullanımına olan ilgiyi büyük ölçüde artırmaktadır. Bu algoritmalar, büyük veri kümeleri arasında ilişkileri belirleyebilmekte, hastalık teşhisi ve prognozlamada önemli bir rol oynayabilmekte ve tedavi yöntemlerinin kişiselleştirilmesine olanak tanımaktadır. Bunun yanı sıra MÖ temelli geliştirilen modeller, tıbbi karar alma süreçlerinde objektif ve veriye dayalı bir yaklaşım sunarak, klinik sonuçları iyileştirmeye ve sağlık hizmetlerini optimize etmeye yardımcı olmaktadır (Chen & Pan, 2018). Gerçekleştirilen bu çalışmada, dünya çapında yaygın olarak karşılaşılan ve insidansı giderek artan diyabet hastalığının erken teşhisi için MÖ algoritmaları ile örnekleme yöntemlerinin birleştirildiği bir model önerilmiştir. Modelde MÖ algoritması olarak literatürde diyabet tanısında sıklıkla tercih edilen KNN, Ridge, XGBoost, EA ve GA algoritmaları kullanılmıştır. Ayrıca, örnekleme yöntemleri olarak SMOTE (Sentetik Azaltma Tekniği) ve Near Miss yöntemleri uygulanmıştır. Elde edilen bulgular, farklı MÖ algoritmalarının diyabet sınıflandırılmasındaki performanslarını değerlendirmek için çeşitli örnekleme yöntemlerinin kullanılabilceğini göstermektedir. Özellikle GA algoritmasının Near Miss alt örnekleme yöntemi ile en iyi performansa sahip olduğu tespit edilmiştir. Bu bulgu, diyabet sınıflandırmasında örnekleme tekniklerinin etkilerini ele almanın önemini vurgulamaktadır. Özellikle, diyabetin erken teşhisi ve doğru sınıflandırılması, hastaların tedavi ve yönetiminde kritik bir öneme sahiptir. Bu bağlamda, elde edilen sonuçlar sağlık uzmanlarına ve

araştırmacılara, diyabet tanısında MÖ algoritmalarının potansiyelini değerlendirme konusunda yol gösterici olacağı düşünülmektedir. Ancak çalışmada veri ön işleme aşamalarında yüksek hesaplama yükü ve zaman maliyeti gibi zorluklarla karşılaşmıştır. Gelecek çalışmalarda, hibrit öznelik seçimi yöntemlerinin kullanılması, hiperparametre optimizasyon tekniklerinin uygulanması ve veri artırma yöntemlerinin entegre edilmesi önerilmektedir. Bu iyileştirmeler, modelin performansını artırmak ve genellenabilirliğini sağlamak adına önemli katkılar sunabilir. Ayrıca, daha geniş ve çeşitli veri setleri kullanılarak modelin farklı veri dağılımlarına karşı dayanıklılığı artırılabilir. Bu tür yaklaşımlar, diyabet teşhisi ve yönetiminde MÖ algoritmalarının etkinliğini ve doğruluğunu artırabilir.

KAYNAKLAR

- Alehegn, M., Raghvendra Joshi, R., & Mulay, P. (2019). Diabetes Analysis And Prediction Using Random Forest, KNN, Naïve Bayes, And J48: An Ensemble Approach. *International Journal of Scientific & Technology Research*, 8(9), 1346-1354.
- Akyol, K., & Şen, B. (2018). Diabetes Mellitus Data Classification by Cascading of Feature Selection Methods and Ensemble Learning Algorithms. *International Journal of Modern Education and Computer Science*, 10(6), 10-16. <https://doi.org/10.5815/ijmeecs.2018.06.02>
- Dal, A., Gümü, İ. H., Güldal, S. & Yavaş, M. (2021). Dengesiz Veriler İçin Ağırlıklı Geometrik Ortalama Tabanlı Yeni Bir Yeniden Örnekleme Yaklaşımı, *Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi*, 8 (15), 343-352. <https://doi.org/10.54365/adyumbd.940539>
- Daghistani, T., & Alshammari, R. (2020). Comparison of statistical logistic regression and randomforest machine learning techniques in predicting diabetes. *Journal of Advances in Information Technology*, 11(2), 78-83. <https://doi.org/10.12720/jait.11.2.78-83>
- Das, H., Naik, B., & Behera, H. S. (2018). Classification of diabetes mellitus disease (DMD): A data mining (DM) approach. *Advances in Intelligent Systems and Computing*, 710, 539-549. Springer Verlag. https://doi.org/10.1007/978-981-10-7871-2_52
- Hacıbeyoğlu, M., Çelik, M., & Erdaş Çiçek, Ö. (2023). En Yakın Komşu Algoritması ile Binalarda Enerji Verimliliği Tahmini. *Necmettin Erbakan Üniversitesi Fen ve Mühendislik Bilimleri Dergisi*, 5(2), 28-37. <https://doi.org/10.47112/neufmbd.2023.10>
- Harman, G. (2021). Destek vektör makineleri ve naive bayes sınıflandırma algoritmalarını kullanarak diyabet mellitus tahmini. *Avrupa Bilim ve Teknoloji Dergisi*, (32), 7-13. <https://doi.org/10.31590/ejosat.1041186>
- IDF Diabetes Atlas. Diabetes around the world in 2021. <https://diabetesatlas.org/> Accessed 04.04.2024
- Jakka, A., & Vakula Rani, J. (2019). Performance evaluation of machine learning models for diabetes prediction. *International Journal of Innovative Technology and Exploring Engineering*, 8(11), 1976-1980. <https://doi.org/10.35940/ijitee.K2155.0981119>
- Kaggle. Pakistani Diabetes Dataset. (2024). <https://www.kaggle.com/datasets/mshoaibishaaq/pakistani-diabetes-dataset> Accessed 04.04.2024
- Korkmaz, M., & Kaplan, K. (2023). Şeker hastalığı teşhisi ve önerilen modellerinin karşılaştırılması. *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 12(1), 1-1.
- Mesquita, F., aurício, J., & Marques, G. (2021). Oversampling techniques for diabetes classification: A comparative study. In 2021 International Conference on e-Health and Bioengineering (EHB) (pp. 1-6). IEEE.
- Mqadi, N. M., Naicker, N., & Adeliyi, T. (2021). Solving misclassification of the credit card imbalance problem using near miss. *Mathematical Problems in Engineering*, 2021, 1-16. <https://doi.org/10.1155/2021/7194728>
- Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In 2020 11th international conference on information and communication systems (ICICS) (pp. 243-248). IEEE.
- Özüler Başer, B., Yangın, M., & Sarıdaş, E. S. (2021). Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 25(1), 112-120. doi: 10.19113/sdufenbed.842460

- Özoğur, H. N., & Orman, Z. (2023). Sağlık Verilerinin Analizinde Veri Ön İşleme Adımlarının Makine Öğrenmesi Yöntemlerinin Performansına Etkisi. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 16(1), 23-33.
- Öztürk, H. (2022). Dengesiz veri setlerinde farklı dengeleme algoritmalarının optimum denge oranlarının sınıflandırma ve regresyon ağaçları yöntemi ile incelenmesi: simülasyon çalışması. Doktora Tezi. Aydın Adnan Menderes Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik, Aydın 88s.
- Özkan, Y., Yürekli, B. S., & Suner, A. (2022). Diyabet tanısının tahminlenmesinde denetimli makine öğrenme algoritmalarının performans karşılaştırması. *Gümüşhane Üniversitesi Fen Bilimleri Dergisi*, 12(1), 211-226. <https://doi.org/10.17714/gumusfenbil.820882>
- Sungur, F., & Bakır, H. (2024). Hiperparametre ayarlama ve veri dengelemenin kalp hastalığı tahmini için kullanılan makine öğrenimi algoritmaları üzerindeki etkilerinin incelenmesi. *Bilişim Teknolojileri Dergisi*, 17(1), 45-58. <https://doi.org/10.17671/gazibtd.1399813>
- Sevli, O. (2022). Diyabet hastalığının farklı sınıflandırıcılar kullanılarak teşhisi. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 38(2), 989-1002. <https://doi.org/10.17341/gazimmfd.880750>
- Shuja, M., Mittal, S., & Zaman, M. (2020). Effective prediction of type ii diabetes mellitus using data mining classifiers and SMOTE. In *Advances in Computing and Intelligent Systems: Proceedings of ICACM 2019* (pp. 195-211). Springer Singapore.
- Turan, T. (2023). Makine Öğrenmesi Algoritmaları İle Su Kalitesi Ve İçilebilirlik Tahmini. *Uluborlu Mesleki Bilimler Dergisi*, 6(2), 65-80.
- Türkmenoğlu, B. K., & Yıldız, O. (2021). Predicting the survival of heart failure patients in unbalanced data sets. In *2021 29th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.
- Turhan S., Özkan, Y., Yürekli, B. S., Suner, A., & Doğu, E. (2020). Comparison of Ensemble Learning Methods for Disease Diagnosis in Presence of Class Unbalanced: Case of Diabetes. *Turkiye Klinikleri Journal of Biostatistics*, 12(1), 16-26. <https://doi.org/10.5336/biostatic.2019-66816>
- Uğurgöl, E., Batbat, T., Yeşilbas, D., Altınkaynak, M., Güven, A., Demirci, E., & Dolu, N. (In press) Doğrusal olmayan EEG dinamikleri ile anksiyete tespiti. *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 13(2), 1-1. <https://doi.org/10.28948/ngumuh.1359809>
- WHO. Diabets. (2024). https://www.who.int/health-topics/diabetes#tab=tab_1 Accessed 04.04.2024
- Yılmaz, H., Buldu, A., Kaya, Y., & Kuncan, F. (2023). Yapay Zeka Yöntemleriyle Erken Evre Diyabet Risk Tahmini. *Journal of Scientific Reports-B*, (007), 59-71.
- Yakut, G., Çay, R. İ., & Öztürk, H. H. (2023). Makine Öğrenimi Teknikleri Kullanılarak Isparta İli İçin Tarımsal Ürün Önerme Sistemi. *Gazi Mühendislik Bilimleri Dergisi*, 9(4-ICAIAME 2023), 174-185. <https://doi.org/10.30855/gmbd.0705S17>
- Yavaş, M., Güran, A., & Uysal, M. (2020). Covid-19 veri kümesinin SMOTE tabanlı örnekleme yöntemi uygulanarak sınıflandırılması. *Avrupa Bilim ve Teknoloji Dergisi*, 258-264. <https://doi.org/10.31590/ejosat.779952>