Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Diyabet Hastalığının Erken Aşamada Tahmin Edilmesi İçin Makine Öğrenme Algoritmalarının Performanslarının Karşılaştırılması[1]

Kemal AKYOL [a,*], Abdulkadir KARACI [a]

[a] *Bilgisayar Mühendisliği Bölümü, Mühendislik ve Mimarlık Fakültesi, Kastamonu Üniversitesi, Kastamonu, TÜRKİYE*

** Sorumlu yazarın e-posta adresi: kakyol@kastamonu.edu.tr*

DOI: 10.29130/dubited.1014508

ÖZ

Şeker hastalığı, kan şekerinde anormalliklere neden olan zararlı hastalıklardan biridir. Bu hastalığın erken teşhisi insan vücudunda oluşabilecek organ bozulmalarını engeller. Yapay zekâ tabanlı çalışmalar medikal alanda etkin bir şekilde gerçekleştirilmektedir. Makine öğrenmesine dayalı bilgisayar destekli uzman sistemler bu hastalığın erken teşhisi için oldukça faydalıdır. Bu çalışmadaki şeker hastalığı problemi, klasik bir denetimli ikili sınıflandırma problemidir. Bu verisetinde 16 öznitelik bulunmakta olup, 200'ü negatif örnek ve 320'si pozitif örnek olmak üzere toplam 520 örnek içermektedir. Önişlemden geçirilen veriseti üzerinde Rastgele Orman, Gradyan Arttırma, K-En Yakın Komşu, Derin Sinir Ağları ve son olarak da Oylama topluluk sınıflandırıcısı kullanılarak inşa edilen modellerin performansları dışarıda tutma ve 5-kat çapraz doğrulama senaryoları çerçevesinde analiz edilmiştir. Her iki senaryoda da, Oylama topluluğu sınıflandırıcısı, deneylerde en iyi performansı sundu. Buna göre, Oylama topluluğu sınıflandırıcısı, tutma tekniğiyle yapılan deneylerde %100'lük bir sınıflandırma doğruluğu ve 5 kat çapraz doğrulamalı deneylerde ortalama %97,31'lik bir sınıflandırma doğruluğu sundu. Sonuç olarak, Oylama topluluğu sınıflandırıcısı kullanılarak diyabeti gerçek zamanlı olarak erken teşhis eden bir uzman sistem tasarlanabilir.

Anahtar Kelimeler: *Şeker hastalığı, makine öğrenmesi, oylama topluluk sınıflandırıcısı*

Comparison of Performances of Machine Learning Algorithms for Predicting Diabetes Mellitus in Early Stage

ABSTRACT

Diabetes mellitus is one of the harmful diseases that cause abnormalities in blood sugar. Early diagnosis of this disease prevents organ deterioration that may occur in the human body. Artificial intelligence-based studies are carried out effectively in the medical field. Computer-aided expert systems based on machine learning are quite useful for the early detection of this disease. The diabetes mellitus problem in this study is a classical supervised binary classification problem. There are 16 attributes in this dataset also it includes a total of 520 samples, 200 of which are negative samples and 320 of which are positive samples. The performances of the models constructed using Random Forest, Gradient Boosting, K-Nearest Neighbour, Deep Neural Networks, and finally voting ensemble classifier are analyzed within the framework of hold-out and 5-fold cross-validation techniques on the dataset pre-processed. In both scenarios, the Voting ensemble classifier presented the best performance in experiments. Accordingly, the Voting ensemble classifier offered a classification accuracy of 100% in experiments with the hold out technique, and an average of 97.31% in experiments with 5-fold cross-validation. As a result, an expert system that early diagnoses diabetes in real-time can be designed by using the Voting ensemble classifier.

Keywords: *Diabetes mellitus, machine learning, voting ensemble classifier*

[1] ICAIAME 2021 konferansında sunulmuş olup, özet metin olarak basılmıştır.

Geliş: 26/10/2021, Düzeltme: 25/11/2021, Kabul: 30/11/2021

I. GİRİŞ

Vücudun kan şekerini yapamadığı bir durum [1,2] olan şeker hastalığı hiperglisemi ile karakterize edilen ve tanımlanan bir grup metabolik bozukluk olarak ifade edilir [3]. Şeker hastalığı olarak bilinen diyabet, kandaki anormal derecede yüksek glikoz seviyeleri ile ilişkili kronik bir hastalıktır. Diyabet, yetersiz insülin üretimi ve hücrelerin insülinin etkisine karşı yetersiz duyarlılığı gibi iki nedenden birine bağlıdır [4]. Bu hastalık, susuzluk, poliüri, görme bulanıklığı ve kilo kaybı gibi karakteristik belirtiler gösterebilmektedir [3,5]. Tedaviye mümkün olan en kısa sürede başlamak ve kronik hiperglisemi ile ilişkili riski azaltmak için erken tanı gereklidir. Poliüri, polidipsi, polifaji, açıklanamayan kilo kaybı, enfeksiyon semptomları ortaya çıktığında tanı kolaylıkla konur [6]. Maniruzzanan ve arkadaşlarına göre şeker hastalığı sessiz bir katildir. Bu hastalığın ana nedeni, glikoz gibi aşırı miktarda metabolitlerin varlığıdır. Yazarlar, 2014 yılında tüm dünyada yaklaşık 387 milyon diyabetik insan olduğunu ifade etmiştir. Dahası, Dünya Sağlık Örgütü'ne göre, bu rakamlar 2030 yılına kadar iki katından fazla olacaktır [7].

Tıbbi araştırmalar, diyabet patolojisinin son yıllarda durma eğiliminde olmadığını ve arttığını göstermiştir [8]. Dünya Sağlık Örgütü 1965 yılından bu yana şeker hastalığının sınıflandırılmasına ilişkin kılavuzları periyodik olarak güncellemekte ve yayınlamaktadır [9]. Diyabetin erken teşhisi, sağlıklı bir yaşam sağlamak için hayati önem taşımaktadır [10]. Diyabetin erkenden tespiti ve analizi, ölüm oranını bir dereceye kadar azaltabilir [11]. Kalıcı tedavisi olmayan bir hastalık olduğu için bu hastalığın erken teşhisi gereklidir [12]. Şeker hastalığında çeşitli faktörlerin rol oynaması ve insan hatasının olması bu hastalığın teşhisini karmaşık hale getirmektedir. Bir kan testi, hastalığın doğru teşhisi için yeterli bilgi sağlamaz [13]. Şeker hastalığının teşhisi, yönetimi ve diğer ilgili klinik yönetimleri açısından makine öğrenimi ve veri madenciliği yaklaşımlarına büyük ilgi duyulmaktadır [1].

Şeker hastalığı, kan şekerinde anormalliklere neden olan zararlı hastalıklardan biridir. Bu hastalığın erken teşhisi insan vücudunda oluşabilecek organ bozulmalarını engeller. Şeker hastalığının teşhisi, bir tıp doktoru tarafından manuel olarak veya otomatik bir cihazla yapılabilir. Ancak, genellikle diyabetin başlangıç aşamasındaki belirtiler o kadar düşüktür ki deneyimli bir doktor bile bunları tam olarak tanımlayamaz. Makine öğrenimi ve yapay zekâ alanındaki ilerlemeler sayesinde erken aşamada hastalık tespiti ve teşhisi manuel diyabet tanımaya kıyasla daha verimlidir [14–17]. Makine öğrenmesine dayalı bilgisayar destekli uzman sistemler bu hastalığın erken teşhisinde oldukça önemli bir yere sahiptir. Bu çalışmanın amacı, makine öğrenmesi algoritmalarını kullanarak diyabetin erken aşamada tanısını gerçekleştirmektir.

Literatürde makine öğrenimi ve yapay zekâ teknikleri aracılığıyla diyabeti otomatik algılama, teşhis ve kendi kendini yönetme konusunda birçok çalışma bulunmaktadır. Örneğin, Bala Manoj Kumar ve ark. [10] Pima diyabet veriseti üzerinde doğru tahmin için denetimsiz bir öğrenme yaklaşımı olan Derin Sinir Ağı (DSA) sınıflandırıcısı ve özellik seçimi için Ekstra Ağaçlar ve Rastgele Orman ile paketlenmiş Özellik Önem Modeli kullanmıştır. Model, %98.16 doğruluk ile diğer güncel yöntemlere kıyasla iyi performans elde etmiştir. Kumari ve ark. [1] Pima diyabet veriseti ve meme kanseri verisetinde hastalığın pozitif ve negatif olarak ikili sınıflandırması için yumuşak oylama sınıflandırıcılı Rastgele orman, lojistik regresyon ve Naive Bayes kullanmışlardır. Maniruzzaman ve ark. [7] diyabet hastalığının tespiti için Gauss süreç sınıflandırıcısının performansını makine öğrenmesinde iyi bilinen sınıflandırma teknikleriyle karşılaştırdılar. Gauss süreç sınıflandırıcısı tabanlı model diğerlerine kıyasla başarılı bir performans sunmuştur. Khaleel ve Al-Bakry [18], şeker hastalığının olup olmadığını tahmin etmek için Pima diyabet veriseti üzerinde Lojistik Regresyon, Naïve Bayes ve K-En Yakın Komşu algoritmalarının performanslarını karşılaştırmış ve Lojistik Regresyon algoritmasının diyabeti tahmin etmede diğer algoritmalara kıyasla daha verimli olduğunu göstermişlerdir. Mercaldo ve ark. [8] diyabet teşhisine yardımcı olmak ve hızlandırmak için diyabetten etkilenen hastaları Dünya Sağlık Örgütü kriterleri çerçevesinde seçilmiş bir dizi özelliği kullanarak sınıflandırabilen bir yöntem önermişlerdir. Howsalya ve ark. [4] diyabet teşhisi için Farthest First kümeleme ve Sıralı Minimal Optimizasyon algoritmalarının

entegre edildiği bir yaklaşımı önermişlerdir. Yazarların önerdikleri çalışma, diyabetik ve diyabetik olmayan örnekler içeren Pima diyabet veriseti üzerinde oldukça başarılı bir performans sunmuştur. Prabha [19], diyabetin kolay tespitini sağlamak için bileklik fotopletismografi sinyaline ve temel fizyolojik parametrelere dayalı bir diyabet tespit sistemi önermiştir. Sistemi geliştirmek için diyabet, prediyabet ve normal koşullara sahip 217 katılımcıdan oluşan bir veriseti elde etmiştir. Mel frekans cepstral katsayılarının elde edildiği verisetindeki öznitelik boyutunu azaltmak için bir hibrit özellik seçim yöntemini kullanmıştır. Reddy ve ark. [11] diyabeti maksimum doğrulukla tahmin eden bir model tasarlamak için Pima diyabet veriseti üzerinde çeşitli makine öğrenme sınıflandırıcılarının performanslarını incelemişlerdir. Viloria ve ark. [13] Kolombiyalı hastalarda belirtilen faktörlere dayalı olarak diyabeti olmayan, diyabete yatkınlığı olan ve diyabetli olmak üzere Destek Vektör Makinesi kullanarak sınıflandırma gerçekleştirmişlerdir. Zou ve ark. [20] Çin'in Luzhou kentindeki hastaneden elde edilen veriler üzerinde diyabeti tahmin etmek için Karar Ağacı, Rastgele Orman ve Sinir Ağı algoritmalarının performanslarını incelemişlerdir. Çalışmalarında sırasıyla boyut azaltma ve öznitelik seçimi yöntemleri olan temel bileşen analizi ve minimum fazlalık maksimum uygunluk düzeyi yöntemlerine kıyasla tüm özniteliklerin olduğu veriseti üzerinde en yüksek doğruluğu elde etmişlerdir. Lai ve ark. [21], 18 ile 90 yaşları arasındaki 13.309 Kanadalı hastanın en son kayıtlarının yanı sıra laboratuvar (yaş, cinsiyet, açlık kan şekeri, vücut kitle indeksi, yüksek yoğunluklu lipoprotein, trigliseritler, kan basıncı ve düşük yoğunluklu lipoprotein) verileri üzerinde oluşturdukları modellerin performanslarını incelemişlerdir. Buna göre, Gradyan Arttırma Makinesi ve Lojistik Regresyon modelleri Rastgele Orman ve Karar Ağacı ile oluşturulan modellere kıyasla daha başarılı olmuştur. Nahzat ve Yağanoğlu [22], en yüksek doğrulukla en iyi sınıflandırıcıyı bulmak için Pima diyabet veriseti üzerinde K-En Yakın Komşu, Rastgele Orman, Destek Vektör Makinesi, Yapay Sinir Ağı ve Karar Ağacı makine öğrenimi algoritmalarının performanslarını incelemişlerdir. Yazarları çalışmasına göre, diğer makine öğrenimi yöntemlerine kıyasla Rastgele Orman, diyabet tahmininde yüksek doğruluk sunmuştur. Khanam ve Foo Pima [12], diyabet veriseti üzerinde diyabet tahmini için çeşitli yöntemlerinin performanslarını karşılaştırmışlardır. İki gizli katmana sahip sinir ağları en başarılı performansı sunmuştur. Sneha ve Gangil [23], diyabetin erken tespitinde yer alan özelliklerin seçimine odaklanmışlardır. Önemli özelliklerin tespiti ve makine öğrenimi ile klinik sonuçlara en yakın sonucu sunan en uygun sınıflandırıcıyı bulmayı amaçlamışlardır. Yazarların çalışmasına göre, Karar ağacı ve Rastgele Orman, sırasıyla %98.20 ve %98.00 ile en yüksek özgüllüğe sahiptir. Naive Bayes, %82.30 ile en iyi doğruluğu sunmuştur.

Görüldüğü üzere literatürde halka açık Pima diyabet veriseti üzerinde gerçekleştirilmiş birçok çalışma bulunmaktadır. Bu çalışmada, diyabetin erken tanısı için hazırlanmış ve güncel olan halka açık diyabet hastalığı veriseti üzerinde sınıflandırıcı algoritmaların performansları ele alınmıştır. Diyabet verilerinin sınıflandırması için yürütülen deneysel çalışmalar, dışarıda tutma ve çapraz doğrulama tekniklerini kullanarak kapsamlı bir karşılaştırma ve yorumlama içermektedir.

Bu çalışmanın geri kalanı aşağıdaki şekilde düzenlenmiştir. Bölüm 2, bu çalışmada kullanılan verisetini ve uygulanan metodolojiyi açıklar. Deneysel çalışmalar ve elde edilen sonuçlar ayrıntılı olarak Bölüm 3'te sunulur ve tartışılır. Son olarak, Bölüm 4'te elde edilen sonuçlar özet olarak sunulmuş ve ayrıca ileride yapılacak çalışmalara yer verilmiştir.

II. METOT

A. VERİSETİ

Çalışmada kullanılan veriseti Islam ve ark. [24] tarafından oluşturulmuştur ve UCI makine öğrenimi deposundan indirilebilir. Bu veriseti, Bangladeş, Sylhet'teki Sylhet Diyabet Hastanesinde 520 hastadan anket yoluyla elde edilen verileri içermektedir. Bu hastalardan 320'si pozitif, 200'ü ise negatiftir. Veriseti diyabetle ilgili semptomları içeren 16 öznitelikten oluşmaktadır. Yaş haricindeki tüm

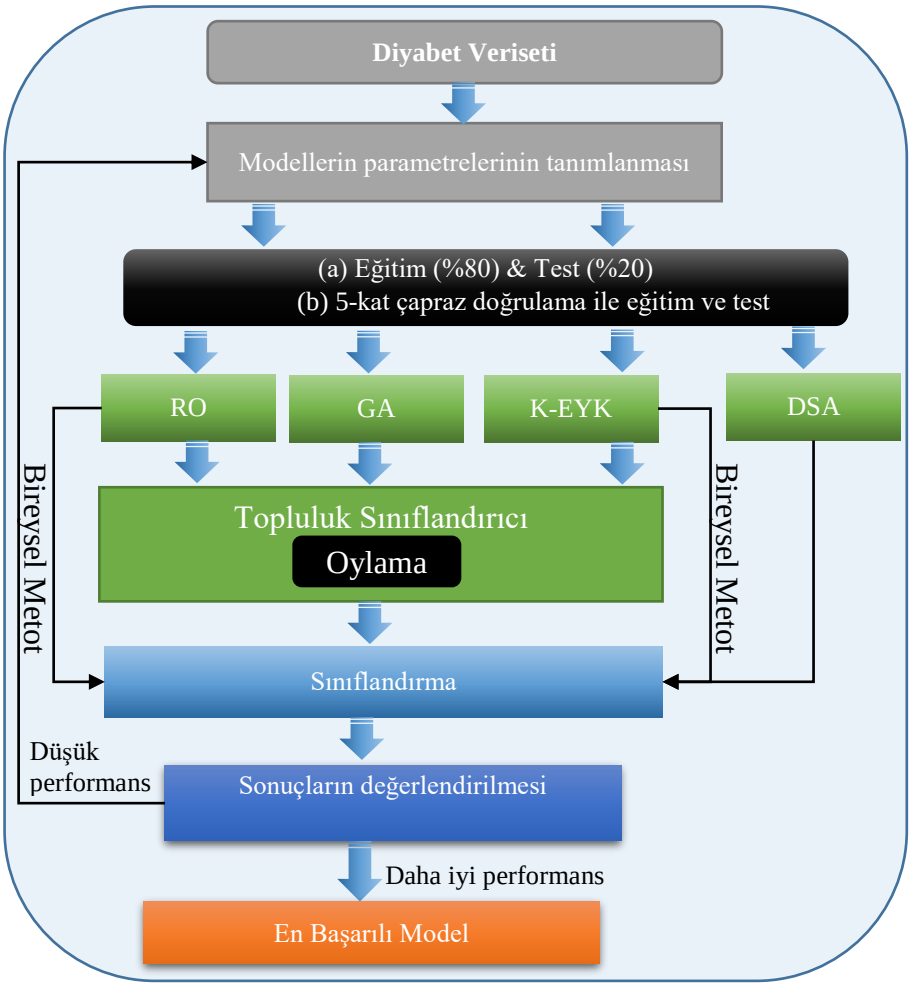
özniteliklerde Evet/Hayır gibi kategorik değerler bulunmaktadır. Bu öznitelikler Tablo 1'de listelenmiştir.

Tablo 1. *Verisetindeki öznitelikler ve değer aralıkları*

No	Öznitelik	Değer Aralığı
1	Yaş	[20-65]
2	Cinsiyet	Erkek, Kadın
3	Poliüri	Evet, Hayır
4	Polidipsi	Evet, Hayır
5	Ani kilo kaybı	Evet, Hayır
6	Zayıflık	Evet, Hayır
7	Polifaji	Evet, Hayır
8	Genital pamukçuk	Evet, Hayır
9	Görsel bulanıklık	Evet, Hayır.
10	Kaşıntı	Evet, Hayır
11	Sinirlilik	Evet, Hayır
12	Gecikmeli iyileşme	Evet, Hayır
13	Kısmi parezi	Evet, Hayır
14	Kas sertliği	Evet, Hayır
15	Alopesi	Evet, Hayır
16	Obezite	Evet, Hayır

B. MODELLERİN İNŞASI

Diyabet hastalığını erken aşamada tahmin etmek için Rastgele Orman (RO), Gradyan Arttırma (GA), K-En Yakın Komşu (K-EYK) makine öğrenmesi algoritmaları ve ayrıca bunların oylama yaklaşımı ile birleştirildiği oylama topluluk sınıflandırıcısı kullanılmıştır. Bunun yanı sıra, Derin Sinir Ağlarının performansı da çalışma kapsamında değerlendirilmiştir. Şekil 1, bu çalışmadaki modellerin inşası ve değerlendirilmesine ait akış şemasını göstermektedir. Öncelikle verisetindeki Evet/Hayır kategorik değerleri 0 ve 1 sayısal değerlerine dönüştürülmüştür. Ardından, her bir model için parametreler belirlenmiş ve bu parametrelere uygun olarak modeller oluşturularak eğitim ve test işlemleri gerçekleştirilmiştir. Eğitim ve test işlemi için veriseti dışarıda tutma ve 5-kat çapraz doğrulama yaklaşımları ile gerçekleştirilmiştir. Dışarıda tutma yaklaşımında veriseti %80 eğitim, %20 test verisi olarak rasgele bölünmüştür. 5-kat çapraz doğrulama yaklaşımında ise veriseti 5 parçaya bölünmüş ve her katta 4 parça eğitim bir parça ise test için kullanılmıştır. Modellere eğitim için verisetindeki 16 öznitelik giriş olarak verilmiş Pozitif (1) ve Negatif (0) şeklinde iki çıkış elde edilmiştir. Eğitim işleminden sonra test işlemi gerçekleştirilmiştir. Model performanslarını arttırmak için parametreler üzerinde ince ayar yapılması bilinen bir olgudur. Bu bakış açısıyla, bu çalışmada her test işleminden sonra mevcut modelin sınıflandırma performansı bir önceki modelin performansı ile karşılaştırılmıştır. Tekraren yeni parametre değerleriyle inşa edilen modeller ile yapılan deneme yanılma çalışmalarında en iyi performansa sahip model bilgisi saklanmıştır. Tablo 2, yapılan denemeler sonucunda en iyi performansın elde edildiği parametre değerlerini göstermektedir.



Şekil 1. *Makine öğrenmesi modellerinin oluşturulması ve değerlendirilmesi için akış şeması*

Tablo 2. *Her bir model için en iyi performansı sunan parametrelerin bilgisi*

Model	**Parametreler**
GA	n_estimators=100, learning_rate=0.02,max_depth=5
RO	n_estimators=4, random_state=55
K-EYK	n_neighbors=4
DSA	Optimizer Algorithm:Adamax, Learning Rate=0.002, beta_1=0.9, beta_2=0.999 Aktivasyon fonksiyonu: Relu Giriş katmanındaki nöron sayısı:16 Çıkış katmanındaki nöron sayısı:2 Gizli katman sayısı:5 Gizli katmanlardaki nöron sayıları:32,64,128,256,512 Epok:300 Kayıp Fonksiyonu: binary_crossentropy

Bu çalışmada kullanılan makine öğrenmesi algoritmaları kısaca aşağıda tanıtılmaktadır.

B. 1. Rastgele Orman (RO)

Doğruluğu artırmak için kullanılan bir topluluk sınıflandırıcısıdır. Tek bir ağaç kullanıldığında ortaya çıkan aşırı öğrenme diğer bir deyişle ezberleme probleminin üstesinden gelmek için geliştirilmiştir. Rastgele orman birçok karar ağacından oluşur ve çeşitli sınıflandırma ve regresyon ağaçlarını bir araya

getirir. Diğer geleneksel sınıflandırma algoritmalarına kıyasla düşük sınıflandırma hatasına sahiptir [25,26].

B. 2. Gradyan Artırma (GA)

GA, regresyon ve sınıflandırma problemleri için kullanılan topluluk makine öğrenme yöntemlerinden biridir. GA'da verilere ağırlık atanır ve bir dizi zayıf öğrenicinin birleşiminden elde edilen verileri eğitmek için bir model oluşturulur. GA, yeni temel öğrenenler olarak ortaya çıkan zayıf öğrenenlerin bir kombinasyonundan eğitim verileri için yeni bir model oluşturmak için yineleme sayıları nedeniyle hala aşırı öğrenmeye sebep olabilir. Bu nedenle GA, ağaç sayısını ve öğrenme hızını tanımlayarak aşırı uyumun üstesinden gelebilir [27].

B. 3. K-En Yakın Komşu (K-EYK)

Örnek tabanlı öğrenmede temel olan bu yöntemde test örneğine en yakın k adet örneğin sınıf etiketlerinin ortalaması ile sınıflandırma işlemi yapılır [28]. K-EYK sınıflandırıcısı, bir girdinin etiketini öklid veya kosinüs mesafesi gibi bir mesafe metriğinde en yakın komşularını bularak ve komşuların etiketlerinden çoğunluk oya göre tahmin eden popüler parametrik olmayan bir sınıflandırıcıdır [29].

B. 4. Derin Sinir Ağı (DSA)

DSA, katman başına çok sayıda doğrusal olmayan nöron içeren çok sayıda sinir ağı katmanından oluşan yapay sinir ağlarıdır. DSA, yapı olarak sığ sinir ağına benzer ancak daha fazla gizli katmana ve daha belirgin hiyerarşi yapısına sahiptir. DSA, birçok önemli uygulamada alternatif makine öğrenimi yöntemlerinden daha iyi performans göstererek önemli ölçüde ilgi görmüştür [30,31].

B. 5. Topluluk Sınıflandırıcı

Topluluk öğrenme algoritmaları, tahmine dayalı analitik çalışmalarda en başarılı makine öğrenme algoritmalarından biridir [32]. Topluluk sınıflandırıcılar temel öğrenici denen birden fazla makine öğrenmesi algoritmasını bir araya getiren sınıflandırıcılardır. Bu yaklaşım, birden çok bireysel sınıflandırıcıyı birleştirerek yüksek doğrulukta topluluk sınıflandırıcıları oluşturmayı amaçlar [33]. Sınıflandırıcı çıktılarının birleştirilmesi, gruptaki en iyi sınıflandırıcıdan daha iyi bir sınıflandırma performansı elde edileceğini garanti etmez. Fakat topluluk sınıflandırıcının doğruluğu tüm bireysel sınıflandırıcıların ortalama doğruluğundan daha az değildir. Topluluk sınıflandırıcıların oluşturulmasında birçok birleştirme yaklaşımı vardır. Bu yaklaşımlardan çoğunluk oylama (Hard Voting) ve ağırlıklı çoğunluk oylama (Soft Voting) basit, yaygın ve etkilidir [34]. Çoğunluk oylama yaklaşımında en çok oyu alan sınıf seçilir. Bu yaklaşım Eşitlik 1'deki gibi tanımlanmaktadır [35].

$$\gamma = \max\big(\varphi(x_i)\big) \text{ Where } \varphi(x_i) = \sum_{k=1}^{K} f_k(x_i) \quad (1)$$

Ağırlıklı çoğunluk oylama yaklaşımında sınıflandırıcının ağırlığı Eşitlik 2'deki gibi hesaplanır. Bu eşitlikte, α_k her bir sınıflandırıcının bireysel doğruluğudur. Bu yöntemde, her sınıflandırıcı kararının ağırlıkları hesaplandıktan sonra, oylamada en yüksek puanı alan sınıf nihai sınıf tahminidir [36]. Bu çalışmada, ağırlıklı çoğunluk oylama yaklaşımı kullanılmıştır.

$$w_k = log\,\frac{\alpha_k}{1-\alpha_k} \quad (2)$$

III. DENEYSEL SONUÇLAR VE TARTIŞMA

Bu bölümde tüm modellerin hem dışarıda tutma hem de 5-kat çapraz doğrulama yaklaşımı ile elde edilen sınıflandırma performansları ortaya konulmuş ve literatürdeki çalışmalarla karşılaştırılmıştır. Modellerin sınıflandırma performanslarının değerlendirilmesinde duyarlılık, özgüllük ve doğruluk metrikleri kullanılmıştır. Bu metrikler [37] sırasıyla Eşitlik 3 ve 5 arasında verilmiştir. Duyarlılık modellerin gerçek pozitifleri ne oranda doğru sınıflandırdığını, özgüllük ise gerçek negatifleri ne oranda doğru sınıflandırdığını ortaya koyar. Yani, duyarlılık gerçek pozitiflerin, özgüllük ise gerçek negatiflerin doğru sınıflandırma oranını vermektedir.

$$Duyarlılık = \frac{DP}{DP+YN} \tag{3}$$

$$\text{Özgüllük} = \frac{DN}{DN+YP} \tag{4}$$

$$\text{Doğruluk} = \frac{DP+DN}{DP+DN+YP+YN} \tag{5}$$

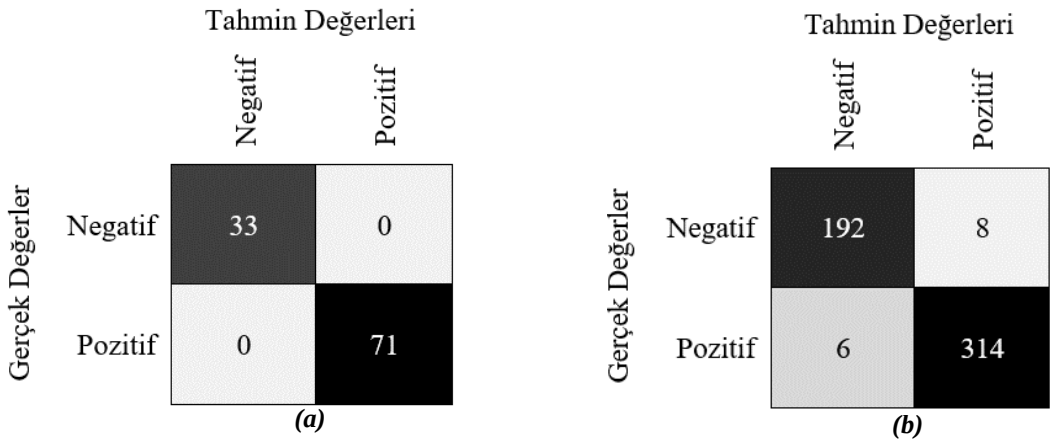
Formüllerdeki parametreler şu şekilde tanımlanmaktadır:
DP: Doğru Pozitif, DN: Doğru Negatif, YP: Yanlış Pozitif, YN: Yanlış Negatif

Oylama topluluk, GA, RO, K-EYK ve DSA modellerinin dışarıda tutma ve 5-kat çapraz doğrulama ile elde edilen sınıflandırma performansları Tablo 3'te gösterilmektedir. Bu değerlere göre en yüksek sınıflandırma performansını hem dışarıda tutma hem de 5-kat çapraz doğrulamada Oylama topluluk sınıflandırıcı vermiştir. Dışarıda tutma tekniğinde hem pozitif hem de negatifleri %100 doğru sınıflandırırken, 5-kat çapraz doğrulamada pozitifleri %98.13 negatifleri ise %96 oranında doğru sınıflandırmıştır. İkinci en yüksek sınıflandırma performansı ise 5-kat çapraz doğrulamada RO, dışarıda tutma tekniğinde ise GA modeline aittir. Üçüncü en yüksek sınıflandırma performansı 5-kat çapraz doğrulamada GA modeline aitken, dışarıda tutma tekniğinde DSA modeline aittir. Burada en düşük performansı ise K-EYK modeli göstermiştir. Eğitim ve test setlerinde sınıflandırıcı algoritmaların performanslarının dengesiz olması makine öğrenmesinde aşırı öğrenme sorunu olarak bilinir. Bu çalışmadaki deneylerde test setlerinde K-En Yakın Komşu sınıflandırıcısı dışındaki tüm sınıflandırıcılardan elde edilen başarıların yüksek olması bu sınıflandırıcıların aşırı öğrenme göstermediğini ifade eder.

Tablo 3. *Modellerin duyarlılık, özgüllük ve doğruluk değerleri*

	Performans Sonuçları (%)					
	5-kat			**Dışarıda Tutma**		
Modeller	**Duyarlılık *(%)***	**Özgüllük *(%)***	**Doğruluk *(%)***	**Duyarlılık *(%)***	**Özgüllük *(%)***	**Doğruluk *(%)***
Oylama Topluluk	98.13	96	97.31	100	100	100
Gradyan Artırma	95.94	94.50	95.38	100	96.97	99.04
Rastgele Orman	96.88	96	96.54	94.37	100	96.15
K-En Yakın Komşu	78.75	96.50	85.58	73.24	81.82	75.96
DSA	95.57	94.36	95.19	98.53	97.22	98.07

Şekil 2, en yüksek sınıflandırma performansı sağlayan Oylama topluluk modeli ile elde edilen karışıklık matrislerini göstermektedir. Bu matrislerden birincisi dışarıda tutma tekniği ile elde edilen sonucu içermekte iken ikincisi 5-kat çapraz doğrulama tekniği ile her bir katta elde edilen karışıklık matrislerinin üst üste bindirilmiş olan sonuç matrisine ait bilgileri içermektedir. Dışarıda tutma tekniğinde bu model test verisetindeki tüm pozitif ve negatif örnekleri doğru sınıflandırırken, 5-kat çapraz doğrulamada 200 negatif örneğin 192'sini, 320 pozitif örneğin ise 314'ünü doğru sınıflandırmıştır.



Şekil 2. *Oylama Topluluk modelinin karmaşıklık matrisi: (a) Dışarıda tutma yöntemi, (b) 5-kat çapraz doğrulama*

Literatürde bu çalışmada kullanılan veriseti üzerinde yapılmış çalışmalar bulunmaktadır. Elde edilen sonuçların önceki çalışmalarla karşılaştırılması önemlidir ve Tablo 4 bu karşılaştırmayı özetlemektedir. Bu çalışmalardan dışarıda tutma tekniğini uygulayan Kumar ve ark. [38] tarafından gerçekleştirilen çalışmada elde edilen sonuç ile bu çalışmada sunulan sonucun aynı olduğu görülmektedir. Le ve ark.'nın [39] çalışmasına kıyasla bizim çalışmamızdaki Oylama topluluk sınıflandırıcısı tabanlı modelin daha başarılı olduğu görülmektedir. Bunun yanı sıra, 10-fold çapraz doğrulamanın uygulandığı çalışmalardaki başarılar ile karşılaştırıldığında önerdiğimiz model %97.31 ile kabul edilebilir seviyede bir sınıflandırma doğruluğu sunmaktadır.

Tablo 4. *Sonuçların önceki çalışmalarla karşılaştırılması*

Çalışma	**Yöntem**	**Veriseti Bölme Tekniği**	**Doğruluk *(%)***
Kumar ve ark. [38]	Gradyan Artırma Karar Ağacı (CatBoost)	Dışarıda tutma (-%80 eğitim, %20 test)	100
Le ve ark. [39]	Adaptif Parçacık Gri Kurt Optimizasyonu ve Çok Katmanlı Algılayıcı	Dışarıda tutma (-%80 eğitim, %20 test)	97
Al-Behadili ve Ku-Mahamud [40]	*Açgözlü Tepe Tırmanması algoritması ile özellik seçimi ve Bulanık mantıkla sınıflandırma*	10-kat çapraz doğruluma	97.692
Özer [41]	Uzun Kısa Dönem Bellek Ağların	10-kat çapraz doğruluma	98.65
Chaves [42]	Yapay Sinir Ağları	10-kat çapraz doğrulama	98.1
Bu çalışma	**Oylama Topluluk Sınıflandırıcısı**	**Dışarıda tutma (-%80 eğitim, %20 test)** **5-kat çapraz doğrulama**	**100.0** **97.31**

IV. SONUÇ

Kandaki anormal derecede yüksek glikoz seviyeleri ile ilişkili kronik bir hastalık olan diyabetin erken evrede tanısı sağlıklı bir yaşam sağlamak için hayati önem taşımaktadır. Bu bağlamda, bu çalışmada bu hastalığın erken tanısı için Oylama topluluk sınıflandırıcısı, K-En Yakın Komşu, Gradyan Arttırma, Rastgele Orman ve Derin Sinir Ağlarına dayalı sınıflandırıcıların performansları ele alınmıştır. Buna göre, Oylama topluluk yaklaşımı %97.31 sınıflandırma doğruluğu ile diğer sınıflandırıcılara kıyasla daha iyi performans göstermiştir. Modellerin performansını doğrulayan dışarıda tutma ve 5-kat çapraz

doğrulama tekniklerinin her ikisinde de Oylama topluluk sınıflandırıcısı, bireysel sınıflandırıcılardan daha yüksek sınıflandırma doğruluğuna sahiptir ve bu verisetindeki performansı oldukça yüksektir. Dışarıda tutma yöntemi ile yapılan deneylerde bu sınıflandırıcı test verisetindeki tüm pozitif ve negatif örnekleri doğru bir şekilde sınıflandırmıştır. 5-kat çapraz doğrulamada ise 200 negatif örnek içerisinden 8 tanesini, 320 pozitif örnek içerisinden sadece 6 tanesini yanlış sınıflandıran bu sınıflandırıcı yüksek doğru pozitif oranı ve düşük yanlış pozitif oranına sahiptir. En iyi performans sunan Oylama Topluluk sınıflandırıcısı tabanlı gerçek zamanlı çalışan bir uzman sistemin şeker hastalığının erken aşamada tespiti çalışmalarına destek vereceği düşünülmektedir. Bu hastalık için önemli özniteliklerin tespiti ve optimizasyon algoritmaları tabanlı deneysel çalışmaların gelecek çalışma olarak yapılması hedeflenmektedir.

TEŞEKKÜR: Yazarlar, kamuya açık olan şeker hastalığı veriseti için Islam ve ark. [24]'e teşekkür eder.

V. KAYNAKLAR

[1] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, 2021.

[2] M. Alehegn and R. Joshi, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach," *International Research Journal of Engineering and Technology*, vol. 4, no.10, pp. 426-436, 2017.

[3] A. Adler *et al.*, "Reprint of: Classification of Diabetes Mellitus," *Diabetes Research and Clinical Practice*, vol. 0, no. 0, p. 108972, In Press, 2021.

[4] R. D. Howsalya Devi, A. Bai, and N. Nagarajan, "A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms," *Obesity Medicine*, vol. 17, p. 100152, 2020.

[5] P. Zimmet, K. G. M. M. Alberti, and J. Shaw, "Global and societal implications of the diabetes epidemic," *Nature*, vol. 414, no. 6865, pp. 782–787, 2001.

[6] J.M. Ekoe, "Diagnosis and Classification of Diabetes Mellitus," *Encyclopedia of Endocrine Diseases (Second Edition)*, vol. 1, pp. 105–109, 2019.

[7] M. Maniruzzaman *et al.*, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Computer Methods and Programs in Biomedicine*, vol. 152, pp. 23–34, 2017.

[8] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques," *Procedia Computer Science*, vol. 112, pp. 2519–2528, 2017.

[9] W. H. O. E. C. on Diabetes Mellitus and W. H. Organization, "Diabetes mellitus : report of a WHO Expert Committee [meeting held in Geneva from 24 to 30 November 1964]." World Health Organization, p. ger published by: Munich : Medizinische Poliklinik, 1965.

[10] P. Bala Manoj Kumar, R. Srinivasa Perumal, R. K. Nadesh, K. Arivuselvan, "Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier," *International Journal of Cognitive Computing in Engineering*, vol. 1, pp. 55–61, 2020.

[11] D. Jashwanth Reddy, B. Mounika, S. Sindhu, T. Pranayteja Reddy, N. Sagar Reddy, G. Jyothsna Sri, et al., "Predictive machine learning model for early detection and analysis of diabetes," *Materials Today: Proceedings,* 2020.

[12] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express,* In Press, 2021.

[13] A. Viloria, Y. Herazo-Beltran, D. Cabrera, and O. B. Pineda, "Diabetes Diagnostic Prediction Using Vector Support Machines," *Procedia Computer Science,* vol. 170, pp. 376–381, Jan. 2020.

[14] J. Chaki, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan, "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review," *Journal of King Saud University - Computer and Information Sciences*, In Press, 2020.

[15] N. Sharma and A. Singh, "Diabetes Detection and Prediction Using Machine Learning/IoT: A Survey,". In: Luhach A., Singh D., Hsiung PA., Hawari K., Lingras P., Singh P. (eds) *Advanced Informatics for Computing Research. Communications in Computer and Information Science*, vol 955. Springer, Singapore, pp 471-479, 2019.

[16] S. Afzali and O. Yildiz, "An Effective Sample Preparation Method for Diabetes Prediction," *International Arab Journal of Information Technology,* vol. 15, no. 6, 2018.

[17] N. Theera-Umpon, I. Poonkasem, S. Auephanwiriyakul, and D. Patikulsila, "Hard exudate detection in retinal fundus images using supervised learning", *Neural Computing and Applications,* vol. 32, pp. 13079–13096, 2020.

[18] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," *Materials Today: Proceedings,* In Press, 2021.

[19] A. Prabha, J. Yadav, A. Rani, and V. Singh, "Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier," *Computers in Biology and Medicine*, vol. 136, pp. 104664, 2021.

[20] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Frontiers in Genetics,* vol. 9, Article 515, 2018.

[21] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocrine Disorders,* vol. 19, Article number: 101, pp. 1–9, Oct. 2019.

[22] S. NAHZAT and M. YAĞANOĞLU, "Diabetes Prediction Using Machine Learning Classification Algorithms," *Avrupa Bilim ve Teknoloji Dergisi,* vol. 24, no. 24, pp. 53–59, 2021.

[23] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big Data,* vol. 6, Article number: 13, pp. 1–19, 2019.

[24] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," In: Gupta M., Konar D., Bhattacharyya S., Biswas S. (eds) Computer Vision and Machine Intelligence in Medical Image Analysis. *Advances in Intelligent Systems and Computing,* vol. 992. pp. 113–125, Springer, Singapore, 2020.

[25] S. Georganos *et al.*, "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling," *Geocarto*

International, vol. 36, no. 2, pp. 121–136, 2019.

[26] N. Farnaaz and M. A. Jabbar, "Random Forest Modeling for Network Intrusion Detection System," *Procedia Computer Science*, vol. 89, pp. 213–217, 2016.

[27] N. Aziz, E. A. P. Akhir, I. A. Aziz, J. Jaafar, M. H. Hasan, and A. N. C. Abas, "A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems," *2020 International Conference on Computational Intelligence*, pp. 11–16, Oct. 2020.

[28] F. Bulut, "Obezite Riski Altındaki Çocukların Örnek Tabanlı Sınıflandırıcı Topluluklarıyla Tespiti," *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, vol. 32, no. 1, pp. 65–76, 2017.

[29] C. Sitawarin and D. Wagner, "On the Robustness of Deep K-Nearest Neighbors," *arXiv:1903.08333*, 2019.

[30] S. Feng, H. Zhou, and H. Dong, "Using deep neural network with small dataset to predict material defects," *Materials & Design*, vol. 162, pp. 300–310, Jan. 2019.

[31] A. Karaci, "Predicting Breast Cancer with Deep Neural Networks," In: Hemanth D., Kose U. (eds) Artificial Intelligence and Applied Mathematics in Engineering Problems. *Lecture Notes on Data Engineering and Communications Technologies*, vol. 43. pp. 996–1003, Springer, Cham, 2019.

[32] G. Bilgin, "Investigation of The Risk of Diabetes in Early Period using Machine Learning Algorithms," *Journal of Intelligent Systems: Theory and Applications*, vol. 4, no. 1, pp. 55–64, 2021.

[33] A. Karaci, O. Ozkaraca, E. Acar, and A. Demir, "Prediction of traumatic pathology by classifying thorax trauma using a hybrid method for emergency services," *IET Signal Processing*, vol. 14, no. 10, pp. 754–764, 2020.

[34] C. Qi and X. Tang, "A hybrid ensemble method for improved prediction of slope stability," *International Journal for Numerical and Analytical Methods in Geomechanics*, vol. 42, no. 15, pp. 1823–1839, 2018.

[35] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, vol. 8, no. 4, pp. e1249, 2018.

[36] F. Moreno-Seco, J. M. Iñesta, P. J. P. de León, and L. Micó, "Comparison of Classifier Fusion Methods for Classification in Pattern Recognition Tasks," In: Yeung DY., Kwok J.T., Fred A., Roli F., de Ridder D. (eds) Structural, Syntactic, and Statistical Pattern Recognition. SSPR /SPR 2006. *Lecture Notes in Computer Science*, vol. 4109. pp. 705–713, Springer, Berlin, Heidelberg, 2006.

[37] D.G. Altman , J.M. Bland, "Diagnostic tests. 1: Sensitivity and specificity," BMJ 1994;308:1552. https://doi.org/10.1136/BMJ.308.6943.1552.

[38] P. S. Kumar, K. Anisha Kumari, S. Mohapatra, B. Naik, J. Nayak, and M. Mishra, "CatBoost ensemble approach for diabetes risk prediction at early stages," *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology*, pp. 1-6, 2021.

[39] T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, "A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic," *IEEE Access*, vol. 9, pp. 7869–7884, 2021.

[40] H. N. K. Al-Behadili and K. R. Ku-Mahamud, "Fuzzy Unordered Rule Using Greedy Hill

Climbing Feature Selection Method: An Application To Diabetes Classification," *Journal of Information and Communication Technology*, vol. 20, no. 3, pp. 391–422, 2021.

[41] İ. Özer, "Uzun Kısa Dönem Bellek Ağlarını Kullanarak Erken Aşama Diyabet Tahmini," *Mühendislik Bilimleri ve Araştırmaları Dergisi*, vol. 2, no. 2, pp. 50–57, 2020.

[42] L. Chaves and G. Marques, "Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study," *Applied Sciences, vol. 11, no. 5, pp. 1-12*, 2021.