Ömer Halisdemir Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi
Yıl: 2025 Cilt-Sayı: 18(1) ss: 292–304

Academic Review of Economics and Administrative Sciences
Year: 2025 Vol-Issue: 18(1) pp: 292–304

https://dergipark.org.tr/tr/pub/ohuiibf

Araştırma Makalesi
Research Article

ISSN: 2564-6931
DOI: 10.25287/ohuiibf.1539535
Geliş Tarihi / Received: 27.08.2024
Kabul Tarihi / Accepted: 28.11.2024
Yayın Tarihi / Published: 31.01.2025

MAKİNE ÖĞRENMESİNDE YENİDEN ÖRNEKLEME: ALGORİTMALARIN PERFORMANSLARINA YANSIMALARI

Ömer Çağrı YAVUZ [1]

Öz

Farklı alanlarda çeşitli uygulamalarda kullanılan makine öğrenmesi teknikleri karmaşık problemlerin çözümünde katkı sağlayarak gelişim göstermektedir. Bu teknikler verilerin işlenmesi, anlamlandırılması ve tahmini gibi çeşitli amaçlarla kullanılmaktadır. Karmaşık problemlerin çözümünde kullanılan sınıflandırma algoritmalarında giriş değerleri üzerinden etiketlenmiş çıkış değerleri tahmin edilmektedir. Ancak bu tür makine öğrenmesi uygulamalarında küme sayılarının dengesiz dağılımlarına bağlı olarak performans kayıpları yaşanmaktadır. Bu performans kayıplarının önüne geçmek amacıyla çeşitli yeniden örnekleme yöntemleri kullanılmaktadır. Alt örnekleme ve aşırı örnekleme olmak üzere iki farklı grupta ele alınan bu yöntemler veri setlerinde yer alan dengesizlikleri ortadan kaldırmak için sıklıkla kullanılır. Alt örnekleme yöntemleri kayıt sayısını sınıf sayısı düşük olan kayıtların sayısına yaklaştırmak amacıyla kullanılır. Aşırı örnekleme yöntemleri ise sınıf sayısı düşük olan kayıtların sayısını yükseltmek amacıyla kullanılır. Bu çalışma kapsamında çeşitli yeniden örnekleme yöntemlerinin makine öğrenmesi algoritmalarının performansları üzerindeki etkisinin ortaya konması amaçlanarak 569 kayıttan oluşan veri seti kullanılmıştır. İyi huylu ve kötü huylu olmak üzere iki farklı sınıftan oluşan göğüs kanseri kayıtlarına çeşitli yeniden örnekleme filtreleri uygulanmıştır. Sonrasında elde edilen veri setlerine dört farklı algoritma uygulanarak elde edilen performans metrikleri karşılaştırılarak sunulmuştur. Yapılan uygulamalar sonucunda yeniden örnekleme yöntemlerinin kullanımının makine öğrenmesi algoritmalarının performanslarına olumlu katkı sağladığı görülmüştür.

Anahtar Kelimeler ***:*** *Makine Öğrenmesi, Yeniden Örnekleme, Sınıflandırma, Performans Metrikleri.*

Jel Sınıflandırılması ***:*** *C53, C67*

[1] Dr. Öğr. Üyesi, Atatürk Üniversitesi, Yapay Zekâ ve Makine Öğrenmesi Bölümü, omercagriyavuz@gmail.com, ORCID: 0000-0002-6655-3754.

***Atıf/Citation (APA 6)*:**
Yavuz, Ö. Ç. (2025). Makine öğrenmesinde yeniden örnekleme: Algoritmaların performanslarına yansımaları. *Ömer Halisdemir Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 18*(1), 292-304. http://doi.org/10.25287/ohuiibf.1539325.

RESAMPLING IN MACHINE LEARNING: IMPLICATIONS FOR ALGORITHM PERFORMANCES

Abstract

Machine learning techniques, used in various applications across different domains, contribute to the development by addressing complex problems. These techniques are utilized for various purposes such as processing, interpreting, and predicting data. In classification algorithms used to solve complex problems, labeled output values are predicted based on input values. However, in such machine learning applications, performance losses occur due to imbalanced distributions of clusters. To mitigate these performance losses, various resampling methods are used. These methods are categorized into two groups: undersampling and oversampling. Undersampling methods are used to approach the number of records to the number of records with low class counts. Oversampling methods, on the other hand, are used to increase the number of records with low class counts. In this study, a dataset consisting of 569 records was used to demonstrate the effect of various resampling methods on the performance of machine learning algorithms. Resampling filters were applied to breast cancer records belonging to two different classes: benign and malignant. Subsequently, performance metrics obtained by applying four algorithms to the resulting datasets were compared. The applications conducted revealed that the use of resampling methods positively contributes to the performance of machine learning algorithms.

Keywords ***:*** *Machine Learning, Resampling, Classification, Performance Metrics***.**

Jel Classification ***:*** *C53, C67*

GİRİŞ

Teknolojinin gelişmesi ve internet kullanımının yaygınlaşmasıyla birlikte veri miktarı artış göstermektedir. Bu artış, büyük miktardaki verilerin işlenmesi, depolanması ve analiz edilmesi gibi süreçlere olan ihtiyacı beraberinde getirmektedir. Üzerinde işlem yapılmayan verilerin değersiz görüldüğü günümüzde veri analizi organizasyonlar için önem arz etmektedir. Gerek stratejik planların yapılandırılmasında gerekse operasyonel faaliyetlerin yürütülmesinde veri bilimi tekniklerinden faydalanılmaktadır. Veri bilimi temelde verilerden değerler çıkarılan tekniklerin bütünüdür (Kotu & Deshpande, 2018: 4). Karmaşık problemlerin çözümünde katkı sağlayan, her geçen gün gelişim gösteren ve veri bilimi alanının temelinde yer alan makine öğrenmesi teknikleri farklı alanlarda çeşitli problemlerin çözümünde kullanılmaktadır.

Makine Öğrenmesi teknikleri verilerin kümelenmesi, işlenmesi, anlamlandırılması ve tahmini gibi çeşitli amaçlarla kullanılmaktadır. Sınıflandırma problemlerinin çözümünde kullanılan algoritmalarla giriş değerleri üzerinden etiketlenmiş çıkış değerleri tahmin edilebilmektedir. Bu tür sınıflandırma problemlerinin çözümünde etiketlenmiş veri setleri kullanılır. Ancak dengesiz veri setleriyle yapılan uygulamalarda küme sayılarının dengesizliklerine bağlı olarak performans kayıpları yaşanabilmektedir. Veri setlerinde yer alan dengesizliklerin giderilerek performans kayıplarının önüne geçmek amaçlanarak çeşitli yeniden örnekleme yöntemleri kullanılmaktadır. Örnek olarak Ghorbani ve Ghousi (2016) çalışmalarında öğrencilerin performansını tahmin etmek üzere oluşturulan iki farklı veri setine yeniden örnekleme yöntemleri uygulayarak çeşitli makine öğrenmesi algoritmalarının performanslarını karşılaştırmayı amaçlamışlardır. Yapılan uygulamalar sonucunda dengesiz veri setlerinin iyi performans göstermediği ve bu durumun çözülmesi gerektiği vurgulanmıştır. Veri dengesizliğinin zorlu bir sorun olduğu belirtilen başka bir çalışmada otizm spektrum bozukluğu kayıtlarından oluşan veri setine çeşitli yeniden örnekleme teknikleri uygulanarak performanslar karşılaştırılmıştır. Bu bağlamda 1.100'den fazla otizm kaydından oluşan veri setine Rastgele Orman ve Naive Bayes algoritmaları uygulanmıştır. Çalışma sonucunda yeniden örneklenmiş veri setleriyle yapılan uygulamalarda performans metriklerinde artış görüldüğü vurgulanmıştır (Alahmari, 2020). Benzer bir çalışmada üç farklı veri seti üzerinde dört farklı yeniden örnekleme yöntemi karşılaştırılmıştır. Bu kapsamda performansların karşılaştırılması amacıyla sınıflandırma algoritmalarından Destek Vektör Makineleri ve Rastgele Orman algoritması kullanılmıştır. Çalışma

sonucunda K Katlamalı Çapraz Doğrulama'nın daha iyi performans gösterdiği belirtilmiştir (Nakatsu, 2020). Naive Bayes ve Destek Vektör Makineleri algoritmalarıyla diyabet hastalığının teşhis edilmesinin amaçlandığı başka bir çalışmada veri setindeki dengesizliklerin giderilmesi amaçlanarak SMOTE tekniği kullanılmıştır. Çalışma sonucunda Vektör Destek Makineleri algoritmasının daha yüksek performans gösterdiği vurgulanmıştır (Harman, 2021). Bir başka çalışmada farklı veri setleri üzerinde yeniden örnekleme yöntemlerinin sınıflandırma algoritmalarının performanslarına etkisinin incelenmesi amaçlanmıştır. Yeniden örnekleme yöntemleri kullanılarak elde edilen farklı veri setlerine Lojistik Regresyon ve Rastgele Orman algoritmaları uygulanarak performans metrikleri karşılaştırılmıştır. Çalışma sonucunda Random Undersampling (RUS) tekniğinin Rastgele Orman algoritmasıyla birlikte diğer yöntemlere göre yüksek performans gösterdiği vurgulanmıştır (Kubus, 2020). Yine dengesiz veri setlerinde performans kayıplarının vurgulandığı benzer bir çalışmada 21 öznitelik içeren ve 7043 müşteriye ait kayıtlardan oluşan veri setine çeşitli yeniden örnekleme yöntemleri uygulanmıştır. Müşteri kaybı tahmini amaçlanarak çeşitli sınıflandırma algoritmaları uygulanmıştır. Çalışma kapsamında yeniden örnekleme uygulanan veri setleri ile orjinal veri setinden elde edilen performans metrikleri karşılaştırılmıştır. Çalışma sonucunda Rastgele Aşırı Örnekleme uygulanan veri seti için Vektör Destek Makineleri algoritmasının performansında %5.7'lik bir artış elde edildiği vurgulanmıştır (Aydın, 2022). Son olarak kalp yetmezliğinde sağkalım analizi amacıyla 299 kayıttan oluşan veri setine çeşitli yeniden örnekleme yöntemleri uygulanarak makine öğrenmesi algoritmalarının performansları karşılaştırılmıştır. Yapılan sınıflandırmalar için Rastgele Orman, KNN ve Ekstra Ağaçlar algoritmaları kullanılmıştır. Çalışma sonucunda yukarı örnekleme için Rastgele Orman algoritması, aşağı örnekleme için Ekstra Ağaçlar algoritmasının daha yüksek performans gösterdiği belirtilmiştir (Türkmenoğlu ve Yıldız, 2021).

Makine öğrenmesi uygulamalarında kullanılan veri setlerinde yer alan dengesizlikler modellerin performansını olumsuz etkileyebildiği gibi güvenilirliği de azaltabilmektedir. Bu durum özellikle sağlık sektöründe karmaşık ve önem arz eden karar süreçlerini olumsuz etkileyebilmektedir. Dengesiz dağılıma sahip veri setlerine uygulanan yeniden örnekleme yöntemleriyle elde edilen performans metrikleri değişiklik gösterebilmektedir. Bu çalışmada iki sınıflı veri seti üzerinde yeniden örnekleme yöntemlerinin makine öğrenmesi algoritmalarının performanslarına etkisinin incelenmesi amaçlanmıştır. Bu doğrultuda 569 kayıttan oluşan veri setine çeşitli yeniden örnekleme yöntemleri uygulanarak elde edilen veri setlerine KNN, Naive Bayes, Rastgele Orman ve Karar Ağacı algoritmaları uygulanmıştır. Farklı veri setleri ile yapılan uygulamalarda yeniden örnekleme yöntemlerinin algoritma performanslarına etkisinin ortaya konulması amaçlanarak performans metrikleri karşılaştırılarak sunulmuştur.

I. YÖNTEM

Çalışma kapsamında iyi huylu ve kötü huylu olarak sınıflandırılan ve 569 kayıttan oluşan veri setine Veri Madenciliği araçlarından biri olan WEKA'nın yeniden örnekleme filtreleri uygulanmıştır. Uygulamalarda Wisconsin Hastanesi bünyesinde William H. Wolberg tarafından paylaşılan veri seti kullanılmıştır (Dua & Graff, 2019). Veri setinde yer alan dağılımların performansa etkisinin değerlendirilmesi amaçlanarak yeniden örnekleme yöntemleri kullanılarak elde edilen veri setlerine K En Yakın Komşu, Naive Bayes, Rastgele Orman ve Karar Ağacı algoritması uygulanmıştır. Yapılan uygulamalar Python sklearn kütüphanesi kullanılarak geliştirilen modelden faydalanılmıştır. Yeniden örnekleme uygulanan dört farklı veri seti için makine öğrenmesi algoritmalarının performansları karşılaştırılmıştır.

I.I. Makine Öğrenmesi Algoritmaları

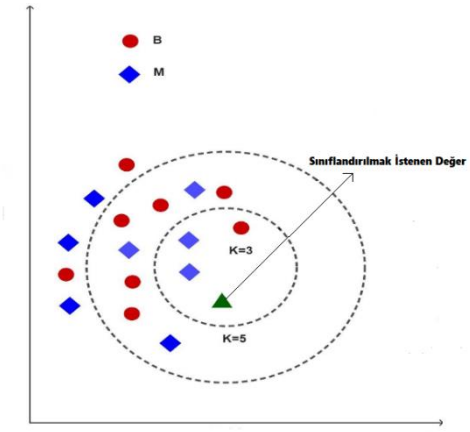
Makine öğrenmesi algoritmaları temelde denetimli öğrenme ve denetimsiz öğrenme olmak üzere iki farklı gruba ayrılmaktadır. Bu çalışma kapsamında ele alınan denetimli öğrenme algoritmalarında mevcut veri setlerinde sınıf değerleri tahmin edilmektedir. Bu tahmin işleminde sınıf değerleri belirli kayıtlardan oluşan veri seti ile öğrenme gerçekleştirilir. Öğrenme işlemi sonrasında giriş değerleri doğrultusunda sınıfı bilinmeyen kayıtlar için tahminler yapılarak atanır (Caruana & Nicelescu-Mizil, 2006: 2).

Bu çalışma kapsamında ele alınan dört farklı denetimli öğrenme algoritması aşağıda özetlenerek anlatılmıştır.

Naïve Bayes algoritması: Bayes teoremine dayanan Naive Bayes algoritmasının temelinde olasılıksal hesaplamalar bulunmaktadır. Olasılıklara dayalı hesaplamalar doğrultusunda elde edilen maksimum değer ilgili sınıfa atanarak sınıflandırma gerçekleştirilir. Bu hesaplamalar için kullanılan formül aşağıda verilmiştir.

$$P(A/B)=(P(B/A)*P(A))/P(B) \quad (1)$$

K En Yakın Komşu algoritması: 1967 yılında Cover ve Hart tarafından geliştirilen K En Yakın Komşu algoritması çeşitli noktalar arasındaki öklid uzaklıklarına dayanır. Şekil 1'de görüldüğü üzere sınıflandırılmak istenen değerin K sayısı kadar komşuya öklid uzaklıkları hesaplanarak ilgili sınıfa atama yapılır (Cover ve Hart, 1967).



Şekil 1: K En Yakın Komşu Algoritması

Karar Ağacı algoritması: Giriş değerlerinin birden fazla homojen kümeye bölünmesine dayalı olarak sınıflandırma problemlerinin çözümünde sıklıkla kullanılır. İstatistiksel bilgi ve analitik altyapı gerektirmediğinden basit kullanım sağlar. Ek olarak veri bilimi uygulamalarının temelini oluşturan ön işleme ve temizleme işlemlerinde kolaylık sağlayarak uç değerlerden etkilenme oranını minimum seviyeye indirir (Sullivian, 2017: 62).

Rastgele Orman Algoritması: Farklı karar ağaçlarının birleşimine dayanan Rastgele Orman algoritması Breiman (2001) tarafından geliştirilmiştir. Oluşturulan her bir ağaç için bağımsız sınıflandırma yapılması amaçlanır. Yapılan sınıflandırılmalar sonrasında oylama yapılarak en yüksek değeri alan sınıf sonuç olarak kabul edilir (Ercire & Ünsal, 2021: 910).

I.II. Yeniden Örnekleme Yöntemleri

Veri setinde yer alan kayıtlarda yer alan sınıf dengesizliğine bağlı durumlarda çeşitli problemlerle karşılaşılabilmektedir. Dengesiz veri setlerinin dengeli hale getirilmesiyle kullanılan algoritmaların başarımları arttırılabilmektedir (Nizam & Akın, 2014: 9). Bu tür dengesizliklerin giderilmesi amacıyla çeşitli yeniden örnekleme yöntemlerinden faydalanılmaktadır. Alt örnekleme ve aşırı örnekleme olmak üzere iki farklı grupta ele alınan yeniden örnekleme yöntemleri dengesizlikleri ortadan kaldırmak için sıklıkla kullanılır (Goy, Gezer, Güngör, 2019: 352). Alt örnekleme yöntemleri farklı sınıfta yer alan ve sayıca değişiklik gösteren kayıtlardan sınıf sayısı yüksek olan kayıt sayısını sınıf sayısı düşük olan kayıtların sayısına yaklaştırmak amacıyla kullanılır. Aşırı örnekleme yöntemleri ise sınıf sayısı düşük olan kayıtların sayısını yükseltmek amacıyla kullanılır (Estabrooks, Jo, Japkowicz, 2004: 24).

Bu çalışma kapsamında yeniden örnekleme yöntemlerinin makine öğrenmesi algoritmaları üzerindeki etkisinin incelenmesi amaçlanarak WEKA'da yer alan filtreler kullanılmıştır. Kullanılan filtreler aşağıda verilmiştir.

Resample: Rastgele alt veri kümesi oluşturmak için kullanılan bu filtre ile veri setinde bulunan kayıt sayısı belirtilebilir (Katore & Umale, 2015: 15).

SMOTE: İnterpolasyona dayalı SMOTE algoritmasında sınıflara ilişkin kayıtlar yapay olarak çoğaltılarak çarpık dağılımlardan kaynaklanan performans değerlerinin giderilmesi amaçlanır. SMOTE algoritmasının çarpık dağılıma sahip veri setleri üzerinde performansa olumlu katkı sağladığı belirtilmektedir (Fernandez, Garcia, Herrera, Chawla, 2018: 866).

SpreadSubSample: Alt örnekleme için kullanılan bu filtre, veri setinde yer alan kayıtların rastgele bir alt kümesini üretir (Gupta, 2017: 4).

Stratified Remove Folds: Algoritmaların başarımını iyileştirme noktasında katkı sağladığı belirtilen bu filtre veri setinde çapraz doğrulama için tanımlanan bir alt küme verir (Mirmozaffari, Golilarz, Band, 2020:16).

II. BULGULAR

Çalışma kapsamında öncelikle elde edilen veri seti üzerinde K En Yakın Komşu, Naive Bayes, Rastgele Orman ve Karar Ağacı algoritmaları için performans metrikleri hesaplanmıştır. Sonrasında çeşitli yeniden örnekleme yöntemleri kullanılarak bu yöntemlerin performans metrikleri üzerindeki etkisi incelenmiştir. K En Yakın Komşu (KNN), Naive Bayes (NB), Rastgele Orman (RF) ve Karar Ağacı (DT) algoritmalarıyla yapılan uygulamalarda elde edilen doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 Skorları Tablo 1'de verilmiştir.

Tablo 1: Performans Metrikleri

	Doğruluk	Çıkış Değeri	Kesinlik	Duyarlılık	F1 Skoru
NB	0.947	0	0.94	0.92	0.93
		1	0.95	0.96	0.96
KNN	0.956	0	0.98	0.94	0.96
		1	0.96	0.99	0.98
RF	0.941	0	0.95	0.92	0.94
		1	0.95	0.97	0.96
DT	0.953	0	0.93	0.89	0.91
		1	0.94	0.96	0.95

Çalışma kapsamında öncelikle göğüs kanseri kayıtlarını içeren orjinal veri setine dört farklı algoritma uygulanmıştır. Orjinal veri seti ile yapılan uygulamalarda en yüksek doğruluk değeri KNN algoritması ile yapılan uygulamada elde edilmiştir. Sonrasında veri setine WEKA aracılığıyla Resample filtresi uygulanarak yeni veri setinden elde edilen performans metrikleri Tablo 2'de verilmiştir.

Tablo 2: Performans Metrikleri (Resample)

	Doğruluk	Çıkış Değeri	Kesinlik	Duyarlılık	F1 Skoru
NB	0.952	0	0.96	0.93	0.95
		1	0.94	0.97	0.96
KNN	0.971	0	1.00	0.96	0.98
		1	0.97	1.00	0.99
RF	0.976	0	0.96	0.98	0.97
		1	0.99	0.97	0.98
DT	0.976	0	0.97	0.98	0.98
		1	0.98	0.97	0.97

Resample filtresi uygulanan veri seti ile yapılan uygulamalarda tüm algoritmalar için doğruluk değerleri artış göstermiştir. Ayrıca çıkış değeri 0 olan kayıtlarının tamamının F1 skorlarında artış görülmüştür. Sonraki aşamada SMOTE filtresi uygulanmıştır. SMOTE uygulanan veri seti ile yapılan uygulamalarda elde edilen performans metrikleri Tablo 3'te verilmiştir.

Tablo 3: Performans Metrikleri (SMOTE)

	Doğruluk	Çıkış Değeri	Kesinlik	Duyarlılık	F1 Skoru
NB	0.942	0	0.93	0.95	0.94
		1	0.95	0.94	0.95
KNN	0.966	0	0.98	0.97	0.97
		1	0.97	0.98	0.98
RF	0.952	0	0.93	0.97	0.95
		1	0.97	0.94	0.95
DT	0.937	0	0.91	0.96	0.93
		1	0.96	0.91	0.93

SMOTE uygulanan veri seti ile yapılan uygulamalarda sadece KNN algoritması için doğruluk değerlerinde artış görülmüştür. Ayrıca çıkış değeri 0 olan kayıtların duyarlılık değerleri ve F1 skorlarında artış görülmüştür. Diğer bir yeniden örnekleme yöntemi olan SpreadSubSample filtresi uygulanan veri seti ile yapılan uygulamalarda elde edilen performans metrikleri Tablo 4'te verilmiştir.

Tablo 4: Performans Metrikleri (SpreadSubSample)

	Doğruluk	Çıkış Değeri	Kesinlik	Duyarlılık	F1 Skoru
NB	0.937	0	0.93	0.93	0.93
		1	0.94	0.94	0.94
KNN	0.954	0	0.92	0.96	0.94
		1	0.97	0.93	0.95
RF	0.937	0	0.89	0.98	0.93
		1	0.98	0.90	0.94
DT	0.905	0	0.88	0.97	0.92
		1	0.96	0.85	0.90

Alt örnekleme yöntemlerinden olan SpreadSubSample uygulanan veri setiyle yapılan uygulamalarda performans metriklerinde genel bir düşüş görülmüştür. Orjinal veri setine son olarak Stratified Remove Folds filtresi uygulanmış olup elde edilen yeni veri seti ile yapılan uygulamalarda elde edilen performans metrikleri Tablo 5'te verilmiştir.

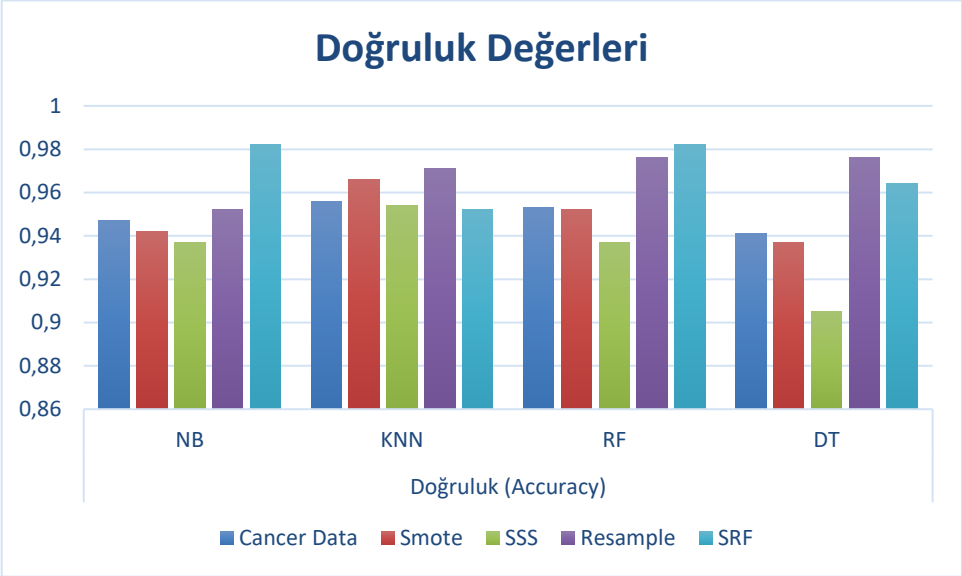
Tablo 5: Performans Metrikleri (StratifiedRemoveFolds)

	Doğruluk	Çıkış Değeri	Kesinlik	Duyarlılık	F1 Skoru
NB	0.982	0	0.96	1.00	0.98
		1	1.00	0.97	0.99
KNN	0.952	0	1.00	0.83	0.90
		1	0.89	1.00	0.94
RF	0.982	0	1.00	0.96	0.98
		1	0.97	1.00	0.99
DT	0.964	0	0.92	0.96	0.94
		1	0.97	0.94	0.96

Stratified Remove Folds uygulanan veri setiyle yapılan uygulamalarda da F1 skorlarında artış görülmüştür. Yapılan uygulamalarda elde edilen doğruluk değerleri değişiminin bir bütün olarak sunulması amaçlanmaktadır. Bu amaç doğrultusunda elde edilen değerler Tablo 6'da, değişimler de Şekil 2'de sunulmuştur.

Tablo 6: Doğruluk (Accuracy) Değerleri

	Doğruluk Değerleri			
	NB	**KNN**	**RF**	**DT**
Cancer Data	0.947	0.956	0.953	0.941
SMOTE	0.942	0.966	0.952	0.937
SpreadSubSample	0.937	0.954	0.937	0.905
Resample	0.952	0.971	0.976	0.976
StrafiedRemoveFolds	0.982	0.952	0.982	0.964



Şekil 2: Doğruluk Değerleri

Diğer performans metriklerinin de bir bütün olarak sunulması amaçlanarak kesinlik değerleri Tablo 7'de, duyarlılık değerleri Tablo 8'de ve F1 skorları da Tablo 9'da sunulmuştur.

Tablo 7: Kesinlik (Precison) Değerleri

	Cancer Data	SMOTE	Resample	SSS	SRF	
0	0.94	0.93	0.96	0.93	0.96	
1	0.95	0.95	0.94	0.94	1	NB
0	0.98	0.98	1	0.92	1	
1	0.96	0.97	0.97	0.97	0.89	KNN
0	0.95	0.93	0.96	0.89	1	
1	0.95	0.97	0.99	0.98	0.97	RF
0	0.93	0.91	0.97	0.88	0.92	
1	0.94	0.96	0.98	0.96	0.97	DT

Tablo 8: Duyarlılık (Recall) Değerleri

	Cancer Data	SMOTE	Resample	SSS	SRF	
0	0.92	0.95	0.93	0.93	1	
1	0.96	0.94	0.97	0.94	0.97	NB
0	0.94	0.97	0.96	0.96	0.83	
1	0.99	0.98	1	0.93	1	KNN
0	0.92	0.97	0.98	0.98	0.96	
1	0.97	0.94	0.97	0.9	1	RF
0	0.89	0.96	0.98	0.97	0.96	
1	0.96	0.91	0.97	0.85	0.94	DT

Tablo 9: F1 Skorları

	Cancer Data	SMOTE	Resample	SSS	SRF	
0	0.93	0.94	0.95	0.93	0.98	NB
1	0.96	0.95	0.96	0.94	0.99	
0	0.96	0.97	0.98	0.94	0.90	KNN
1	0.98	0.98	0.99	0.95	0.94	
0	0.94	0.95	0.97	0.93	0.98	RF
1	0.96	0.95	0.98	0.94	0.99	
0	0.91	0.93	0.98	0.92	0.94	DT
1	0.95	0.93	0.97	0.9	0.96	

SONUÇ VE DEĞERLENDİRME

Yeniden örnekleme yöntemlerinin makine öğrenmesi algoritmalarının performansına etkisinin incelendiği bu çalışmada öncelikle göğüs kanseri kayıtlarından oluşan veri setine Naive Bayes, KNN, Rastgele Orman ve Karar Ağacı algoritması uygulanmıştır. Dört farklı algoritma ile yapılan uygulamalarda 0,94'ün üzerinde doğruluk değeri elde edilmiştir. Bu doğruluk değerleri, hastalık teşhisine yönelik tahminlere ihtiyaç duyulan benzer çalışmalarda makine öğrenmesi algoritmalarının uygulanabilirliğini ortaya koymaktadır.

539 kayıttan oluşan ve dengesiz dağılım gösteren veri setine WEKA'da sunulan dört farklı yeniden örnekleme filtresi uygulanarak dört farklı veri seti elde edilmiştir. Elde edilen veri setleriyle yapılan uygulamalarda SpreadSubSample uygulanan veri setinde bütün algoritmaların doğruluk değerlerinde düşüş görülmüştür. Aksine Resample filtresi uygulanan veri seti ile yapılan uygulamalarda algoritmaların doğruluk değerleri artış göstermiştir. Alt örnekleme grubunda yer alan iki filtrenin sonucu farklı yönde etkilediği görülmüştür. Bu durum göz önünde bulundurularak benzer çalışmalarda birden fazla yeniden örnekleme yönteminin kullanılmasının algoritma performanslarının değerlendirilmesinde katkı sağlayacağı söylenebilir.

Yeniden örnekleme uygulanan veri setleriyle yapılan uygulamalarda algoritmaların performanslarında kayda değer farklılıklar görülmüştür. Örnek olarak SpreadSubSample uygulanan veri setiyle yapılan uygulamalarda Karar Ağacı algoritması için 0,905 doğruluk değeri elde edilirken Resample uygulanan veri setiyle yapılan uygulamada aynı algoritma ile 0,976 doğruluk değeri elde edilmiştir. Benzer şekilde SpreadSubSample uygulanan veri setiyle yapılan uygulamada Naive Bayes algoritması için 0,937 doğruluk değeri elde edilirken Stratified Remove Folds uygulanan veri setiyle yapılan uygulamada 0,982 doğruluk değeri elde edilmiştir. Dahası SMOTE uygulanan veri setiyle yapılan uygulamalarda Naive Bayes, Karar Ağacı ve Rastgele Orman algoritmalarının doğruluk değerlerinde düşüş görülürken KNN algoritmasının doğruluk değerinde artış görülmüştür. Ayrıca ağaç yapısına sahip algoritmalarla yapılan uygulamalarda benzer sonuçlar elde edildiği görülmüştür. Bu doğrultuda doğruluk değerlerinin kullanılan algoritmaların yapısına göre değişiklik gösterdiği

söylenebilir. Ek olarak bu tür uygulamalarda birden fazla algoritmanın kullanılmasının sonuçların etkinliği açısından katkı sağlayacağı söylenebilir.

Toplamda beş farklı setiyle yapılan uygulama sonuçları incelendiğinde performans metriklerinin farklı yönde değişiklik gösterdiği görülmüştür. Örnek olarak SpreadSubSample uygulanan veri setiyle yapılan uygulamada Naive Bayes algoritması için duyarlılık değeri artış gösterirken kesinlik değeri düşüş göstermektedir. Benzer şekilde SMOTE uygulanan veri setiyle yapılan uygulamada Rastgele Orman algoritması için kesinlik değeri artış gösterirken F1 skorunda düşüş görülmüştür. Benzer çalışmalarda birden fazla performans metriğinin hesaplanması ve değerlendirilmesi sonuçların güvenilirliği açısından katkı sağlayacaktır.

KAYNAKÇA

Alahmari, F. (2020). A comparison of resampling techniques for medical data using machine learning. *Journal of Information & Knowledge Management*, *19*(01), 2040016.

Aydın, M. A. (2022). Müşteri Kaybı Tahmininde Sınıf Dengesizliği Problemi. Politeknik Dergisi, 2022, 25 (1), 351-360.

Caruana, R., & Niculescu-Mizil, A. (2006). Denetimli öğrenme *algoritmalarının ampirik bir karşılaştırması. 23. Uluslararası Makine Öğrenimi Konferansı Bildiri Kitabı, s. 161-168.*

Cover, T., & Hart, P. *(1967).* Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

Dua, D., & Graff, C. (2019). "UCI ML Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, *School of Information and Computer Science*, 2019.

Ercire, M., Ünsal, A. (2021). Kısa süreli güç kalitesi bozulmalarının dalgacık analizi ve rastgele orman yöntemi ile sınıflandırılması. *Uludağ Üniversitesi Mühendislik Fakültesi Dergisi*, 26(3), 903-920.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1), 18-36.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61, 863-905.

Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access*, 8, 67899-67911.

Goy, G., Gezer, C., & Gungor, V. C. (2019, September). Credit Card Fraud Detection with Machine Learning Methods. *In 2019 4th International Conference on Computer Science and Engineering (UBMK)*, pp. 350-354. IEEE.

Gupta, V. (2017). Classification of satisfaction level based on survey questions and features selection using decision trees.

Harman G. (2021). Destek vektör makineleri ve naive bayes sınıflandırma algoritmalarını kullanarak diabetes mellitus tahmini. *Avrupa Bilim ve Teknoloji Dergisi*, (32), 7-13.

Katore, L. S., & Umale, J. S. (2015). Comparative study of recommendation algorithms and systems using WEKA. International *Journal of Computer Applications*, 110 (3).

Kotu, V., & Deshpande, B. (2018). Data science: concepts and practice. Morgan Kaufmann.

Kubus, M. (2020). Evaluation of resampling methods in the class unbalance problem. *Econometrics*, 24(1), 39-50.

Mirmozaffari, M., Golilarz, N. A., & Band, S. S. (2020). Machine learning algorithms based on an optimization model.

Nakatsu, R. T. (2020). An evaluation of four resampling methods used in machine learning classification. *IEEE Intelligent Systems*, 36(3), 51-57.

Nizam, H., & Akın, S. S. (2014). Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması. *XIX. Türkiye'de İnternet Konferansı*, 1-6.

Türkmenoğlu, B. K.,& Yıldız, O. (2021). Predicting the survival of heart failure patients in unbalanced data sets. In 2021 29th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

W. Sullivan, "*ML For Beginners Guide Algorithms: Supervised & Unsupervised Learning, Decision Tree & Random Forest Introduction*", CreateSpace Independent Publishing Platform, 2017.

Etik Beyanı **:** *Bu çalışmanın tüm hazırlanma süreçlerinde etik kurallara uyulduğunu yazarlar beyan eder. Aksi bir durumun tespiti halinde ÖHÜİİBF Dergisinin hiçbir sorumluluğu olmayıp, tüm sorumluluk çalışmanın yazar(lar)ına aittir.*

Teşekkür **:** *Yayın sürecinde katkısı olan hakemlere ve editör kuruluna teşekkür ederiz.*

Ethics Statement **:** *The authors declare that ethical rules are followed in all preparation processes of this study. In case of detection of a contrary situation, ÖHÜİİBF Journal does not have any responsibility and all responsibility belongs to the author (s) of the study.*

Acknowledgement **:** *We thank the referees and editorial board who contributed to the publishing process.*