Journal of the Faculty of Engineering and Architecture of Gazi University 38:2 (2023) 989-1001

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Diagnosis of diabetes mellitus using various classifiers

Onur Sevli*

Department of Computer Engineering, Faculty of Engineering and Architecture, Burdur Mehmet Akif Ersoy University, 15030, Burdur, Türkiye

Highlights:

- Early diagnosis of Diabetes Mellitus
- The effect of resampling methods on classification success
- Performance comparison of six different machine learning algorithms
-

Keywords:

- Diabetes diagnosis
- Machine learning
- Resampling

Article Info:

Research Article
Received: 15.02.2021
Accepted: 01.05.2022

DOI:

10.17341/gazimmfd.880750

Correspondence:

Author: Onur Sevli
e-mail: onursevli@mehmetakif.edu.tr
phone: +90 248 213 4130

Graphical/Tabular Abstract

In this study, performance comparisons were performed as shown in Figure A, using various classifiers with resampling techniques for the diagnosis of diabetes.

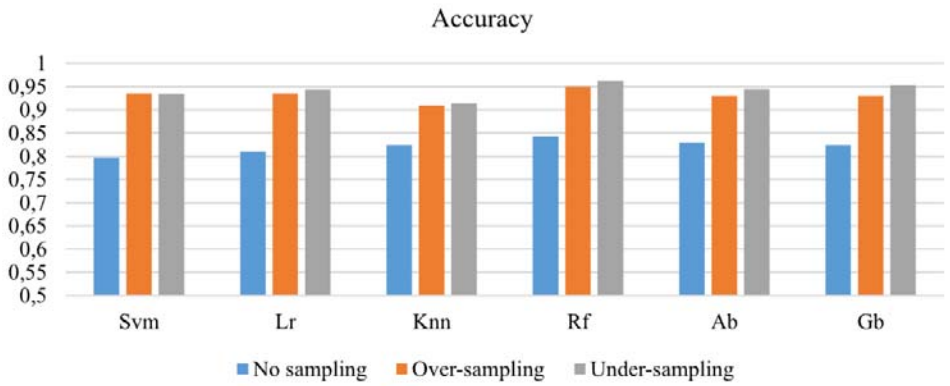


Figure A. Accuracy values of the classifiers obtained with and without resampling

Purpose:
This study aims to evaluate diabetes risk according to individuals' lifestyles and family history. Accordingly, it includes classification studies for early diagnosis of diabetes with different machine learning algorithms and resampling techniques on the Pima Indian Diabetes dataset.

Theory and Methods:
Support Vector Machine, Logistic Regression, K-Nearest Neighbor, Random Forest, AdaBoost and Gradient Boosting methods were used for classification. In addition to the non-sampling success of these six different methods used in this study, fourteen different resampling techniques were applied for each independently, and the successes of ninety different classification processes were reported.

Results:
The resampling techniques generally increased the success of the classifiers. The highest scores were obtained in the classification process where the Random Forest was used with InstanceHardnessThreshold under-sampling technique. The best results obtained were 96.29% as accuracy, 98.07% as precision, 100% as recall, 96.22% as F1 Score, and 96.29% as AUC. Following the success achieved with Random Forest, the highest accuracy values were obtained as 95.32% and 94.49% with Gradient Boosting and AdaBoost using the same resampling method. The results obtained were compared with recent studies conducted with various machine learning techniques on the same data set in the literature. It was revealed that the results obtained in this study were higher than in other studies.

Conclusion:
Diabetes is one of the common diseases with rapidly increasing prevalence worldwide. Machine learning techniques play the role of an intelligent decision support system that helps experts in the diagnosis of different diseases. Balancing datasets with appropriate resampling techniques reduces the problem of overfitting and improves the accuracy of the classification. As a result of the combination of resampling and ensemble learning, it was observed that the accuracy becomes even higher.

Journal of the Faculty of Engineering and Architecture of Gazi University 38:2 (2023) 989-1001

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Diyabet hastalığının farklı sınıflandırıcılar kullanılarak teşhisi

Onur Sevli*
Burdur Mehmet Akif Ersoy Üniversitesi, Mühendislik Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, 15030, Burdur, Türkiye

Ö N E Ç I K A N L A R

- Diyabet hastalığının erken teşhisi
- Yeniden örnekleme yönteminin sınıflandırma başarısına etkisi
- Altı farklı makine öğrenmesi algoritmasının performans karşılaştırması

Makale Bilgileri
Araştırma Makalesi
Geliş: 15.02.2021
Kabul: 01.05.2022

DOI:
10.17341/gazimmfd.880750

Anahtar Kelimeler:
Diyabet teşhisi, makine öğrenmesi, yeniden örnekleme

ÖZ

Diyabet dünya genelinde görülme oranı giderek artan, yaygın sağlık sorunlarından biridir. Kronik bir hastalık olan diyabet kontrol altına alınmadığı takdirde göz, kalp, böbrek gibi organlara zarar verebilmekte ve ölümlere neden olabilmektedir. Diyabetin erken teşhisi oluşabilecek komplikasyonları önleme ve yaşam kalitesini arttırma açısından önemlidir. Medikal alanda yaygın kullanılan makine öğrenmesi teknikleri farklı hastalıkların teşhisinde uzmanlara yardımcı akıllı bir karar destek sistemi rolü üstlenmektedir. Bu çalışma, diyabetin erken teşhisine yönelik olarak altı farklı makine öğrenmesi tekniği ile Pima Indian Diabetes veri seti üzerinde gerçekleştirilen sınıflandırma çalışmalarını içermektedir. Gerçekleştirilen sınıflandırmalardaki temel hedeflerden biri tahmin doğruluğunu arttırmaktır. Bu çalışmada sınıflandırıcıların başarılarını arttırmak için veri seti üzerinde on dört farklı yeniden örnekleme yöntemi kullanılmıştır. Her bir makine öğrenmesi modeli için örnekleme olmaksızın ve yeniden örnekleme yapılarak, toplam doksan sınıflandırma işlemi gerçekleştirilmiştir. Her bir sınıflandırma işleminin başarısı beş farklı performans metriği ile raporlanmıştır. En başarılı sonuç %96,296 doğrulukla, InstanceHardnessThreshold az örnekleme tekniği ile birlikte Rastgele Orman modelinin kullanıldığı sınıflandırma işleminde elde edilmiştir. Yeniden örnekleme tekniklerinin genel olarak sınıflandırıcıların başarılarını arttırdığı ve kolektif öğrenme yöntemleri ile birlikte kullanıldığında daha başarılı sonuç verdiği görülmüştür. Literatürdeki diğer benzer çalışmalarla karşılaştırıldığında bu çalışmada elde edilen sonuçların diğerlerinden daha yüksek olduğu sonucuna varılmıştır.

Diagnosis of diabetes mellitus using various classifiers

H I G H L I G H T S

- Early diagnosis of Diabetes
- The effect of resampling method on classification success
- Performance comparison of 6 different machine learning algorithms

Article Info
Research Article
Received: 15.02.2021
Accepted: 01.05.2022

DOI:
10.17341/gazimmfd.880750

Keywords:
Diabetes diagnosis, machine learning, resampling

ABSTRACT

Diabetes is one of the common health problems with an increasing incidence worldwide. Diabetes is a chronic disease that can damage organs such as the eyes, heart, and kidneys, as well as cause mortality if not taken under control. Early diagnosis of diabetes is important in terms of preventing complications and increasing the quality of life. Machine learning techniques, which are widely used in the medical field, play the role of an intelligent decision support system that helps experts in the diagnosis of different diseases. This study includes classifications performed on the Pima Indian Diabetes dataset with six different machine learning techniques for the early diagnosis of diabetes. One of the main goals of the classifications carried out is to increase the prediction accuracy. In this study, fourteen different resampling methods were used on the dataset to increase the success of the classifiers. A total of ninety classifications were carried out without sampling and resampling for each machine learning model. The success of each classification process was reported with five different performance metrics. The highest performance was obtained with an accuracy of 96.296% in the classification using the Random Forest with the InstanceHardnessThreshold under-sampling technique. It was observed that resampling techniques generally increased the success of the classifiers and were more successful when used together with ensemble learning methods. Compared to the other similar studies in the literature, it was shown that the results obtained in this study were higher than the others.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *onursevli@mehmetakif.edu.tr / Tel: +90 248 213 4130

1. Giriş (Introduction)

Diyabet (Diabetes Mellitus) dünya genelinde yaygın olarak görülen sağlık sorunlarından biridir. İnsülinin mutlak veya göreceli yokluğuna bağlı olarak gelişen diyabet, kandaki glikoz seviyesinin kabul edilebilir sınırların üzerinde seyrettiği, vücudun glikoz kullanımını etkileyen metabolik bir bozukluktur [1]. Uluslararası Diyabet Federasyonu'nun (IDF) istatistiklerine göre, dünya genelinde 451 milyon diyabet hastası bulunmakta ve hastalığın görülme oranı giderek artmaktadır. Vaka artışları göz önüne alındığında bu sayının 2045 yılına gelindiğinde 693 milyona ulaşacağı tahmin edilmektedir [2]. Bu durum diyabetin önemsenmesi gereken bir hastalık olduğunu açıkça göstermektedir. Diyabet hastalığı kontrol altına alınmazsa farklı sağlık sorunlarına hatta ölümcül sonuçlara neden olabilir. Dünya Sağlık Örgütü tarafından 2016 yılında yayımlanan global diyabet raporuna göre, erken ölüm nedenleri içerisinde diyabet hastalığı yedinci sırada yer almaktadır. 2016 yılınca yaklaşık 1.6 milyon ölümün doğrudan diyabet kaynaklı olduğu raporlanmıştır [3]. Diyabet hastalığında pankreas yeteri kadar insülin üretemez veya hücre ve dokular üretilen insülini etkin şekilde kullanamaz [4]. Bunun sonucunda kandaki glikoz miktarı anormal seviyelere ulaşır. Sağlıklı bireylerde kandaki glikoz miktarı 70 ile 99 mg/dL düzeyindedir. 100-125 mg/dL glikoz konsantrasyonuna sahip bireyler ise prediyabetik olarak değerlendirilir. Diyabet hastalığı Tip-1, Tip-2 ve gebelik döneminde görülen diyabet olmak üzere üç grupta incelenebilir. Tip-1 diyabet pankreastaki fonksiyonel bozukluktan dolayı insülin salgılanamaması sonucu ortaya çıkar. Bu nedenle Tip-1 diyabet hastaları vücudun ihtiyaç duyduğu insülini karşılamak için dışarıdan insülin dozları almak zorundadırlar. Tip-2 diyabet ise insülin direnci ve insülin sekresyon eksikliği ile ortaya çıkan diyabet türüdür. Prediyabetik bireyler birer Tip-2 diyabet adayıdır. Diyabeti hastalığını tamamen tedavi eden şu ana kadar bilinen kesin bir çözüm yoktur ancak erken teşhis hastalığın ilerlemesini ve ortaya çıkacak komplikasyonları önlemek açısından önemlidir.

Vücuttaki yüksek glikozun yol açtığı hiperglisemi, kardiyovasküler anormalliklere ve ayrıca gözler, böbrekler, sinir sistemi gibi çeşitli organ, doku ve sistemlerin işleyişinde sorunlara yol açabilmektedir [5]. Diyabetin erken teşhisi sadece tedaviye erken başlama açısından değil, hastalığın ortaya çıkma potansiyelinin yüksek olduğu durumlarda gerekli önlemlerin alınması açısından da önemlidir. Yaşam tarzında yapılacak değişikliklerle diyabetin etkilerinin önlenebilmesi mümkündür. Hastalığın teşhisi, alan uzmanlarının bilgi ve tecrübesine dayanır. Ancak insanlar karar verme konusunda her zaman istikrarlı olamayabilir. Bu nedenle, uzmanlar için diyabet hastalığının teşhisini kolaylaştıran erken tanı araçlarının geliştirilmesi, sağlık hizmetlerinin kalitesini arttırmak açısından önem taşımaktadır. Sağlık alanında büyük miktarda veri toplanmaktadır ve etkili kararlar verebilmek için veriler arasındaki kritik örüntülerin ortaya konması gerekmektedir. Ancak bu örüntülerin tümüyle, insan eliyle ortaya konması son derece zordur. Hastalıkların erken ve yüksek doğrulukla teşhis edilebilmesi için, veriler arasındaki örüntüleri başarı ile ortaya koyabilen gelişmiş destek mekanizmalarına ihtiyaç vardır [6]. Teknolojik alanda meydana gelen hızlı gelişmeler, hastalıkların erken aşamada teşhis edilmesi ve raporlanması konusunda yeni olanaklar sunmaktadır. Medikal alanda kullanımı giderek yaygınlaşan makine öğrenmesi, görülme sıklığı günden güne artan diyabetin erken ve etkin teşhisinde alternatif çözümler vadetmektedir. Makine öğrenmesi, bilgisayarların mevcut verilerden öğrenerek, yeni durumlar hakkında tahminde bulunmasına olanak sağlayan algoritmalar bütünüdür. Yapay zekânın bir alt dalı olan makine öğrenmesi istatistik ile de yakından ilgilidir. Yapay zekâ alanındaki hızlı gelişmelerle birlikte, makine öğrenmesi medikal alanda geniş perspektife yayılan çözümler üretmeye devam etmektedir. Makine öğrenmesi teknikleri, rutin kontrol sonuçlarına bağlı olarak diyabet hastalığının teşhisi konusunda uzmanlara yüksek başarımlı karar desteği sağlayacak bir potansiyele sahiptir [7].

Literatürde diyabet hastalığının teşhisine yönelik, farklı veri setleri üzerinde, çeşitli makine öğrenmesi teknikleri kullanılarak gerçekleştirilen çalışmalar mevcuttur. Bu çalışmalarda genel olarak Lojistik Regresyon (LR), Naive Bayes (NB), K-En Yakın Komşu (KNN), Destek Vektör Makinesi (DVM), Karar Ağaçları (KA), Rastgele Orman (RO), Gradient Boosting (GB) ve Yapay Sinir Ağları (YSA) modelleri kullanılmaktadır. Lai vd. diyabet teşhisi için, yaşları 18 ile 90 arasında değişen Kanadalı hastaların laboratuvar ölçümlerini analiz etmişlerdir. Yaş, cinsiyet, açlık kan şekeri, kan basıncı, vücut kitle indeksi, yüksek ve düşük yoğunluklu lipoprotein ve trigliserit ölçümlerini içeren 13309 adet veri örneğini, LR ve GB yöntemleri ile sınıflandırmışlardır. LR kullanılarak elde edilen eğri altında kalan alan (Area Under the Curve - AUC) %84, duyarlılık %73,4, GB kullanarak elde edilen AUC %84,7 ve duyarlılık ise %71,6'dır [8]. Kopitar vd., Slovenyada'ki on farklı sağlık merkezinden toplanan elektronik sağlık kayıtları üzerinde çeşitli makine öğrenmesi modelleri ile Tip-2 diyabet hastalığının teşhisine yönelik tahmin çalışması gerçekleştirmişlerdir. Yaş ortalamaları 54,45 ±11,69 olan, 3723 hastaya ait 61 farklı özellik içeren veri seti üzerinde Lineer Regresyon, RO ve GB yöntemleri kullanılarak gerçekleştirilen tahminlerin başarıları ortalama karesel hata (RMSE) olarak raporlanmıştır. En başarılı model 0.838 ile en düşük RMSE değerine sahip Lineer Regresyon olurken bunu 0,842 ile RO ve 0,846 ile GB izlemiştir [9]. Maniruzzaman vd., ABD Sağlık ve Sosyal Hizmetler Bakanlığı (HHS) tarafından 2009-2012 yılları arasında yürütülen "Ulusal Sağlık ve Beslenme İnceleme Anketi"nden türetilen bir veri setini kullanarak, makine öğrenmesi ile diyabet hastalığı teşhisine yönelik bir çalışma gerçekleştirmişlerdir [10]. 657 diyabetik ve 5904 sağlıklı bireyden oluşan, toplam 6561 kayıt içeren veri seti üzerinde NB, KA, Adaboost ve RO yöntemleri ile gerçekleştirilen sınıflandırmada ortalama %90,62 doğruluk elde edilmiştir. RO sınıflandırıcı ile LR tabanlı özellik seçicinin birlikte kullanılması sonucu doğruluk değerinin %94,25'e yükseldiği raporlanmıştır. Zhang vd., Çin'in Henan eyaletindeki beş ayrı kırsal bölgede yaşları 18 ile 79 arasında değişen 39259 katılımcıdan toplanan veriler ile diyabet riskini karakterize etmeye yönelik bir çalışma gerçekleştirmişlerdir [11]. Güler ve Übeyli 14 özellik ve toplam 760 kayıttan oluşan veri seti üzerinde dört ayrı algoritma ile eğitilen çok katmanlı perseptron sinir ağları ile diyabet hastalığının teşhisine yönelik bir çalışma gerçekleştirmişlerdir. Hızlı yayılım algoritmasının en başarılı algoritma olduğunu ortaya koymuşlardır [12] LR, KA, YSA, DVM, RO ve GB olmak üzere altı farklı makine öğrenmesi yöntemiyle gerçekleştirilen çalışmada en iyi sonuç 0,872 AUC değeri ile GB yönteminden elde edilmiştir. Muhammad vd., Nijerya'nın Kano eyaletindeki bir hastanede diyabet hastalığının tespitine yönelik olarak toplanan, 9 özellik ve 383 kayıttan oluşan veriler üzerinde, farklı makine öğrenmesi teknikleri ile tahmin çalışması gerçekleştirmişlerdir [13]. LR, DVM, KNN, RO, NB ve GB yöntemleri ile gerçekleştirilen sınıflandırmalarda sırasıyla %80,88, %85,29, %82,35, %88,76, %77,94 ve %86,76 doğruluk değerine ulaşılmıştır. En yüksek başarı RO sınıflandırıcı ile elde edilmiştir.

Diyabet konusunda, çeşitli internet veri depoları ve harici kaynaklarda farklı veri setlerine ulaşmak mümkündür. Amerikan Ulusal Diyabet, Sindirim ve Böbrek Hastalıkları Enstitüsü (National Institute of Diabetes and Digestive and Kidney Diseases-NIDDK) tarafından oluşturulan Pima Indian Diabetes (PIMA) veri seti ise literatürdeki çalışmalarda yaygın olarak kullanılmaktadır. Sisodia vd., KA, DVM ve NB yöntemlerini kullanarak PIMA veri seti üzerinde diyabet hastalığının erken tanısına yönelik bir çalışma gerçekleştirmişlerdir. Kullandıkları yöntemler ile elde ettikleri doğruluk değerleri sırasıyla %73,6, %51,3 ve %76,30'dur [14]. Zou vd., YSA ve RO kullanarak

bir sınıflandırma çalışması gerçekleştirmişler, sırayla %76,67 ve %76,04 oranında doğruluk elde etmişlerdir. Aynı çalışmada, veriler üzerinde boyut indirgeme uygulandığında ise doğruluk değerleri %71,44 ve %74,75 olarak rapor edilmiştir [15]. Wei vd. parametre optimizasyonu ve 10 kat çapraz doğrulama kullanarak, dört farklı makine öğrenmesi modeli ile karşılaştırmalı bir sınıflandırma işlemi gerçekleştirmişlerdir. Elde edilen en iyi doğruluk değerleri LR için %77,47, DVM için %77,60, KA için %76,30 ve NB için %75,78'dir [16]. Kohli ve Arora; AdaBoost, KA, LR, RO ve DVM kullanarak gerçekleştirdikleri çalışmada sırayla %80,52, %74,03, %84,42, %81,82 ve %85,71 doğruluk elde etmişlerdir [17]. Mir ve Dhage ise NB, DVM ve RO kullanarak gerçekleştirdikleri çalışmalarında %77, %79,13 ve %76,5 doğruluğa ulaşmışlardır [18]. Varma ve Panda diyabet hastalığının tahminlenmesine yönelik gerçekleştirdikleri çalışmada DVM, KNN ve RO modellerini karşılaştırmışlardır. Kullandıkları üç ayrı yöntemde sırayla %72,17, %73,57 ve %74,67 doğruluk elde etmişlerdir [19]. Radja ve Emanuel, PIMA veri setini farklı boyutlarda veri kümelerine bölerek NB, DVM ve KA yöntemlerinin performanslarını incelemişlerdir. En yüksek doğruluk değerini %77,3 olarak DVM ile elde etmişlerdir [20]. Yahyaoui vd. DVM ve RO ile gerçekleştirdikleri sınıflandırmada %83,67 ve %65,38 doğruluk elde etmişlerdir [21]. Benbelkacem ve Atmani medikal çalışmalarda yaygın kullanılan RO yöntemini, farklı sayıda karar ağacından oluşan varyasyonlar ile PIMA veri seti üzerinde uygulayarak optimum çözümü bulmaya çalışmışlardır. 40 ağaçtan oluşan modelin 0,21 hata oranı ile en iyi performansı gösterdiği sonucuna varmışlardır [22]. Birjais vd.; NB, LR ve GB yöntemleri ile gerçekleştirdikleri sınıflandırmada %77, %79,2 ve %86 doğruluk elde etmişlerdir [23]. Wang vd., sınıflandırma öncesinde veriler üzerinde iki aşamalı önişleme gerçekleştirmiş, önce kayıp verilerin üretilmesi ve ardından Adaptive Synthetic (ADASYN) algoritması ile aşırı örnekleme işlemi uygulamışlardır. Daha sonra RO ve 5 kat çapraz doğrulama kullanarak gerçekleştirdikleri sınıflandırmada %87,1 doğruluğa ulaşmışlardır [24]. Srivastava vd. sınıflandırma işleminde YSA kullanmışlar ve %92 doğruluk elde etmişlerdir [25]. Yuvaraj ve SriPreethaa ilk olarak veri seti üzerinde bilgi kazancı (information gain – IG) yöntemi ile özellik indirgeme gerçekleştirmiş ve ardından üç farklı makine öğrenmesi yöntemi ile sınıflandırma yapmışlardır. KA ile %88, NB ile %91 ve RO ile %94 doğruluk elde etmişlerdir [26]. Battineni vd. LR, RO ve NB yöntemleri ile farklı çapraz doğrulama değerleri kullanarak gerçekleştirdikleri sınıflandırma çalışmalarında sırayla %83, %82, %81 doğruluk oranlarına ulaşmışlardır [27]. Agarwal ve Saxena, 10 kat çapraz doğrulama ile LR, DVM, NB, KA ve KNN olmak üzere beş farklı makine öğrenmesi yönteminin sınıflandırma başarısını karşılaştırmışlardır. En yüksek doğruluğu %81 ile LR sınıflandırıcı ile elde etmişlerdir [28]. Livington vd., Fuzzy sınıflandırıcılar kullanarak gerçekleştirdikleri sınıflandırma çalışmasında %83 doğruluğa ulaşmışlardır [29]. Naz ve Ahuja, NB ve YSA yöntemlerinin sınıflandırma başarılarını karşılaştırmışlar, NB için %76,33, YSA için %90,34 doğruluk elde etmişlerdir [30]. Hasan vd.; KNN, KA, RO, NB, AdaBoost ve XGBoost yöntemleri ile sınıflandırma gerçekleştirmişler ve elde ettikleri en yüksek AUC değeri %95 olmuştur [31]. Tigga ve Garg., farklı makine öğrenmesi teknikleri ile gerçekleştirdikleri çalışmalarında en yüksek doğruluğu %94,1 olarak RO ile elde etmişlerdir. Kaur ve Kumari ise DVM, KNN ve YSA ile gerçekleştirdikleri sınıflandırma çalışmalarında %89, %88 ve %86 doğruluğa ulaşmışlardır [32]. Patil vd. diyabetin erken teşhisi konusunda DVM, KA ve RO ile gerçekleştirdikleri karşılaştırmalı çalışmada %74,9, %75,32 ve %77,05 doğruluk elde etmişlerdir [33]. Pranto vd. PIMA veri seti üzerinde çeşitli makine öğrenmesi teknikleri ile oluşturdukları modeller ile Bangladeşli kadınların diyabet durumlarını tahminlemeye yönelik bir çalışma gerçekleştirmişlerdir. 3 kat çapraz doğrulama kullanarak, KNN, KA, RO ve NB yöntemleri ile gerçekleştirdikleri sınıflandırmalar sonucunda %75,7, %73,1, %77,9 ve %72,1 doğruluk elde etmişlerdir [34]. Reddy vd. sınıflandırma için DVM, KNN, LR, NB ve GB tekniklerini kullanmışlar ve sırasıyla %79,15, %87,61, %80,64, %77,34 ve %87,31 doğruluk elde etmişlerdir [35]. Nusrat vd.; KA, RO ve GB kullanarak gerçekleştirdikleri çalışmalarında %73,69, %74,50 ve %76,30 doğruluğa ulaşmışlardır [36]. Köse, zeki optimizasyon tabanlı destek vektör makineleri ile gerçekleştirdiği diyabet teşhisine yönelik çalışmasında Parçacık Sürü Optimizasyonu (PSO), Genetik Algoritma (GA), Guguk Kuşu Araması (GKA), Bakteriyel Yiyecek Arama Optimizasyonu (BYAO) ve Çiçek Tozlaşma Optimizasyonu (ÇTO) algoritmalarını kullanmıştır. PSO tabanlı DVM ile %74,61, GA ile %75,22, GKA ile %94,42, BYAO ile %90,59 ve ÇTO ile %91,77 doğruluğa ulaşmıştır [37].

Bu çalışmada farklı makine öğrenmesi yöntemleri kullanılarak diyabet hastalığının erken teşhisine yönelik bir çözüm sunulmuştur. Kullanılan farklı yöntemlerin başarıları karşılaştırılmış ve ayrıca sınıflandırıcıların tahmin başarılarını arttırmaya yönelik farklı yeniden örnekleme teknikleri uygulanmıştır.

2. Materyal ve Metot (Material and Method)

Bu çalışma, bireylerin diyabet riskini yaşam tarzlarına ve aile geçmişlerine göre değerlendirmeyi amaçlamakta ve bu doğrultuda PIMA veri seti üzerinde farklı makine öğrenmesi algoritmaları ile diyabet hastalığının erken teşhisine yönelik sınıflandırma çalışmalarını içermektedir. Sınıflandırma için Destek Vektör Makinesi, Lojistik Regresyon, K-En Yakın Komşu, Rastgele Orman, AdaBoost ve Gradient Boosting yöntemleri kullanılarak performansları farklı metrikler açısından değerlendirilmiştir. Sınıflandırma çalışmalarındaki ana amaçlardan biri tahmin başarısını arttırmaktır. Bu doğrultuda, çalışmada kullanılan altı farklı yöntemin varsayılan durumundaki başarıları yanında, her biri için on dört farklı yeniden örnekleme metodu bağımsız şekilde uygulanarak, doksan ayrı sınıflandırma işleminin başarıları raporlanmıştır.

2.1. Veri Seti (Dataset)

Diyabet hastalığının erken teşhisine yönelik bir çözüm sunmayı amaçlayan bu çalışmada, literatürde yaygın olarak kullanılan Pima Indian Diabetes (PIMA) veri seti kullanılmıştır. PIMA veri seti, Amerika Ulusal Diyabet ve Sindirim ve Böbrek Hastalıkları Enstitüsü tarafından 1965 yılından başlayarak yürütülen uzun vadeli bir araştırmanın ürünüdür. Kamuya açık olarak paylaşılan PIMA veri setine, University of California Irvine (UCI) Machine Learning Repository isimli veri deposundan erişilebilmektedir [38]. Veri seti içerisindeki örnekler, Arizona bölgesinde yaşayan 21 ile 81 yaş arası Pima Indian kökenli kadınlara aittir. Veri seti 768 adet örnekten oluşmakta; her bir örnek 8 adet girdi özellik ve bir adet teşhis sınıfı ile temsil edilmektedir. Sınıf değeri bireyin sağlıklı ya da diyabet hastası olmasına bağlı olarak iki değere sahiptir. Veri seti içerisinde 500 adet sağlıklı, 268 adet diyabet hastasına ait kayıt bulunaktadır. Veri setindeki özelliklerin detayları Tablo 1'de verilmiştir. Veri setinde yer alan özelliklerin istatistiki karakteristikleri ise Tablo 2'de yer almaktadır.

Veri seti özelliklerinin korelasyon ilişkileri Şekil 1'de verilmiştir. Renk skalasında açık renkler yüksek korelasyonu, koyu renkler ise düşük korelasyonu göstermektedir.

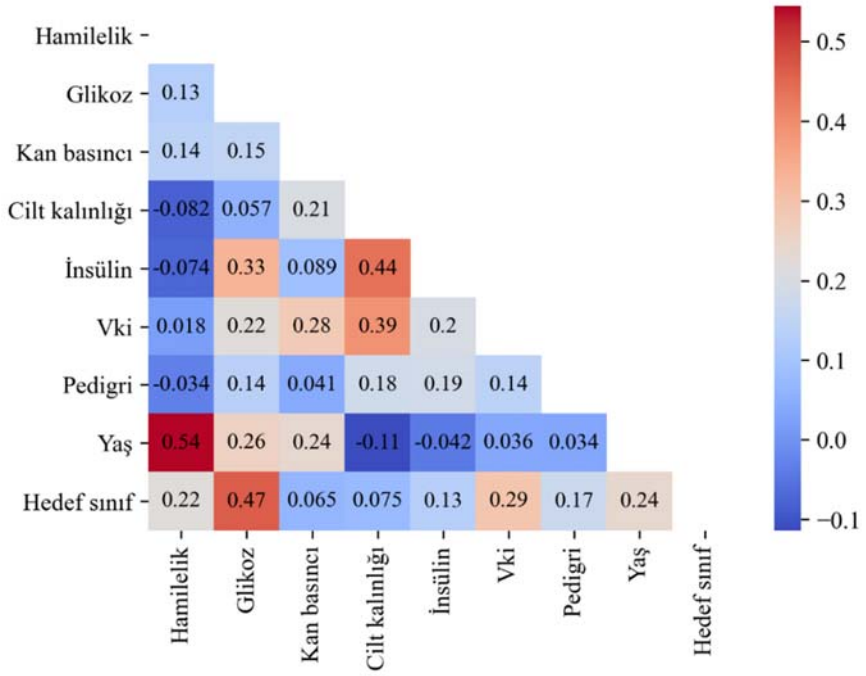
Şekil 1 incelendiğinde, glikoz seviyesi, vücut kitle indeksi (VKİ), yaş ve hamilelik sayısının, hedef sınıf ile daha yüksek korelasyona sahip olduğu görülmektedir. Hedef ile arasında 0,5'in üzerinde korelasyona sahip olan herhangi bir özellik bulunmadığından dolayı tahminleme konusunda tek başına baskın bir özellik yoktur.

Tablo 1. Veri setinde yer alan özellikler (Features of the dataset)

Özellik Adı	Birimi	Açıklama
Pregnancies (Hamilelik)	Adet	Hamile kalma sayısı
Glucose (Glikoz)	mg/dL	Plazmadaki glikoz konsantrasyonu
Blood Pressure (Kan basıncı)	mmHg	Diyastolik kan basıncı değeri (küçük tansiyon)
Skin Thickness (Cilt kalınlığı)	mm	Üç parçalı arka üst kol kası (triceps) cilt kalınlığı
Insulin (İnsulin)	µU/ml	2 saat serum insülin miktarı
BMI (Vücut kitle indeksi)	kg / $(m)^2$	Vücut kitle indeksi (VKİ)
Pedigree (Pedigri)	Sayısal	Diyabet pedigri fonksiyonu ile elde edilen değer
Age (Yaş)	Yıl	Yaş
Outcome (Hedef sınıf)	Kategorik	Sınıf etiketi (0 - sağlıklı, 1 - diyabet)

Tablo 2. Veri seti özelliklerine ait tanımlayıcı istatistikler (Descriptive statistics for data set features)

Özellik	Ortalama	Standart Sapma	En Küçük	En Büyük
Pregnancies (Hamilelik)	3,845052	3,369578	1	17
Glucose (Glikoz)	120,894531	31,972618	0	199
Blood Pressure (Kan basıncı)	69,105469	19,355807	0	122
Skin Thickness (Cilt kalınlığı)	20,536458	15,952218	0	99
Insulin (İnsulin)	79,799479	115,244002	0	846
BMI (Vücut kitle indeksi)	31,992578	7,884160	0	67,100000
Pedigree (Pedigri)	0,471876	0,331329	0,078000	2,420000
Age (Yaş)	33,240885	11,760232	21	81



Şekil 1. Korelasyon matrisi (Confusion matrix)

2.2. Kullanılan Sınıflandırma Yöntemleri (Classification Methods)

Sınıflandırma problemleri, mevcut veri seti üzerinden yapılan çıkarımlar doğrultusunda, yeni bir gözlemin hangi sınıfa ait olduğunun belirlenmesi ile ilgilidir. Sınıflandırma işlemleri için kullanılabilecek farklı yöntemler mevcuttur. Bu çalışmada diyabet hastalığı teşhisi için, literatürde sınıflandırma çalışmalarında sıklıkla kullanılan Destek Vektör Makinesi, Lojistik Regresyon, K-En Yakın Komşu, Rastgele Orman, AdaBoost ve Gradient Boosting sınıflandırma yöntemleri kullanılmıştır.

Destek vektör makinesi (DVM), Vapnik vd. tarafından literatüre kazandırılan, istatistiksel öğrenme teorisine dayalı, sınıflandırma ve regresyon analizi için kullanılan gözetimli bir makine öğrenmesi tekniğidir [39]. DVM, her biri iki kategoriden birine ait olarak işaretlenmiş eğitim veri setinden öğrenerek, yeni örnekleri bu iki sınıftan birine olasılıklı olmayacak şekilde atayan bir model oluşturur. Veri örneklerinin yer aldığı düzlemde, sınıfları birbirinden ayırmak için, iki sınıfın üyelerinden en uzak mesafede olacak şekilde bir karar sınırının çizilmesi sağlanır. DVM'nin, aşırı uydurma sorununa yatkınlığının az olması ve yüksek doğruluk sağlaması kullanım yaygınlığını arttırmaktadır.

Lojistik Regresyon (LR), bağımlı değişkenin süreksiz olduğu ikili sınıflandırma problemlerinde kullanılan istatistiksel bir modeldir. Bilgisayar bilimi, pek çok uygulamalı bilimde ve gerçek dünya

problemlerinde yaygın olarak kullanılmaktadır. Lojistik regresyon ikili bağımlı değişken ile bir dizi bağımsız değişken arasındaki ilişkiyi açıklamaya yönelik tahminleyici bir analizdir. Bu işi yerine getirmek için bir lojistik fonksiyon (logit fonksiyonu) kullanır. Bir olayın gerçekleşme olasılığı Eş. 1'deki formül ile ifade edilir. Olayın gerçekleşmeme olasılığı 1-p olmak üzere logit fonksiyonu ise Eş. 2'ye göre hesaplanır.

$$p = \frac{e^{a+bx}}{1+e^{a+bx}} \quad (1)$$

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \quad (2)$$

K-En Yakın Komşu (KNN) algoritması, literatüre Fix ve Hodges tarafından kazandırılan, sınıflandırma ve regresyonda yaygın olarak kullanılan, parametrik olmayan yöntemidir [40]. KNN sınıflandırmada, çıktı bir sınıfın üyelik değeridir. Sınıfı belirlenmek istenen nokta, K adet en yakın komşusuna bakılarak en yaygın olan sınıfa atanır. Sınıfı tahmin edilecek her bir örnek için, veri setindeki tüm örnekler arasında en yakın komşuluğun aranması nedeniyle, veri setinin büyümesi halinde işlem yükü artmaktadır. KNN algoritması mesafeye dayalı olduğundan, eğitim verilerinin normalizasyonu sınıflandırıcının doğruluğunu önemli ölçüde yükseltebilir.

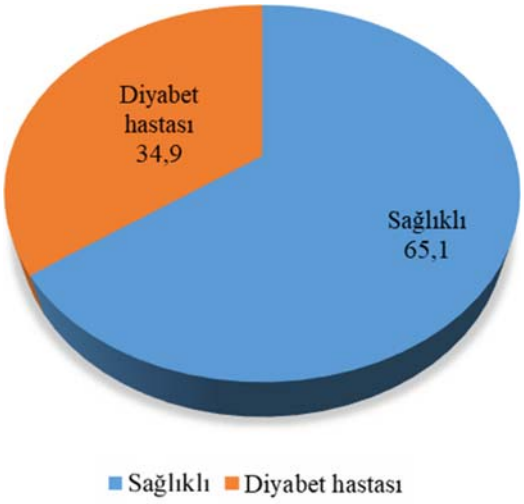
Rastgele Orman (RO), eğitim esnasında çok sayıda karar ağacı oluşturarak, her bir ağacın ürettiği sonuçların modu veya ortalamasını alarak çıktı sınıfı belirleyen bir kolektif öğrenme algoritmasıdır. Ho [41] tarafından oluşturulan yönteme dayanan RO, daha sonra Breiman [42] tarafından geliştirilerek literatüre kazandırılmıştır. RO, geleneksel karar ağaçlarında yaygın olan aşırı uydurma problemine hem veri seti hem özellikleri çok sayıda parçaya bölüp birden çok ağaç üzerinde işleme yoluyla çözüm getirir.

AdaBoost (AB), Freund ve Schapire [43] tarafından formüle edilen adaptif bir meta algoritmadır. Bireysel öğrenicilerin ve kararlarının birleştirilmesi mantığına dayanan kolektif bir öğrenme yöntemidir. Eğitim sürecinde bireysel öğrenicilerin durumları ağırlıklandırılarak, yapılan güncellemelerle nihai modelin güçlü bir öğrenmeye yakınsaması sağlanır. AB, kaynak tüketiminin etkin ve tahmin hızının yüksek olması nedeni ile kolektif modeller içerisinde yaygın olarak tercih edilir.

Gradient Boosting (GB), AB yöntemine benzer şekilde bireysel zayıf öğrenicilerin güçlü öğreniciler haline getirilmesi mantığı ile çalışır. Pek çok modeli aşamalı, eklemeli ve sıralı bir şekilde eğitir. AB ve GB arasındaki en temel fark, zayıf öğrenicilerin eksiklerinin nasıl belirlendiğidir. AB modeli, yüksek ağırlıklı veri noktalarını kullanarak eksiklikleri tespit ederken, GB ise bu işi kayıp fonksiyonundaki gradyanları kullanarak gerçekleştirir.

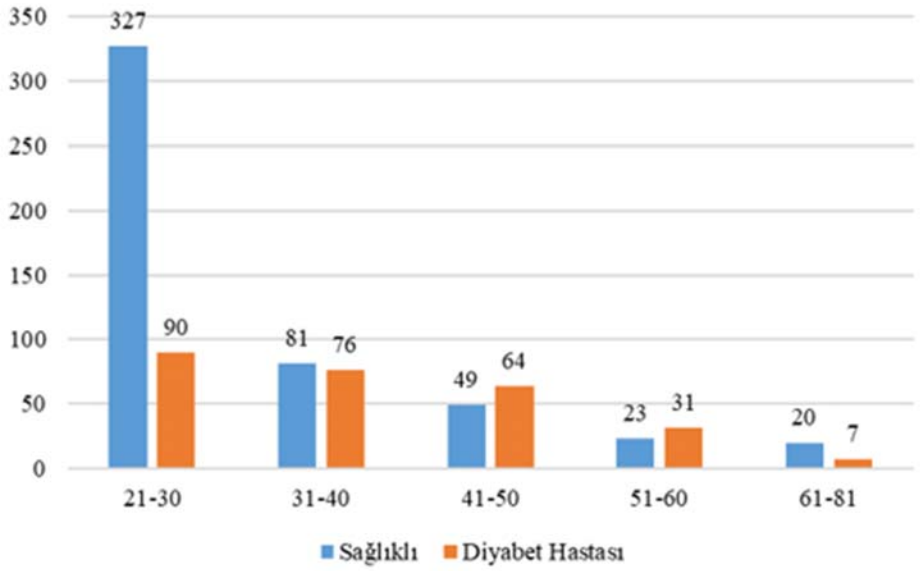
3. Diyabet Teşhisine Yönelik Sınıflandırma
(Classification for Diagnosis of Diabetes)

8 özellik ve 1 sınıf değeri ile tanımlanan, 768 örnekten oluşan veri setinde yer alan özelliklerin tümü sayısal değerlerdir. Sınıf değeri ise sağlıklı (0) ve diyabet hastası (1) olmak üzere sayısal-kategorik türdedir. Veri setindeki hasta ve sağlıklı birey dağılımları Şekil 2'de gösterilmiştir.



Şekil 2. Veri setindeki diyabetli ve sağlıklı kayıtların dağılımları (Distribution of diabetic and healthy records in the dataset)

Sağlıklı Veri setindeki kayıtların yaşlara göre hastalık dağılım grafiği ise Şekil 3'te verilmiştir. Veriler incelendiğinde diyabet hastalığı



Şekil 3. Yaşlara göre hastalık dağılımları (Disease distributions by age)

sayıca daha çok 21 ile 50 yaş aralığında görülmektedir. 41-50 ile 51-60 yaş aralığında ise diyabetli bireyler, sağlıklı bireylere oranla %130 seviyelerindedir. Veri setindeki hasta ve sağlıklı veri dağılımlarında dengeli bir durum olmadığı görülmektedir. Sağlıklı bireylere ait kayıtlar %65'in üzerinde bir orana sahip olduğundan, kullanılan sınıflandırma modellerinin bu sınıfı aşırı öğrenmeye daha yatkınlaşması muhtemeldir. Bu duruma çözüm getirmek için her bir sınıflandırıcı ile birlikte yeniden örnekleme teknikleri kullanılmıştır. Yeniden örnekleme, az örnekleme ve fazla örnekleme olmak üzere iki şekilde uygulanabilmektedir. Bu çalışmada fazla örnekleme için SMOTE [44], KMeansSMOTE [45], RandomOverSampler [46], ADASYN [47], BorderlineSMOTE [48] ve SVMSMOTE [49]; az örnekleme için EditedNearestNeighbours [50], AllKNN [51], InstanceHardnessThreshold [52], NearMiss [53], NeighbourhoodCleaningRule [54], OneSidedSelection [55], RandomUnderSampler [56] ve TomekLinks [57] teknikleri uygulanmıştır. Veri seti üzerinde farklı sınıflandırıcıların her biri için bahsi geçen yeniden örnekleme yöntemleri ayrı ayrı uygulanarak performans ölçümleri gerçekleştirilmiştir.

3.1. Performans Metrikleri (Performance Metrics)

Bir sınıflandırıcının başarısını ölçmek için Tablo 3'te verilen karmaşıklık matrisinden elde edilen değerlerden yararlanılır. Karmaşıklık matrisinden elde edilen değerler kullanılarak farklı başarı metrikleri üretilir. Bu çalışmada kullanılan metrikler Tablo 4'te verilmiştir.

Doğruluk metriği modelin genel başarısını ifade eder. Kesinlik modelin pozitif olarak sınıflandırdığı örneklerdeki isabet oranıdır. Duyarlılık ise gerçek pozitif değerlerden kaçının doğru şekilde belirlendiğini gösterir. F1 skoru, kesinlik ve duyarlılık arasındaki dengeyi ifade eder. Bu metrikler yanında alıcı işlem karakteristiği (Receiver Operating Characteristic - ROC) eğrileri, farklı sınıflar için bir olasılık eğrisidir. X ekseninde yanlış pozitif oranı, Y ekseninde ise doğru pozitif oranın yer aldığı bu eğri, kullanılan sınıflandırıcının tahminde ne kadar iyi olduğunu açıklar. Eğri altında kalan alan (Area Under the Curve - AUC), [0,1] aralığında değer alır ve model performansının bir özeti kabul edilir (Şekil 4). AUC değerinin 1'e yaklaşması veri setindeki sınıfların daha başarılı şekilde ayırt edilebildiğini gösterir.

4. Sonuçlar ve Tartışmalar (Results and Discussions)

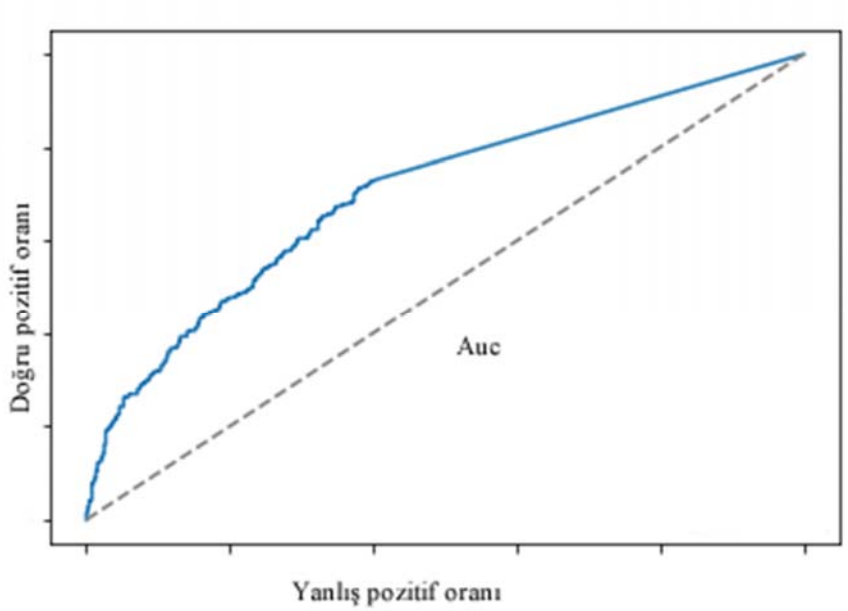
Çalışmada kullanılan veri seti dengesiz bir dağılıma sahiptir. PIMA veri seti üzerinde DVM, LR, KNN, RO, AB ve GB sınıflandırıcıları ve her bir sınıflandırıcı için Tablo 4'te adı geçen yeniden örnekleme teknikleri ayrı ayrı uygulanarak performans ölçümleri

Tablo 3. Karmaşıklık matrisi (Confusion matrix)

		Gerçek sınıf	
		Hastalık var	Hastalık yok
Tahmin edilen sınıf	*Hastalık var*	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
	Hastalık yok	Yanlış Negatif (YN)	Doğru Negatif (DN)

Tablo 4. Performans metrikleri (Performance metrics)

Metrik	Matematiksel ifadesi
Doğruluk	(DP + DN) / (DP + YP + YN + DN)
Kesinlik	DP / (DP + YP)
Duyarlılık	DP / (DP + YN)
F1 Skoru	2 * kesinlik * duyarlılık / (kesinlik + duyarlılık)



Şekil 4. ROC eğrisi ve AUC (ROC curve and AUC)

gerçekleştirilmiştir. Her bir sınıflandırıcının hiperparametreleri ızgara arama yöntemi ile belirlenmiştir. Her bir yöntem için kullanılan parametreler ve değerleri Tablo 5'te verilmiştir.

Yeniden örnekleme yapılarak ve yapılmadan gerçekleştirilen her bir sınıflandırma işleminin sonucu, Tablo 4'te bahsi geçen metrikler ile ölçümlenip, raporlanmıştır. Sınıflandırma işlemlerinde 5 kat çapraz doğrulama uygulanmış ve elde edilen sonuçların ortalama değerleri verilmiştir. Örnekleme tekniklerinin sınıflandırıcılar üzerindeki etkisini gösteren sonuçlar DVM için Tablo 6'da, LR için Tablo 7'de, KNN için Tablo 8'de, RO için Tablo 9'da, AB için Tablo 10'da ve GB için Tablo 11'de yer almaktadır.

Tablolarda yer alan ölçümler değerlendirildiğinde, bazı yeniden örnekleme tekniklerinin belirli sınıflandırıcıların başarısını örnekleme olmayan duruma göre düşürdüğü görülse de genel anlamda değerlendirildiğinde yeniden örnekleme tekniklerinin başarıyı arttırdığı gözlenmiştir. Performans metrikleri açısından yeniden örnekleme teknikleri ile %14'e kadar başarı artışlarının olduğu görülmektedir. Az ve fazla örnekleme tekniklerinin başarı durumları değişkenlik göstermekle birlikte, bu çalışmada az örnekleme tekniklerinin daha başarılı sonuç verdiği görülmüştür. Ayrıca uygulanan fazla örnekleme teknikleri içerisinde KMeansSMOTE, az örnekleme teknikleri içerisinde ise InstanceHardnessThreshold tekniğinin genel olarak daha yüksek başarı artışı sağladığı görülmüştür. Kullanılan tüm sınıflandırıcılar ve örnekleme yöntemlerinin en iyi sonuçları özet olarak Tablo 12'de verilmiştir.

Tablo 12'deki sonuçlar incelendiğinde kolektif öğrenme yöntemleri olan Rastgele Orman, Gradient Boosting ve AdaBoost yeniden örnekleme ile birleşince en yüksek skorların elde edildiği görülmüştür. Rastgele Orman sınıflandırıcı ile tüm parametreler

Tablo 5. Sınıflandırıcılar için kullanılan parametreler ve değerleri (Parameters used for the classifiers and their values)

Sınıflandırıcı	Kullanılan parametre ve değeri
DVM	kernel='rbf', C=2
KNN	n_neighbors=13
RO	n_estimators =100
LR	max_iter=250
AB	n_estimators=100
GB	n_estimators=100, learning_rate=1.0, max_depth=1

Tablo 6. Destek Vektör Makinesi için sınıflandırma sonuçları (Classification results for Support Vector Machine)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0,797386	0,789474	0,566038	0,659341	0,743019
	SMOTE	0,775	0,802198	0,73	0,764398	0,775
	KMeansSMOTE	*0,935*	0,884956	*1,0*	*0,938967*	*0,935*
Fazla	RandomOverSampler	0,785	0,806452	0,75	0,777202	0,785
Örnekleme	ADASYN	0,758974	0,744898	0,768421	0,756477	0,759211
	BorderlineSMOTE	0,76	0,720339	0,85	0,779817	0,76
	SVMSMOTE	0,83	0,757812	0,97	0,735135	0,755
	EditedNearestNeighbours	0,910891	0,907407	0,924528	0,915888	0,910181
	AllKNN	0,924731	0,925926	0,943396	0,934579	0,921698
	InstanceHardnessThreshold	0,934579	*0,96*	0,905660	0,932039	0,934312
Az	NearMiss	0,757009	0,8	0,679245	0,734694	0,756289
Örnekleme	NeighbourhoodCleaningRule	0,894231	0,905660	0,888889	0,897196	0,894444
	OneSidedSelection	0,823944	0,888889	0,603774	0,719101	0,779415
	RandomUnderSampler	0,794393	0,833333	0,740741	0,784314	0,794899
	TomekLinks	0,816901	0,885714	0,584906	0,704545	0,769981

Tablo 7. Lojistik Regresyon için sınıflandırma sonuçları (Classification results for Logistic Regression)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0,810458	0,785714	0,622642	0,694737	0,766321
	SMOTE	0,805	0,827957	0,77	0,797927	0,805
	KMeansSMOTE	0,935	0,884956	*1,0*	0,938967	0,935
Fazla	RandomOverSampler	0,825	0,828283	0,82	0,824121	0,825
Örnekleme	ADASYN	0,74359	0,710280	0,8	0,752475	0,745
	BorderlineSMOTE	0,755	0,742857	0,78	0,760976	0,755
	SVMSMOTE	0,855	0,808696	0,93	0,865116	0,855
	EditedNearestNeighbours	0,900990	0,905660	0,905660	0,905660	0,900747
	AllKNN	0,892473	0,892857	0,925926	0,909091	0,886040
	InstanceHardnessThreshold	*0,944444*	*0,979592*	0,905660	*0,941176*	*0,943739*
Az	NearMiss	0,738318	0,719298	0,773585	0,745455	0,738644
Örnekleme	NeighbourhoodCleaningRule	0,864078	0,854545	0,886792	0,870370	0,863396
	OneSidedSelection	0,830986	0,837209	0,679245	0,75	0,800297
	RandomUnderSampler	0,794393	0,796296	0,796296	0,796296	0,794375
	TomekLinks	0,830986	0,837209	0,679245	0,75	0,800297

Tablo 8. K-En Yakın Komşu için sınıflandırma sonuçları (Classification results for K Nearest Neighbor)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0,823529	0,825	0,622642	0,709677	0,776321
	SMOTE	0,825	0,787611	0,89	0,835681	0,825
	KMeansSMOTE	0,91	0,847458	*1,0*	0,917431	0,91
Fazla	RandomOverSampler	0,815	0,794393	0,85	0,821256	0,815
Örnekleme	ADASYN	0,784615	0,715447	0,926316	0,807339	0,788158
	BorderlineSMOTE	0,8	0,727273	0,96	0,827586	0,8
	SVMSMOTE	0,81	0,787037	0,85	0,817308	0,81
	EditedNearestNeighbours	0,872549	0,886792	0,870370	0,878505	0,872685
	AllKNN	*0,913978*	*0,941176*	0,905660	*0,923077*	*0,915330*
	InstanceHardnessThreshold	0,887850	0,901961	0,867925	0,884615	0,887666
Az	NearMiss	0,728972	0,777778	0,648148	0,707071	0,729734
Örnekleme	NeighbourhoodCleaningRule	0,884615	0,903846	0,870370	0,886792	0,885185
	OneSidedSelection	0,845070	0,860465	0,698113	0,770833	0,815349
	RandomUnderSampler	0,78504	0,767857	0,811321	0,788991	0,785290
	TomekLinks	0,845070	0,860465	0,698113	0,770833	0,815349

Tablo 9. Rastgele Orman için sınıflandırma sonuçları (Classification results for Random Forest)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0,843137	0,837209	0,679245	0,75	0,804623
	SMOTE	0,925	0,897196	0,96	0,927536	0,925
	KMeansSMOTE	0,95	0,909091	*1,0*	0,952381	0,95
Fazla	RandomOverSampler	0,945	0,900901	*1,0*	0,947867	0,945
Örnekleme	ADASYN	0,882051	0,852941	0,915789	0,883249	0,882895
	BorderlineSMOTE	0,92	0,881818	0,97	0,923810	0,92
	SVMSMOTE	0,925	0,889908	0,97	0,928230	0,925
	EditedNearestNeighbours	0,920792	0,924528	0,924528	0,924528	0,920597
	AllKNN	0,925532	0,943396	0,925926	0,934579	0,925463
	InstanceHardnessThreshold	*0,962963*	*0,980769*	0,944444	*0,962264*	*0,962963*
Az	NearMiss	0,794393	0,775862	0,833333	0,803571	0,794025
Örnekleme	NeighbourhoodCleaningRule	0,923077	0,960000	0,888889	0,923077	0,924444
	OneSidedSelection	0,866197	0,857143	0,777778	0,815534	0,849116
	RandomUnderSampler	0,813084	0,793103	0,851852	0,821429	0,812718
	TomekLinks	0,859155	0,836735	0,773585	0,803922	0,841849

Tablo 10. AdaBoost için sınıflandırma sonuçları (Classification results for AdaBoost)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0,830065	0,8	0,679245	0,734694	0,794623
	SMOTE	0,86	0,839623	0,89	0,864078	0,86
	KMeansSMOTE	0,93	0,890909	*0,98*	0,933333	0,93
Fazla	RandomOverSampler	0,895	0,876190	0,92	0,897561	0,895
Örnekleme	ADASYN	0,815385	0,831461	0,778947	0,804348	0,814474
	BorderlineSMOTE	0,845	0,828571	0,87	0,848780	0,845
	SVMSMOTE	0,83	0,784483	0,91	0,842593	0,83
	EditedNearestNeighbours	0,871287	0,870370	0,886792	0,878505	0,870480
	AllKNN	0,935484	0,944444	0,944444	*0,944444*	0,933761
	InstanceHardnessThreshold	*0,944954*	*0,960784*	0,924528	0,942308	*0,944407*
Az	NearMiss	0,757009	0,787234	0,698113	0,74	0,756464
Örnekleme	NeighbourhoodCleaningRule	0,893204	0,903846	0,886792	0,895238	0,893396
	OneSidedSelection	0,836879	0,829787	0,722222	0,772277	0,815134
	RandomUnderSampler	0,794393	0,807692	0,777778	0,792453	0,794549
	TomekLinks	0,823944	0,833333	0,660377	0,736842	0,790863

açısından en yüksek skorların elde edildiği görülmektedir. Bu sınıflandırıcı ile elde edilen en yüksek doğruluk değeri ise InstanceHardnessThreshold az örnekleme tekniği ile kullanıldığı durumda sağlanmıştır. Rastgele Orman ile tüm sınıflandırmalarda elde edilen en iyi sonuçlar doğruluk için %96,29, kesinlik için %98,07, duyarlılık için %100, F1 Skoru için %96,22 ve AUC için %96,29'dur. Rastgele Orman sınıflandırıcıdan sonra en yüksek ölçümler Gradient Boosting, ardından AdaBoost ile elde edilmiştir. %95,32 ve %94,49 olarak elde edilen doğruluk değerleri yine InstanceHardnessThreshold az örnekleme tekniği ile sağlanmıştır. Yeniden örnekleme tekniği kullanılarak, bu çalışmada test edilen sınıflandırıcılar içerisinde en düşük doğruluk %91,39 olarak K-En

Tablo 11. Gradient Boosting için sınıflandırma sonuçları (Classification results for Gradient Boosting)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0,823529	0,809524	0,641509	0,715789	0,780755
Fazla Örnekleme	SMOTE	0,85	0,843137	0,86	0,851485	0,85
	KMeansSMOTE	0,93	0,883929	*0,99*	0,933962	0,93
	RandomOverSampler	0,86	0,846154	0,88	0,862745	0,86
	ADASYN	0,784615	0,752381	0,831579	0,79	0,785789
	BorderlineSMOTE	0,815	0,805825	0,83	0,817734	0,815
	SVMSMOTE	0,82	0,813725	0,83	0,821782	0,82
Az Örnekleme	EditedNearestNeighbours	0,862745	0,870370	0,870370	0,870370	0,862269
	AllKNN	0,913978	0,910714	0,944444	0,927273	0,908120
	InstanceHardnessThreshold	*0,953271*	*0,944444*	0,962264	*0,953271*	*0,953354*
	NearMiss	0,747664	0,754717	0,740741	0,747664	0,747729
	NeighbourhoodCleaningRule	0,893204	0,875	0,924528	0,899083	0,892264
	OneSidedSelection	0,836879	0,860465	0,685185	0,762887	0,808110
	RandomUnderSampler	0,803738	0,796296	0,811321	0,803738	0,803809
	TomekLinks	0,823944	0,791667	0,716981	0,752475	0,802311

Tablo 12. Elde edilen en iyi sonuçlar (The best results obtained)

Sınıflandırıcı	Örnekleme Yöntemi	En İyi Doğruluk	En İyi Kesinlik	En İyi Duyarlılık	En İyi F1 Skoru	En İyi AUC
Destek Vektör Makinesi	Örnekleme yok	0,797386	0,789474	0,566038	0,659341	0,743019
	Fazla Örnekleme	*0,935*	0,884956	*1,0*	*0,938967*	*0,935*
	Az Örnekleme	0,934579	*0,96*	0,943396	0,934579	0,934312
Lojistik Regresyon	Örnekleme yok	0,810458	0,785714	0,622642	0,694737	0,766321
	Fazla Örnekleme	0,935	0,884956	*1,0*	0,938967	0,935
	Az Örnekleme	*0,944444*	*0,979592*	0,925926	*0,941176*	*0,943739*
K-En Yakın Komşu	Örnekleme yok	0,823529	0,825	0,622642	0,709677	0,776321
	Fazla Örnekleme	0,91	0,847458	*1,0*	0,917431	0,91
	Az Örnekleme	*0,913978*	*0,941176*	0,905660	*0,923077*	*0,915330*
Rastgele Orman	Örnekleme yok	0,843137	0,837209	0,679245	0,75	0,804623
	Fazla Örnekleme	0,95	0,909091	*1,0*	0,952381	0,95
	Az Örnekleme	*0,962963*	*0,980769*	0,944444	*0,962264*	*0,962963*
AdaBoost	Örnekleme yok	0,830065	0,8	0,679245	0,734694	0,794623
	Fazla Örnekleme	0,93	0,890909	*0,98*	0,933333	0,93
	Az Örnekleme	*0,944954*	*0,960784*	0,944444	*0,944444*	*0,944407*
Gradient Boosting	Örnekleme yok	0,823529	0,809524	0,641509	0,715789	0,780755
	Fazla Örnekleme	0,93	0,883929	*0,99*	0,933962	0,93
	Az Örnekleme	*0,953271*	*0,944444*	0,962264	*0,953271*	*0,953354*

yakın Komşu sınıflandırıcı ile elde edilmiştir. Tüm sınıflandırıcılar için kesinlik değerleri %94 ile %98 arasında seyretmektedir. Duyarlılık değerleri ise genel olarak %98, %99 ve %100 seviyesindedir. Bu durum tüm sınıflandırıcıların doğru pozitifleri tespit ve yanlışları eleme kabiliyetlerinin son derece yüksek olduğunu göstermektedir.

Bu çalışmada elde edilen sonuçlar ile literatürde aynı veri setini kullanan ve çeşitli makine öğrenmesi teknikleri ile sınıflandırmalar gerçekleştiren son dönemdeki diğer çalışmaların sonuçları karşılaştırılmalı olarak Tablo 13'te verilmiştir.

Bu çalışmada, uygulanan yeniden örnekleme teknikleri ile birlikte kullanılan DVM, LR, KNN, RO, AB ve GB sınıflandırıcılarının, literatürdeki diğer çalışmalarda kullanılan emsal yöntemden daha yüksek başarı sağladığı görülmüştür. Tablo 13'te yer alan diğer çalışmalarda elde edilen en yüksek doğruluk değeri %94,42, bunun ardından %94,10 ve %94'tür. Tabloda verilen 23 çalışmanın ortalama başarısı ise %83,65'tir. Bu çalışmada elde edilen en düşük doğruluk değeri %91,39, önerilen altı modelin ortalama doğruluğu ise %94,24'tür. Bu çalışmada uygulanan modeller ile elde edilen doğruluk değerlerinin geneli, literatürdeki diğer çalışmalarda elde edilen doğruluk değerlerinden daha yüksektir. Bu çalışmada en yüksek doğruluğu sağlayan Rastgele Orman ve bunu takip eden Gradient Boosting sınıflandırıcının her ikisinin de InstanceHardnessThreshold az örnekleme tekniği ile birlikte, Tablo 13'te verilen diğer çalışmaların tümünden daha yüksek doğruluk sağlandığı görülmektedir. Literatürdeki çalışmalarda Destek Vektör Makinesi yönteminin çoğu çalışmada daha yüksek doğruluk sağladığı ortaya konmuştur. Bu çalışmada ise uyguladığı kolektif yaklaşımdan

Tablo 13. Elde edilen bulguların literatürdeki çalışmalar ile karşılaştırılması (Comparison of the findings with the studies in the literature)

Referans	Kullanılan yöntem(ler)	En iyi doğruluk (%)
Agarwal ve Saxena, 2020 [28]	LR, DVM, NB, KA, KNN	81,16
Livington vd., 2020 [29]	Fuzzy sınıflandırıcı	83,00
Naz ve Ahuja, 2020 [30]	NB, YSA	90,34
Tigga ve Garg, 2020 [58]	RO	94,10
Kaur ve Kumari, 2020 [32]	DVM, KNN, YSA	89
Patil vd., 2020 [33]	DVM, KA, RO	77,05
Pranto vd., 2020 [34]	KNN, KA, RO, NB	77,9
Reddy vd., 2020 [35]	DVM, KNN, LR, NB, GB	87,61
Nusrat vd., 2020 [36]	KA, RO, GB	76,30
Varma ve Panda, 2019 [19]	DVM, KNN, RO	74,67
Radja ve Emanuel, 2019 [20]	NB, DVM, KA	77,3
Yahyaoui vd., 2019 [21]	DVM, RO	83,67
Birjais vd., 2019 [23]	NB, LR, GB	86
Wang vd., 2019 [24]	RO - ADASYN	87.1
Srivastava vd., 2019 [25]	YSA	92
Yuvaraj ve SriPreethaa, 2019 [26]	KA, NB, RO	94
Battineni vd. ,2019 [27]	LR, RO, NB	83
Köse, 2019 [37]	Zeki Optimizasyon-DVM	94,42
Sisodia, 2018 [14]	KA, DVM, NB	76,30
Zou vd., 2018 [15]	YSA, RO	76,67
Wei vd., 2018 [16]	LR, DVM, KA, NB	77,60
Kohli ve Arora, 2018 [17]	AB, KA, LR, RO, DVM	85,71
Mir ve Dhage, 2018 [18]	NB, DVM, RO	79,13
	Literatür Ortalaması	*83,65*
	Önerilen DVM	93,5
	Önerilen LR	94,44
	Önerilen KNN	91,39
	Önerilen RO	*96,29*
	Önerilen AB	94,49
	Önerilen GB	95,32
	Önerilen modellerin ortalaması	*94,24*

dolayı, aşırı uydurma sorununa daha az duyarlı Rastgele Orman modelinin ve bunun yanında diğer iki kolektif öğrenme modeli Gradient Boosting ve AdaBoost yöntemlerinin daha başarılı sonuç verdiği görülmüştür.

5. Sonuçlar (Conclusions)

Makine öğrenmesi sağlık alanında elde edilen büyük veri kümeleri içerisinde yer alan kritik desenleri ortaya koyarak hastalıkların erken teşhisine olanak sağlamaktadır. Bu çalışmada dünya genelinde yaygın olarak görülen ve görülme sıklığı giderek artan diyabet hastalığının erken teşhisine yönelik makine öğrenmesi tekniklerini kullanan bir çözüm önerilmiştir. Diyabet tedavi edilmediği zaman farklı organ ve dokularda ciddi hasar ve ölümcül sonuçlara neden olabilmektedir. Diyabet riskinin erken teşhisi durumunda ise önlenmesi ve oluşturacağı komplikasyonların azaltılması mümkündür. Bu çalışmada literatürde diyabet teşhisi konusunda yaygın olarak kullanılan PIMA veri seti üzerinde altı farklı makine öğrenmesi yöntemi ile sınıflandırmalar gerçekleştirilmiştir. PIMA veri seti dengeli bir yapıya sahip değildir. Bu nedenle kullanılan altı sınıflandırıcı ile birlikte on dört farklı yeniden örnekleme yöntemi kullanılmıştır. DVM, LR, KNN, RO, AB ve GB makine öğrenmesi modellerinin her birinin yeniden örnekleme yapılmadan ve on dört ayrı yeniden örnekleme işlemi kullanılarak elde edilen sınıflandırma başarıları karşılaştırılmıştır. Test edilen toplam doksan ayrı sınıflandırmanın doğruluk, kesinlik, duyarlılık, F1 skoru ve AUC olmak üzere beş metrik ile karakterize edilen dört yüz elli ayrı ölçümü karşılaştırıldığında yeniden örnekleme yöntemlerinin genelinin sınıflandırma başarısını arttırdığı gözlemlenmiştir. Bu çalışmada elde edilen ölçümler literatürde son yıllarda yapılan ve aynı veri seti üzerinde çeşitli makine öğrenmesi yöntemlerini uygulayan diğer çalışmalar ile karşılaştırılmıştır. Bu çalışmada, InstanceHardnessThreshold az örnekleme tekniği kullanılarak, RO sınıflandırıcı ile elde edilen %96,29 doğruluk ve yine aynı örnekleme tekniği ve GB sınıflandırıcı ile elde edilen %95,32 doğruluk literatürdeki diğer çalışmalardan %2 ile %22 arasında daha yüksek sonuç vermiştir. Veri setindeki sınıf dengesini sağlayan ve bu yolla aşırı uydurma sorununa çözüm üreten yeniden örneklemenin

sınıflandırıcı başarısını arttırıcı etkisi olduğu görülmüştür. Kolektif öğrenme algoritmaları ile yeniden örneklemenin birleşmesi daha yüksek bir başarı artışı getirmiştir. Bu çalışmada kullanılan fazla örnekleme tekniklerinden KMeansSMOTE ve az örnekleme tekniklerinden InstanceHardnessThreshold daha yüksek başarı artışı sağlamıştır. Ayrıca genel olarak az örnekleme yöntemlerinin daha çok başarı artışı sağladığı görülmüştür.

Veri üzerinden öğrenen makine öğrenmesi algoritmalarının başarılarını belirleyen temel etken üzerinde çalışılan veridir. Veri setindeki sınıf dengesizlikleri modelin başarısını, sonuçların genellenebilirliğini olumsuz etkiler. Eldeki veri setinin uygun örnekleme teknikleri kullanılarak dengeli hale getirilmesi kullanılan modelin daha başarılı sınıflandırmalar yapmasına olanak tanımaktadır. Bu çalışmada elde edilen sonuçlar yeninden örnekleme tekniklerinin performansları hakkında bir görüş oluşturmuştur. Ayrıca farklı makine öğrenmesi modellerinin başarılarını karşılaştırma olanağı tanımıştır. Gelecek çalışmalarda, farklı gerçek dünya problemleri üzerinde bu çalışmadan edilen tecrübeler doğrultusunda çözümler üretilmesi amaçlanmaktadır.

Kaynaklar (References)

1. Bilgehan B., Kavalcıoğlu C., Continuous wavelet transform (CWT) based filtering method for treating type 1 diabetes with continuous glucose monitoring (CGM) signals, Journal of the Faculty of Engineering and Architecture of Gazi University, 35 (2), 581–594, 2019.
2. Cho N.H., Shaw J.E., Karuranga S., Huang Y., Rocha Fernandes J.D., Ohlrogge A.W., Malanda B., IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045, Diabetes Research and Clinical Practice, 138 (1), 271–281, 2018.
3. Roglic G., WHO Global report on diabetes: A summary, International Journal of Noncommunicable Diseases, 1(1), 3-8, 2016.
4. American Diabet Association, Diagnosis and classification of diabetes mellitus, Diabetes Care, 32 (1), 62–67, 2009.
5. Swapna G., Vinayakumar R., Soman K.P., Diabetes detection using deep learning algorithms, ICT Express, 4(4), 243–246, 2018.
6. Palaniappan S., Awang R., Intelligent heart disease prediction system using data mining techniques, IEEE/ACS International Conference on Computer Systems and Applications, Doha, Qatar, 108–115, March 31-April 4, 2008.
7. Kavakiotis I., Tsave O., Salifoglou A., Maglaveras N., Vlahavas I., Chouvarda I., Machine learning and data mining methods in diabetes research, Computational and Structural Biotechnology Journal, 15 (1), 104–116, 2017.
8. Lai H., Huang H., Keshavjee K., Guergachi A., Gao X., Predictive models for diabetes mellitus using machine learning techniques, BMC Endocrine Disorders, 19 (1), 2-9, 2019.
9. Kopitar L., Kocbek P., Cilar L., Sheikh A., Stiglic G., Early detection of type 2 diabetes mellitus using machine learning-based prediction models, Scientific Reports, 10 (1),1-12, 2020.
10. Maniruzzaman M., Rahman M.J., Ahammed B., Abedin M.M., Classification and prediction of diabetes disease using machine learning paradigm, Health Information Science and Systems, 8 (1), 1–14, 2020.
11. Zhang L., Wang Y., Niu M., Wang C., Wang Z., Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study, Scientific Reports, 10 (1), 1–10, 2020.
12. Güler İ., Übeyli E.D., Diabetes Diagnosis by Multilayer Perceptron Neural Networks, Journal of the Faculty of Engineering and Architecture of Gazi University, 21 (2), 319–326, 2006.
13. Muhammad L., Algehyne E.A., Usman S.S., Predictive supervised machine learning models for diabetes mellitus, SN Computer Science, 1 (5), 1–10, 2020.
14. Sisodia D., Sisodia D.S., Prediction of Diabetes using Classification Algorithms, Procedia Computer Science, 132(1), 1578–1585, 2018.
15. Zou Q., Qu K., Luo Y., Yin D., Ju Y., Tang H., Predicting Diabetes Mellitus With Machine Learning Techniques, Front Genet, 9 (1), 515–515, 2018.
16. Wei S., Zhao X., Miao C., A comprehensive exploration to the machine learning techniques for diabetes identification, IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, 291–295, February 5-8, 2018.
17. Kohli P.S., Arora S., Application of Machine Learning in Disease Prediction, 4th International Conference on Computing Communication and Automation (ICCCA), India, 1–4, December 14-15, 2018.
18. Mir A., Dhage S.N., Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare, Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), India, 1–6, August 16-18, 2018.
19. Varma K.M., Panda D., Comparative analysis of Predicting Diabetes Using Machine Learning Techniques, Journal of Emerging Technologies and Innovative Research, 6 (6), 522–530, 2019.
20. Radja M., Emanuel A.W.R., Performance Evaluation of Supervised Machine Learning Algorithms Using Different Data Set Sizes for Diabetes Prediction, 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia, 252–258, October 23-24, 2019.
21. Yahyaoui A., Jamil A., Rasheed J., Yesiltepe M., A decision support system for diabetes prediction using machine learning and deep learning techniques, 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 1–4, November 6-7, 2019.
22. Benbelkacem S., Atmani B., Random Forests for Diabetes Diagnosis, International Conference on Computer and Information Sciences (ICCIS), Aljouf, Saudi Arabia, 1–4, April 3-4, 2019.
23. Birjais R., Mourya A.K., Chauhan R., Kaur H., Prediction and diagnosis of future diabetes risk: a machine learning approach, SN Applied Sciences, 1 (9), 1–8, 2019.
24. Wang Q., Cao W., Guo J., Ren J., Cheng Y., Davis D.N., DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values, IEEE Access, 7 (1), 102232–102238, 2019.
25. Srivastava S., Sharma L., Sharma V., Kumar A., Darbari H., Prediction of diabetes using artificial neural network approach, Engineering Vibration, Communication and Information Processing, Editors: Ray K., Sharan S.N., Rawat S., Jain S.K., Srivastava S., Bandyopadhyay A., Springer, India, 679–687, 2019.
26. Yuvaraj N., SriPreethaa K., Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster, Cluster Computing, 22 (1), 1–9, 2019.
27. Battineni G., Sagaro G.G., Nalini C., Amenta F., Tayebati S.K., Comparative machine-learning approach: A follow-up study on type 2 diabetes predictions by cross-validation methods, Machines, 7 (4), 1-12, 2019.
28. Agarwal A., Saxena A., Comparing Machine Learning Algorithms to Predict Diabetes in Women and Visualize Factors Affecting It the Most—A Step Toward Better Health Care for Women, International Conference on Innovative Computing and Communications, New Delhi, India, 339–350, February 20-22, 2020.
29. Livington M., Sujihelen L., Senthilsingh C., Predictive Design to Analyze Diabetes using Machine Learning Classifier, Solid State Technology, 63 (5), 6862–6871, 2020.
30. Naz H., Ahuja S., Deep learning approach for diabetes prediction using PIMA Indian dataset, Journal of Diabetes & Metabolic Disorders, 19 (1), 391–403, 2020.
31. Hasan M.K., Alam M.A., Das D., Hossain E., Hasan M., Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers, IEEE Access, 8 (1), 76516–76531, 2020.
32. Kaur H., Kumari V., Predictive modelling and analytics for diabetes using a machine learning approach, Applied Computing and Informatics, 18 (2), 90-100, 2020.
33. Patil R., Majumder L., Jain M., Patil V., Diabetes Disease Prediction Using Machine Learning, International Journal of Research in Engineering, Science and Management, 3 (6), 292–295, 2020.
34. Pranto B., Mehnaz S.M., Mahid E.B., Sadman I.M., Rahman A., Momen S., Evaluating machine learning methods for predicting diabetes among female patients in bangladesh, Information, 11 (8), 374-393, 2020.
35. Reddy D.J., Mounika B., Sindhu S., Pranayteja T., Reddy N., Sri G.J., Swaraja K., Meenakshi K., Kora P., Predictive machine learning model for early detection and analysis of diabetes," Materials Today: Proceedings, 1-7, 2020.

36. Nusrat F., Uzbaş B., Baykan Ö.K., Prediction of Diabetes Mellitus by using Gradient Boosting Classification, European Journal of Science and Technology, Special Issue(ICCEES), 268–272, 2020.

37. Köse U., Diabetes Diagnosis with Intelligent Optimization Based Support Vector Machines, Journal of Polytechnic, 22 (3), 557–566, 2019.

38. UCI. Pima Indian Diabetes Dataset. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/index.php. Access date January 09, 2021.

39. Vapnik V., Golowich S.E., Smola A., Support vector method for function approximation, regression estimation, and signal processing, Advances in Neural Information Processing Systems, 281–287, 1997.

40. Fix E., Hodges J.L., Discriminatory analysis-nonparametric discrimination: Small sample performance, California University of Berkeley, 1952.

41. Ho T.K., Random decision forests, Proceedings of 3rd international conference on document analysis and recognition, Montreal, 278–282, August 14-16, 1995.

42. Breiman L., Random Forests, Machine Learning, 45 (1), 5–32, 2001.

43. Freund Y., Schapire R.E., Experiments with a new boosting algorithm, 13th International Conference on Machine Learning, Italy, 148–156, July 3-6, 1996.

44. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P., SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research, 16 (1), 321–357, 2002.

45. Last F., Douzas G., Bacao F., Oversampling for imbalanced learning based on k-means and smote, Information Sciences, 465, 1-20, 2018.

46. Drummond C., Holte R.C., C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, Workshop on learning from imbalanced datasets II, Washington DC, 1–8, August 21, 2003.

47. He H., Bai Y., Garcia E., Li S., ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, International Joint Conference on Neural Networks, Hong Kong, China, 1322–1328, June 1-8, 2008.

48. Han H., Wang W.Y., Mao B.H., Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, Advances in Intelligent Computing, Hefei, China, 878–887, August 23-26, 2005.

49. Nguyen H., Cooper E., Kamei K., Borderline over-sampling for imbalanced data classification, International Journal of Knowledge Engineering and Soft Data Paradigms, 3 (1), 4–21, 2011.

50. Wilson D.L., Asymptotic Properties of Nearest Neighbor Rules Using Edited Data, IEEE Transactions on Systems, Man, and Cybernetics, 2(3), 408–421, 1972.

51. Tomek I., An Experiment with the Edited Nearest-Neighbor Rule, IEEE Transactions on Systems, Man, and Cybernetics, 6 (6), 448–452, 1976.

52. Smith M.R., Martinez T., Giraud-Carrier C., An instance level analysis of data complexity, Machine Learning, 95 (2), 225–256, 2014.

53. Mani I., Zhang I., KNN approach to unbalanced data distributions: a case study involving information extraction, Workshop on learning from imbalanced datasets, Washington DC, 1-7, August 21, 2003.

54. Laurikkala J., Improving Identification of Difficult Small Classes by Balancing Class Distribution, 8th Conference on Artificial Intelligence in Medicine in Europe, Cascais, Portugal, 63–66, July 1-4, 2001.

55. Kubat M., Matwin S., Addressing the curse of imbalanced training sets: one-sided selection, Fourteenth International Conference on Machine Learning, San Francisco, USA, 179–186, July 8-12, 1997.

56. Prusa J., Khoshgoftaar T.M., Dittman D.J., Napolitano A., Using random undersampling to alleviate class imbalance on tweet sentiment data, IEEE international conference on information reuse and integration, San Francisco, USA, 197–202, August 13-15, 2015.

57. Tomek I., Two modifications of CNN, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 6 (1), 769–772, 1976.

58. Tigga N.P., Garg S., Prediction of Type 2 Diabetes using Machine Learning Classification Methods, Procedia Computer Science, 167 (1), 706–716, 2020