

## Yeniden Örnekleme Metotlarının Kredi Kartı Sahtecilik Tespiti için Topluluk Öğrenmesine Kapsamlı Analizi

Ali Kemal AY<sup>1\*</sup>, Esra Nergis YOLAÇAN<sup>2</sup>

<sup>1</sup>Afyon Kocatepe Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Afyonkarahisar.

<sup>2</sup>Eskişehir Osmangazi Üniversitesi, Mühendislik Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Eskişehir.

Sorumlu yazar e-posta: aykemal.1993@gmail.com ORCID ID: <http://orcid.org/0000-0002-4061-4395>

e-posta: yolacan@ogu.edu.tr ORCID ID: <http://orcid.org/0000-0002-0008-1037>

Geliş Tarihi: 01.02.2022

Kabul Tarihi: 22.09.2022

### Öz

#### Anahtar kelimeler

Dengesiz Veri Kümesi;  
Yeniden Örnekleme  
Metotları; Topluluk  
Öğrenmesi; Kredi Kartı  
Sahtecilik Tespiti;  
Makine Öğrenmesi

Kredi kartı aracılığıyla hızlı ve kolay satın alma işlemleri sahtecilik işlemlerinin artmasına neden olmuştur. Son yıllarda makine öğrenmesi yöntemlerinin kullanımı sahtecilik tespiti işlemlerinde önemli bir pay oluşturmuştur. Sahtecilik tespiti işlemlerinde karşılaşılan yaygın problemlerden birisi veri kümelerinin dengesiz olmasıdır. Dengesizlik problemi için kullanılan yeniden örnekleme metotları kullanıldıkları aşamalar bakımından çalışmadan çalışmaya farklılık gösterebilmektedir. Bu çalışma başlıca topluluk öğrenmesi yöntemleri olmak üzere çeşitli makine öğrenmesi yöntemlerini kullanarak yeniden örnekleme metotlarının kullanıldıkları aşamalara göre yarattığı etkileri karşılaştırmaktadır. Karşılaştırma sonucunda, çapraz doğrulama metodu aracılığıyla yeniden örnekleme metotlarının eğitim ve test veri kümelerine ayrı ayrı yapılmasının en doğru sonucu verdiği gösterilmiştir. Bununla birlikte bu çalışmada kullanılan XGB, LGBM, RF, FNN ve diğer metotların metrik değerlerine dayanan bir başka kıyaslamada ise XGB ve FNN metotları %99 duyarlılık, kesinlik ve doğruluk ile en yüksek değerlere ulaşmışlardır.

## Comprehensive Analysis of Resampling Methods on Ensemble Learning for Credit Card Fraud Detection

### Abstract

#### Keywords

Imbalanced Dataset;  
Resampling Methods;  
Ensemble Learning;  
Credit Card Fraud  
Detection; Machine  
Learning

Rapid and easy purchases via credit cards have led to a rise in fraudulent transactions. In recent years, machine learning methods have been an important part of fraud detection processes. One of the common problems encountered in processes of fraud detection is the imbalance in datasets. Resampling methods used for the problem of imbalance may differ from study to study in terms of the stages these methods are applied. This study compares the effects of resampling methods according to these stages, using various ensemble learning methods, including a few machine learning and deep learning methods. The comparison utilizing cross-validation technique shows that applying the resampling methods separately to the training and test datasets method gives the most accurate result. However, in another comparison by metric scores of XGB, LGBM, RF, FNN, and other methods used in this study, XGB and FNN techniques give the highest values with 99% recall, precision, and accuracy.

© Afyon Kocatepe Üniversitesi

### 1. Giriş

İnternet teknolojilerinin gelişmesiyle son yıllarda kredi kartı kullanımı önemli oranda artmıştır. Kullanım oranlarının artması ayrıca kötü niyetli kişiler tarafından yapılan dolandırıcılık saldırılarını da artırmıştır. Bu tür saldırılar dramatik ölçüde kayıplara neden olmaktadır. The Nilson verilerine

göre, 2019 yılında, dünya çapındaki dolandırıcılık kayıpları 28,65 milyar dolara ulaşmış ve bu rakamın 2027 yılına kadar 38,50 milyar dolara ulaşacağı tahmin edilmektedir (Int. Kyn. 1). Federal Trade Commission'a, 2019 yılında bildirilen başlıca sahtecilik türleri ve bunlara karşılık gelen kayıpları içeren grafikte, yaklaşık 1,7 milyon sahtecilik ve

bunun yaklaşık 250 bin (%15) kadarının ödeme yöntemleri ile ilgili olduğu belirtilmektedir (Int. Kyn. 2). En çok bildirilen 2. sahtecilik türü olarak kredi kartı sahteciliği, 135 milyon dolar kayba sahip yaklaşık 53 bin bildiri ile bu listede yer almaktadır.

Son yıllarda sahtecilik tespiti için makine öğrenmesi yöntemleri kullanımında artış görülmektedir. Makine öğrenmesi yöntemleri uygulanırken karşılaşılan problemlerden birisi üzerinde çalışılan veri kümesinin, European Cardholders (EC), dengesiz olmasıdır. Dengesiz veri kümesinde, bir sınıfa ait veriler diğer sınıfa ait verilerden önemli derecede, sayıca fazladır. Dengesizlik problemi veri sayısı az olan sınıfın makine öğrenmesi metotları tarafından yeterince öğrenilmemesine neden olmaktadır. Bu problemin üstesinden gelebilmek için yeniden örneklemeye (YÖ) metotları kullanılmaktadır.

EC veri kümesi ile yapılan çalışmalar incelenildiğinde, Sohony *et al.* (2018), sahtecilik tespiti için topluluk öğrenmesi (Ensemble Learning) yöntemini önermektedir. İlgili çalışmada YÖ'nin sağladığı faydalardan bahsedilmesine karşı kullanılmadığı görülmektedir. Bununla birlikte topluluk öğrenmesi için gerekli yapıyı, katman ve nöron sayıları bakımından farklılık gösteren 3 tane ileri beslemeli sinir ağları (FNN) ile ağaç sayılarında değişiklik gösteren 2 tane de rastgele orman (RF) metotlarından oluşturmaktadır. Sohony *et al.* (2018) önerdikleri topluluk öğrenmesi metodu ile %86,73 duyarlılık (recall), %85,85 kesinlik (precision) ve %99,95 doğruluk (accuracy) değerlerine ulaşmaktadır.

Taneja *et al.* (2019), alt örneklemeye (down-sampling), yukarı örneklemeye (up-sampling), sentetik azınlık örneklemeye tekniği (SMOTE) ailesi ve uyarlanabilir-sentetik (ADA-SYN) gibi çeşitli yeniden örneklemeye metotlarını dengeli bir veri kümesine ulaşabilmek için önermektedir. İlgili çalışma, RF metodu kullanarak, SVM-SMOTE (Destek Vektör Makinesi) ile %80 duyarlılık, %91 kesinlik ve %85 F Skor (F-Score) değerlerine ulaşmıştır ancak SVM-SMOTE metodunun işlemsel olarak maliyetli bir metot olduğunu paylaşmaktadır.

Husejinović (2020), yeniden örneklemeye metotları kullanmadan, Naïve Bayes (NB), C4.5 Karar Ağacı (DT) ve bir de temel öğretici metot olarak C4.5 DT metodunu kullanan torbalama (Bagging) metotlarını karşılaştırmaktadır. Husejinović (2020) tüm metotlarda normal veriler için Kesinlik-Duyarlılık (Precision-Recall) eğrisi altında alan değeri oranını %99 ve üzeri, sahtecilik verileri için sırasıyla %8, %74,5 ve %82,5 değerlerini paylaşmaktadır.

Isabella *et al.* (2021), dengeli veri kümesine ulaşabilmek için rastgele aşırı örneklemeye (ROS) metodunu, sahtecilik tespiti için Binary Cross Entropy – Gradient Boost Hybrid Ensemble Classifier (BCE-GBHEC), AdaBoost, k en yakın komşu (KNN), AdaBoost+SVM metotlarını kullanmaktadır. Isabella *et al.* (2021) önerdiği BCE-GBHEC metodu ile aralarındaki en yüksek doğruluk değerine %97,23 ile ulaşmaktadır. Diğer metotlar ise KNN %96,89, AdaBoost %96,87 ve AdaBoost+SVM %96,01 doğruluk değerlerine sahiptir.

Kredi kartı sahtecilik tespiti (KKST) literatürü incelendiğinde, EC veri kümesi ile yapılmış, topluluk öğrenmesinin geçtiği çalışmalar olarak yukarıda incelenen 4 çalışma bulunmuştur. İncelenen çalışmalardan sadece ikisinde, Taneja *et al.* (2019) ve Isabella *et al.* (2021), YÖ metotları kullanılarak başarı seviyesinde artış sağlandığı paylaşılmıştır ancak ilgili çalışmalar dahil olmak üzere, KKST literatüründe EC veri kümesi ile yapılmış YÖ metotları kullanılan, 2020 ve 2021 yıllarında yapılmış çalışmalar Çizelge 1'de incelenmektedir. Toplam 20 çalışmanın 4 tanesinde aşırı örneklemeye metotları, eğitim, test veri kümesi ayırımından önce uygulanmasıyla beraber diğer 6 çalışmada ne zaman, nasıl uygulandığı bilinmemektedir. Kalan 10 tane çalışma da ise aşırı örneklemeye metotları, eğitim, test ayırımından sonra kullanılmaktadır. Bilinmeyen çalışmalar, bilinenlerin yüzdeleri dikkate alınarak dağıtıldığında çalışmaların yaklaşık %30'unda aşırı örneklemeye, eğitim test ayırımından önce yapılmıştır denebilir. Aşırı örneklemeye metotlarının tüm veri kümesi üzerine uygulanması veya eğitim, test veri kümeleri ayırımı yapıldıktan sonra ayrı ayrı uygulanması durumlarında farklı sonuçlar doğmaktadır. Örneğin, aşırı örneklemeye tüm veri kümesine uygulandıktan sonra, eğitim, test

kümesi ayrımı yapıldığı durumda, başarı değerlendirme aşamasında, örneklemenin uygulanmadığı haline kıyasla bir artış görülmektedir ancak bu artış, tıpkı dengesiz veri kümelerinde doğruluk metriği kullanımının meydana getirdiği gibi olumlu fakat yanıltıcı etki yaratmaktadır. Diğer çalışmaların aksine, bu çalışmada, KKST literatüründeki en yaygın XGB (Extreme Gradient Boosting), LGBM (Light Gradient Boosting Machine) gibi topluluk öğrenmesi metotları başlıca olmak üzere diğer öğrenme yöntemleri de kullanılarak YÖ metotlarının uygulandıkları aşamalara göre, başarı seviyesine etkileri detaylıca incelenmektedir. Bu incelemeyle birlikte, YÖ ve özneliliklerle ilgili önerilen çeşitli yöntemlerin kombinasyonu bu çalışmaya özgünlük katmaktadır. Ayrıca bu kombinasyondan yararlanılarak hızlı ve başarılı tespit yapabilen çeşitli metotlar karşılaştırılmaktadır. Çalışmanın ikinci kısmında materyal ve metot, üçüncü kısmında bulgular ve tartışma ve son kısmında ise çalışmanın sonuçlarına değinilmektedir.

## 2. Materyal ve Metot

Bu kısımda çalışmada kullanılacak YÖ metotları, makine öğrenmesi metotları, veri kümeleri ve değerlendirme metrikleri genel olarak açıklanmıştır (2.1-2.4). Son olarak, önerilen metot başlığında kısaca teknikleri açıklanan bu metotların bu çalışmaya özgü nasıl kullanıldığının detayları anlatılmaktadır.

### 2.1 Yeniden Örneklemeye

YÖ, Efron (1982), veri kümesindeki anlamı bozmadan veri miktarını azaltabilmek, anlamı daha net elde edebilmek için veri miktarını çoğaltabilmek gibi çeşitli amaçlarda kullanılır. Bu çalışmada örneklemeye metotlarından SMOTE ve rastgele örnek azaltma (RUS) metotları kullanılmaktadır.

SMOTE, Chawla *et al.* (2002) tarafından geliştirilen, azınlık sınıfı için sentetik veri üretme tekniğidir. Aşırı örneklemeye metodunda, örneklerin birbirinin direkt kopyalanmasından kaynaklanan aşırı öğrenme problemine karşı önerilmektedir.

RUS verileri azaltmaya yönelik bir yöntemdir. Çoğunluk sınıfı verileri rasgele seçilir, istenilen

denge oranı elde edilene kadar seçilen bu örnekler veri kümesinden atılır.

**Çizelge 1.** 2020 ve 2021 yıllarında EC veri kümesi ile yapılmış YÖ metotları kullanan çalışmalar.

Yazarlar ve Yılları	Kullanılan YÖ Metotları	Aşırı Örneklemeye Eğitim-Test Ayrımından Önce mi?
bin Alias <i>et al.</i> (2021)	SMOTE, TOMEK	Evet
Tran and Dang (2021)	SMOTE, ADASYN	Belirtilmemiş
Isabella <i>et al.</i> (2021)	ROS	Evet
Wibowo and Fatchah (2021)	ROS, ADASYN, SMOTE, Borderline-SMOTE	Hayır
Bej <i>et al.</i> (2020)	LoRAS, SMOTE, Borderline-1 SMOTE, ADASYN	Hayır
Shamsudin <i>et al.</i> (2020)	ROS(ADASYN, Borderline, SVM-SMOTE) + RUS	Belirtilmemiş
Riffi <i>et al.</i> (2020)	SMOTE	Evet
Rtayli and Enneya (2020)	SMOTE	Belirtilmemiş
Alam <i>et al.</i> (2020)	SMOTE, ADASYN, RUS	Hayır
Nguyen <i>et al.</i> (2020)	RUS, SMOTE, NearMiss	Belirtilmemiş
Janbandhu <i>et al.</i> (2020)	SMOTE, ADASYN	Hayır
Wang <i>et al.</i> (2020)	Oversampling, Mini-Batch Under-Sampling	Hayır
Aung <i>et al.</i> (2020)	SMOTE	Hayır
Mrozek <i>et al.</i> (2020)	RUS, SMOTE	Belirtilmemiş
Shivanna <i>et al.</i> (2020)	SMOTE	Evet
Shah <i>et al.</i> (2020)	ROS, RUS, SMOTE	Belirtilmemiş
Tingfei <i>et al.</i> (2020)	SMOTE, GAN	Hayır
Mînăstireanu and Meşniță (2020)	ROS, RUS	Hayır
Gulati (2020)	SMOTE, Neighborhood Cleaning Rule	Hayır
Zhang <i>et al.</i> (2020)	RUS, NearMiss, KNN	Hayır

### 2.2 Makine Öğrenmesi

Makine öğrenmesi yöntemleri çeşitli alanlarda yaygın şekilde kullanılmaktadır (Alam *et al.*, 2020; Çınarer vd., 2021; Garg *et al.*, 2019; Mochida *et al.*, 2019). KKST literatüründe en yaygın makine öğrenmesi metotları RF, KNN, SVM, Lojistik Regresyon (LR) ve bunlara ek olarak Çok Katmanlı Algılayıcı (MLP), FNN, XGB, LGBM gibi farklı topluluk

metotları çalışmada kullanılarak sonuçları karşılaştırılmaktadır.

RF, Breiman (2001), birden çok DT ile oluşturulan bir orman gibi düşünülebilir. Her bir DT için n boyutlu eğitim veri kümesinden, n tane eleman yinelemeli seçilir dolayısıyla orijinal eğitim veri kümesinden farklı bir küme ortaya çıkarak, her bir DT bir tahminde bulunur. Bu tahminler oylama yöntemiyle çoğunluğa sahip olan sınıf son tahmin olarak karar verilir.

SVM, Vapnik (2000), farklı sınıflar altında bulunan örnekleri ayıran bir ayırma çizgisini (hiperdüzlem) tanımlaya çalışan bir metottur. Dolayısıyla yeni örnekler geldikçe, örnekler çizginin hangi tarafına düştüğüne göre, SVM sınıflandırma yapmaktadır.

LR, Wright (1995), bir ya da daha özneteliği dikkate alarak ikili olasılık değeri hesaplamak için bir fonksiyonel teknik kullanmaktadır. Genel olarak, iki durumun olduğu senaryolarda kullanıldığı görülür.

MLP, McCulloch ve Pitts (1943), ileri beslemeli yapay sinir ağına dayanmaktadır, en az 3 temel katmandan (girdi, gizli ve çıktı) oluşur. Girdi katmanı, işlenecek girdi değerlerini alır. Çıktı katmanı, tahmin yada sınıflandırma için gerekli işlemleri uygulamaktadır. Rastsal sayıdaki gizli katman ise girdi ve ağırlık değerlerine göre bir hesaplama yapmaktadır. Girdi ve çıktı katmanı arasındaki gizli katman sayısı artırılabilir ki bu doğruluğu yükseltebilirken eğitim ve tahmin süresini de uzatabilmektedir.

KNN, Cover ve Hart (1967) tespit edilecek örneğin, k tane en yakın komşusunu bulmaktadır ve bu komşular arasında hangi sınıf ağırlıkta ise tespit sonucunu o sınıfa vermektedir.

LGBM, Ke vd. (2017), diğer gradyan artırma (gradient boosting) yapılarına kıyasla doğruluk değerini düşürmeden daha hızlı tahmin yapabilmeyi sağlayan bir DT metoduna dayanan bir yapıdır.

XGB, Chen ve Guestrin (2016), DT metoduna dayanan, gradyan artırma yapısını kullanan bir topluluk öğrenmesi metodudur.

### 2.3 Veri Kümesi

European Cardholders, (Int Kyn. 3) Eylül 2013 yılında, iki gün içerisinde Avrupalı kredi kartı sahipleri tarafından gerçekleştirilmiş işlemleri içermektedir. Bunlardan 492 tanesi dolandırıcılıkla ilgili olmakla beraber toplamda 284807 veri vardır. Ayrıca veri kümesi, PCA tekniği dönüşümünün sonucuyla sadece sayısal değerlerden oluşmaktadır. Gizlilik durumları nedeniyle (Time and Amount) öznetelikleri hariç, diğer öznetelikler V1,V2, ...,V28 şeklinde kodlanarak paylaşılmıştır.

### 2.4 Değerlendirme Metrikleri

Bu çalışmada kullanılan metriklerin formülleri aşağıda paylaşılmıştır (1-4). Metriklerin hesaplanmasında kullanılan karmaşıklık matrisi Çizelge 2’de paylaşılmaktadır.

Çizelge 2. Karmaşıklık Matrisi

		Gerçek	
		1	0
Tahmin	1	TP	FP
	0	FN	TN

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (1)$$

$$Kesinlik = \frac{TP}{TP+FP} \times 100 \quad (2)$$

$$Duyarlılık = \frac{TP}{TP+FN} \times 100 \quad (3)$$

$$F\_Skor = \frac{2 \times Kesinlik \times Duyarlılık}{Duyarlılık + Kesinlik} \times 100 \quad (4)$$

İşlem karakteristik eğrisi altındaki alan (AUROC) için bir değer döndürmektedir. Bu eğri ise sınıflandırma için kullanılacak farklı eşik değerleri ile bulunan doğru ve yanlış pozitif oranlarla oluşmaktadır.

PR eğrisi altındaki alan (AUPR) için bir değer döndürmektedir. ROC eğrisinden farkı ise dengesiz veri kümeleri için genellikle önemsiz olan doğru tespit edilmiş normal işlemler değerini hesaba katmadan başarı değerlendirmesi yapan kesinlik ve duyarlılık metriklerinden oluşturulmasıdır.

### 2.5 Önerilen Metot

Bu çalışmada ilk olarak EC veri kümesine, veri ön işlemleri uygulanarak daha az veri ile daha hızlı ve başarılı bir tespit için makine öğrenmesi yöntemlerine zemin oluşturulmuştur. Veri

önışlemleri için, Şekil 1'de gösterilen sınıfla özneliklerin korelasyon ısı haritasından yararlanılmıştır. Şekil 1'de görüldüğü üzere en yüksek korelasyona sahip 4 öznelik sırasıyla, V17, V14, V12 ve V10'dur. Başarı seviyesini yükseltebilmek için, bu 4 öznelik için eşik değerleri kullanılmıştır. Burada öznelik sayısının 4 olması pozitif ya da negatif yönde 0.2 korelasyon değerinin üstünde kalan öznelik sayısının 4 olmasındandır. Şekil 1'de en yüksek ilişkili özneliğin korelasyon değeri -0.33 olduğu için 0.2 değeri korelasyon için düşük olsa da tercih edilmiştir. Çizelge 3'te paylaşılan deneme, yanılma ile belirlenen eşik değerleri, sahtecilik sınıf verilerinin mümkün olabildiği kadar çoğunu kapsamaktadır. Bu eşik değerlerinin belirlenmesi ile kalan veri sayıları şöyledir: 41491 normal, 492 sahtecilik ve toplam veri sayısı 41983'tür. Ayrıca diğer öznelik değerleriyle uyumlu olabilmesi adına Amount özneliğine normalleştirme işlemleri uygulanmıştır.

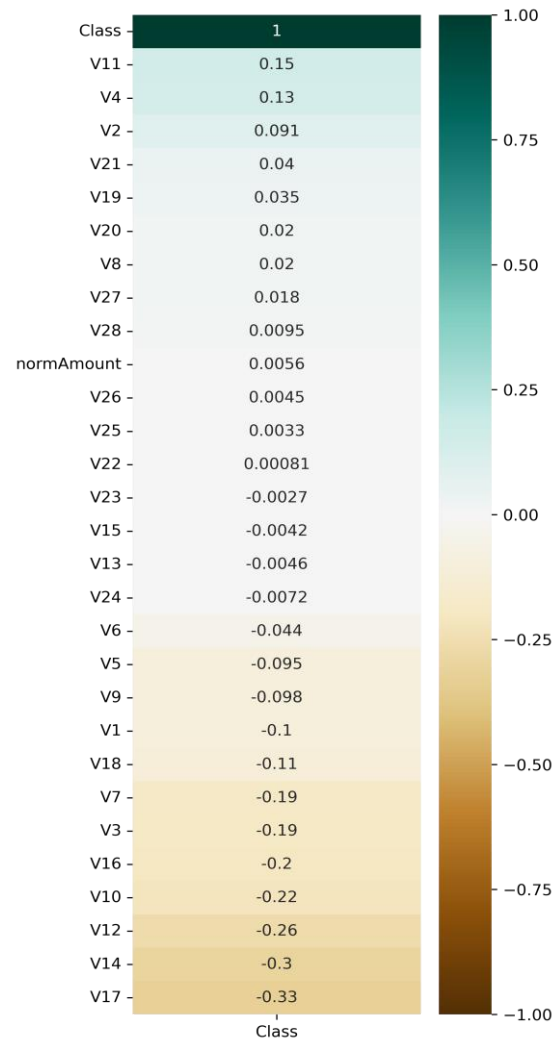
Uygulanan veri önışlemlerinden sonra, KKST için bazı topluluk öğrenmesi yöntemleri de dahil olmakla birlikte çeşitli makine öğrenmesi yöntemleri ve YÖ metotlarının kullanıldığı aşamalar karşılaştırılmaktadır. Bu çalışmada kullanılan her bir makine öğrenmesi yöntemi 4 farklı deney ortamında eğitilip, test edilmiştir: (i) ilk deney ortamında herhangi bir YÖ metodu kullanılmamaktadır, (ii) ikincide sadece test veri kümesine uygulanarak dengeli hale gelmiştir, (iii) üçüncüde eğitim ve test veri kümelerine ayrı ayrı uygulanarak her ikisi de dengeli olmuştur, (iv) dördüncüde ise tüm veri kümesine YÖ metotları uygulanıp veri kümesi dengeli hale geldikten sonra eğitim, test veri kümesi ayrımı yapılmaktadır.

Her bir makine öğrenmesi yöntemi Python yazılımı ile kullanılırken kendine özgü parametrelere sahip olmaktadır. RF metodu için karar ağacı sayısını alan  $n\_estimators$  ya da bir dallanma kalitesini ölçmek için kullanılan yöntemleri girdi olarak alan  $criterion$ , LR metodu için nihai yakınsamayı çözmek için gerekli maksimum iterasyon sayısı bu parametrelere örnek olarak verilebilir. Bu parametrelerin aldıkları değerler problemin türüne göre değişiklik gösterebilmektedir. Bu çalışma kapsamında ise en iyi parametre değerlerine ulaşabilmek için

GridSearchCV ile farklı parametre değerleri test edilmiştir. Örneğin GridSearchCV ile RF metodu  $n\_estimators$  parametresi için 100, 400, 600 ve 800 değerleri denenmiş parametrenin varsayılan değeri (100) ile denenilen değerler arasında önemli farklar görülmemiştir. Bu durum diğer metotların denenilen parametre değerleri için de geçerli olduğu için bu çalışmada her bir metot parametresi varsayılan değerleri ile kullanılmıştır.

**Çizelge 3.** En yüksek korelasyona sahip 4 öznelik için eşik değerleri.

	V17	V14	V12	V10
Eşik Değeri	< 0.5	< 0.5	< 0.2	< 0.2



**Şekil 1.** Sınıfla özneliklerin korelasyon ısı haritası.

### 3. Bulgular ve Tartışma

Veri kümesinin sabit bir kısmını eğitim kümesine, kalanını test kümesine ayırmanın (Hold-out) taraflı bir sonuç doğuracağından, 5 katlamalı çapraz

doğrulama (Cross Validation) yöntemi tercih edilmiştir.

Çizelge 4 ve 5'te her metoda ait birinci ve ikinci satırlarının karşılaştırılması test veri kümesinin dengesiz ve dengeli olduğu durumlardaki farklılıklarını yansıtmaktadır. Tüm metotların birinci satırlarındaki doğruluk ve kesinlik ve bazı diğer değerleri ikinci satırlarda olması gereken asıl değerlerine düşerek, dengesiz veri kümesinde hatalı kullanılan bazı metriklerin yanıltıcı algısını ortaya koymaktadır. İkinci ve üçüncü satırlar arasındaki karşılaştırma ise eğitim kümesine YÖ metotları uygulamanın katkılarını göstermektedir. Eğitim kümesinde YÖ metotları kullanılarak, eğitilen tüm metotlar, Kesinlik metriği hariç diğer tüm metriklerde artış sağlamaktadır. Bununla birlikte YÖ metotlarının en çok katkı sağladığı öğrenme metodu LR ve sonrasında SVM metotlarıdır. Bu karşılaştırmaların en önemlisi olan, metotların üçüncü ve dördüncü satırlarının karşılaştırılması ise YÖ metotlarının eğitim, test ayırımından önce veya sonra uygulanmasının doğuracağı bir başka yanıltıcı artışı ifade etmektedir. Aşırı örnekleme metodu ve türevleri, çapraz doğrulama için gerekli eğitim, test ayırımından önce tüm veri kümesine uygulandığında metot gereği yaratılan birbirine benzer veya kopya verilerin ayırımından sonra eğitim ve test veri kümelerine dağıtılma ihtimali vardır. Benzer verilerin iki tarafa dağıtılması, öğrenme yönteminin eğitim kümesinde gördüğü örneklerin aynısı ya da çok benzeriyle tekrar test edilmesi anlamına gelmesidir. Buda bir bakıma modele kopya vererek başarı değerlendirmesini yükseltmektir. Metotların birinci satırlarında doğruluk vb. diğer metrik değerlerinin olması gerektiğinden fazla olması gibi dördüncü satırlarındaki değerlerde yanlış algı yaratmaktadır. Dolayısıyla dört satırdan en doğru, gerçekçi değerlerin üçüncü satırda olduğu çıkarımı yapılmıştır. Tüm bunlarla birlikte test veri kümesine direkt aşırı örnekleme ve türevleri metotların kullanılmasının da yanıltıcı sonuçlar üreteceğinden kullanılmaması tercih edilmiştir.

Çizelge 4'ün üçüncü satırları dikkate alınarak, öğrenme yöntemleri ve Duyarlılık metriği bakımından karşılaştırma yapıldığında Şekil 2'de görüldüğü üzere en yüksek değere FNN %95 ve en

düşük değere ise SVM ve RF metotları %88 ile sahiptir. Bununla birlikte, Stacking(XGB, FNN) yönteminin diğer topluluk öğrenmesi (RF, LGBM, XGB) yöntemlerine kıyasla başarı seviyelerinde önemli bir fark bulunmamaktadır ancak bu yöntemler eğitim, test amacıyla birden çok metot kullandığı için işlem maliyetleri diğer metotlara kıyasla yüksektir.

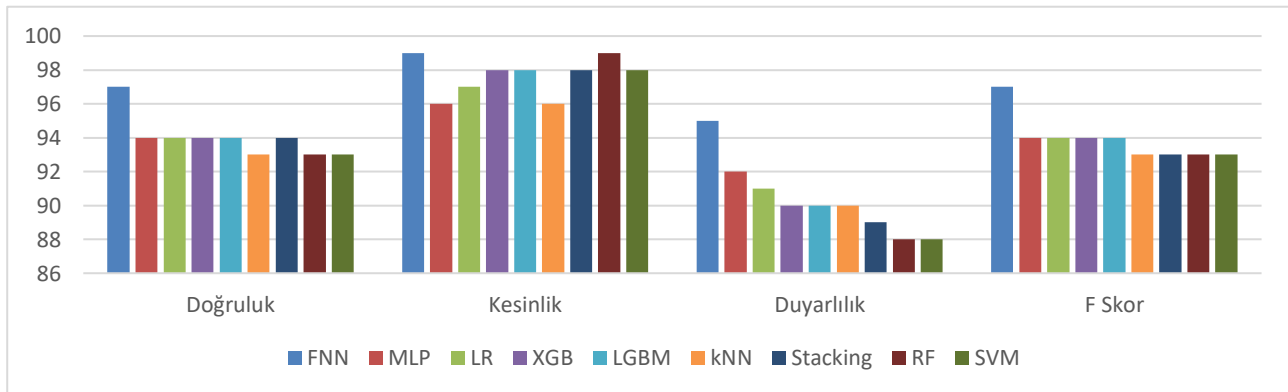
Veri ön işlemlerinin etkisinin görülebilmesi için de uygulanmadığı sonuçlar Çizelge 4'te ve uygulandığı sonuçlar Çizelge 5'te paylaşılmaktadır. Çizelge 5'te paylaşılan her bir metodun üçüncü satırları Şekil 3'te temel metrikler kapsamında karşılaştırılmıştır. Şekil 2 ve 3 incelendiğinde, veri ön işlemleri nedeniyle F-Skor metriğinde en fazla düşüş SVM metodunda %4, en az düşüş ise MLP metodunda %1 ile görülmektedir. Genel olarak %2 değer düşüşüyle birlikte orijinal veri kümesine kıyasla daha az veri ile neredeyse aynı başarılı modeller elde edilmiştir. Bu sonuçların devamında ise metotlar arasında en başarılı sonuçlar veren FNN metodunun veri ön işlemsiz ve ön işlemler durumlarındaki karmaşıklık matrisleri (Çizelge 6 ve 7) ve sonrasında ROC (Şekil 4 ve 6) ve PR (Şekil 5 ve 7) eğrileri paylaşılmıştır.

#### 4. Sonuç

Bu çalışmada, 2020 ve 2021 yıllarında EC veri kümesi ile yapılmış YÖ metotları kullanan 20 farklı çalışma incelendiğinde, YÖ metotları kullanıldıkları aşamalar bakımından farklılık göstermektedir. Bu farklılıklar ortaya koyulduktan sonra en doğru sonuçların eğitim ve test veri kümesine ayrı ayrı YÖ metotlarını uygulayarak bulunacağı gösterilmiştir. Ayrıca test veri kümesine uygulanacak aşırı örnekleme metotlarının da yanıltıcı bir etki yarattığı paylaşılmıştır. Gelecek çalışmada örnekleme metotlarının önemini gösterebilmek için farklı öğrenme yöntemleri kullanılabilir.

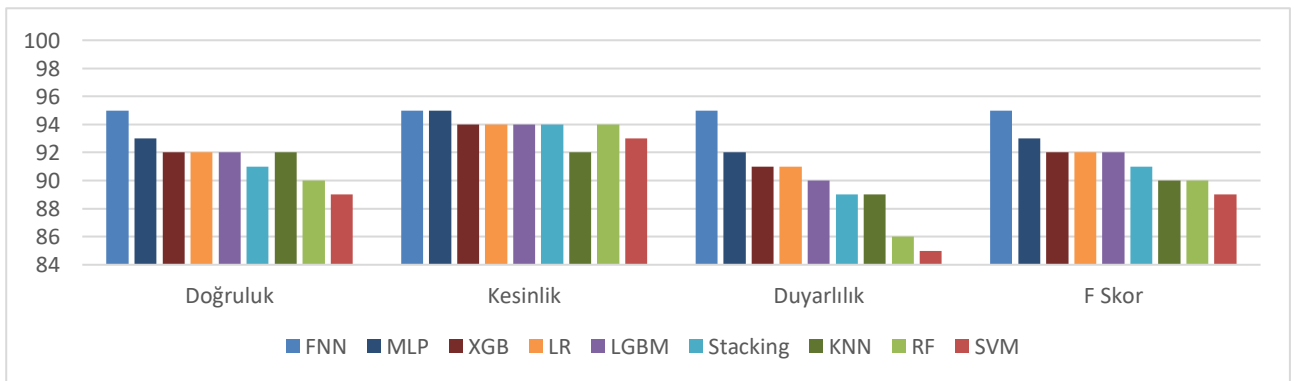
**Çizelge 4.** Veri ön işlemsiz yeniden örneklemeye metotlarının kullanıldığı aşamalara göre öğrenme metotlarına etkisi.

Metot	YÖ Metotları		Değerlendirme Metrikleri (%)					
	Eğitim	Test	Doğruluk	Kesinlik	Duyarlılık	F Skor	AUROC	AUPR
XGB	-	-	100	94	79	86	90	87
	-	RUS (1)	90	100	79	89	90	95
	SMOTE(0.0034), RUS(1)	RUS (1)	94	98	90	94	94	96
	SMOTE(0.0034), RUS(1)		97	99	95	97	97	98
LGBM	-	-	100	47	38	42	69	42
	-	RUS (1)	68	100	38	55	69	84
	SMOTE(0.0034), RUS(1)	RUS (1)	94	98	90	94	94	96
	SMOTE(0.0034), RUS(1)		97	98	96	97	97	98
Stacking (XGB, FNN)	-	-	100	95	77	85	88	86
	-	RUS (1)	86	100	71	83	86	93
	SMOTE(0.0034), RUS(1)	RUS (1)	94	98	89	93	93	96
	SMOTE(0.0034), RUS(1)		97	99	95	97	97	98
RF	-	-	100	95	78	86	89	87
	-	RUS (1)	89	100	79	88	89	95
	SMOTE(0.0034), RUS(1)	RUS (1)	93	99	88	93	93	96
	SMOTE(0.0034), RUS(1)		96	99	94	96	96	98
kNN	-	-	100	94	79	86	90	86
	-	RUS (1)	90	100	79	88	90	95
	SMOTE(0.0034), RUS(1)	RUS (1)	93	96	90	93	93	95
	SMOTE(0.0034), RUS(1)		96	96	96	96	96	97
SVM	-	-	100	95	68	79	84	81
	-	RUS (1)	84	100	68	81	84	92
	SMOTE(0.0034), RUS(1)	RUS (1)	93	98	88	93	93	96
	SMOTE(0.0034), RUS(1)		95	98	91	94	95	97
LR	-	-	100	87	62	73	81	75
	-	RUS (1)	81	100	62	77	81	91
	SMOTE(0.0034), RUS(1)	RUS (1)	94	97	91	94	94	96
	SMOTE(0.0034), RUS(1)		95	97	92	95	95	97
FNN	-	-	100	88	81	84	90	84
	-	RUS (1)	89	100	79	88	89	95
	SMOTE(0.0034), RUS(1)	RUS (1)	97	99	95	97	97	98
	SMOTE(0.0034), RUS(1)		98	99	97	98	98	99
MLP	-	-	100	91	76	83	88	84
	-	RUS (1)	88	100	76	86	88	94
	SMOTE(0.0034), RUS(1)	RUS (1)	94	96	92	94	94	96
	SMOTE(0.0034), RUS(1)		97	97	96	97	97	98

**Sekil 2.** Veri ön işlemsiz şekilde temel metrikler ile metotların karşılaştırılması.

**Çizelge 5.** Veri ön işlem ile yeniden örneklemeye metotlarının kullanıldığı aşamalara göre öğrenme metotlarına etkisi.

Metot	YÖ Metotları		Değerlendirme Metrikleri (%)					
	Eğitim	Test	Doğruluk	Kesinlik	Duyarlılık	F Skor	AUROC	AUPR
XGB	-	-	96	90	79	84	90	86
	-	RUS (1)	90	96	78	86	89	93
	SMOTE(0.1), RUS(1)	RUS (1)	92	94	91	92	92	94
	SMOTE(0.1), RUS(1)		95	96	94	95	95	96
LGBM	-	-	96	44	38	41	69	43
	-	RUS (1)	68	96	38	54	68	85
	SMOTE(0.1), RUS(1)	RUS (1)	92	94	90	92	92	93
	SMOTE(0.1), RUS(1)		95	96	94	95	95	96
Stacking (XGB, FNN)	-	-	95	92	78	84	88	86
	-	RUS (1)	86	95	73	83	85	92
	SMOTE(0.1), RUS(1)	RUS (1)	91	94	89	91	91	93
	SMOTE(0.1), RUS(1)		95	96	95	95	95	96
RF	-	-	96	92	79	85	88	87
	-	RUS (1)	89	96	78	86	89	95
	SMOTE(0.1), RUS(1)	RUS (1)	90	94	86	90	90	91
	SMOTE(0.1), RUS(1)		94	96	93	94	94	95
kNN	-	-	94	92	79	85	90	86
	-	RUS (1)	89	94	79	86	88	94
	SMOTE(0.1), RUS(1)	RUS (1)	92	92	89	90	92	93
	SMOTE(0.1), RUS(1)		93	94	92	93	93	95
SVM	-	-	94	91	67	77	81	81
	-	RUS (1)	81	94	67	78	82	89
	SMOTE(0.1), RUS(1)	RUS (1)	89	93	85	89	89	91
	SMOTE(0.1), RUS(1)		95	98	91	94	95	97
LR	-	-	96	86	62	72	80	74
	-	RUS (1)	80	96	66	78	79	90
	SMOTE(0.1), RUS(1)	RUS (1)	92	94	91	92	92	94
	SMOTE(0.1), RUS(1)		95	96	93	95	95	97
FNN	-	-	97	85	81	83	89	84
	-	RUS (1)	89	97	82	89	89	95
	SMOTE(0.1), RUS(1)	RUS (1)	95	95	95	95	95	97
	SMOTE(0.1), RUS(1)		97	98	97	97	97	99
MLP	-	-	96	89	76	82	88	84
	-	RUS (1)	88	96	77	85	85	92
	SMOTE(0.1), RUS(1)	RUS (1)	93	95	92	93	93	95
	SMOTE(0.1), RUS(1)		96	97	96	96	96	96

**Şekil 3.** Veri ön işlemleri şeklinde temel metrikler ile metotların karşılaştırılması.

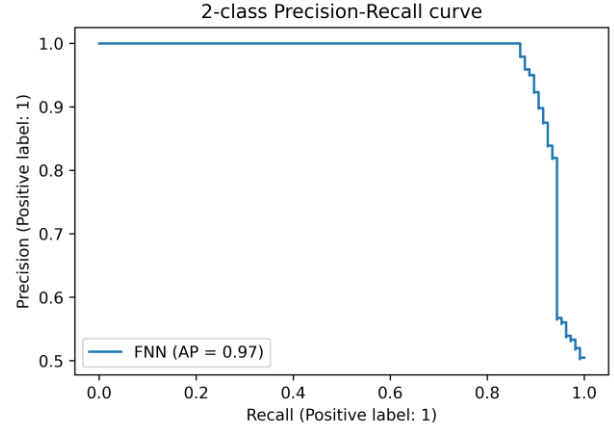


**Çizelge 6.** Karmaşıklık matrisi veri ön işlemsiz FNN

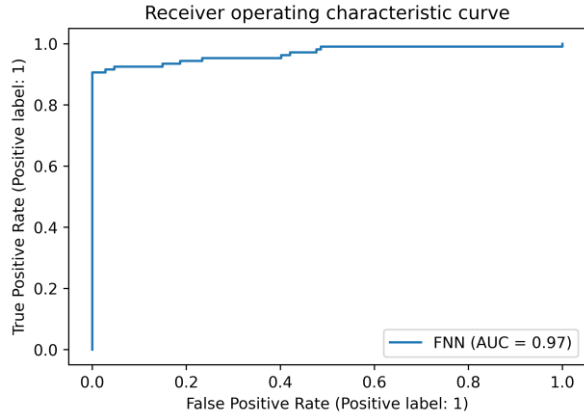
		Gerçek	
		1	0
Tahmin	1	102	1
	0	5	106

**Çizelge 7.** Karmaşıklık matrisi veri ön işlemlili FNN

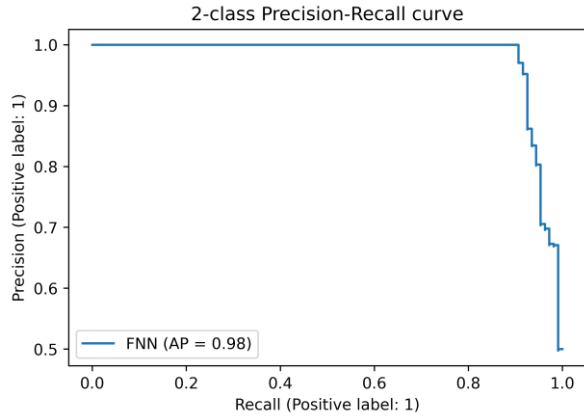
		Gerçek	
		1	0
Tahmin	1	106	1
	0	1	106



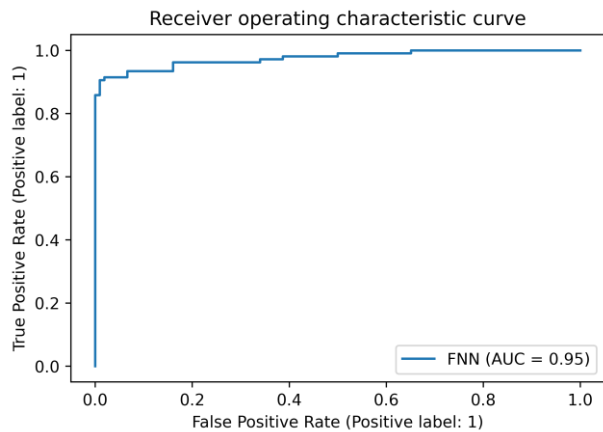
**Şekil 7.** PR eğrisi veri ön işlemlili FNN



**Şekil 4.** ROC eğrisi veri ön işlemsiz FNN



**Şekil 5.** PR eğrisi veri ön işlemsiz FNN



**Şekil 6.** ROC eğrisi veri ön işlemlili FNN

## 5. Kaynaklar

Alam, T.M., Shaukat, K., Hameed, I.A., Luo, S., Sarwar, M.U., Shabbir, S., Li, J. and Khushi, M., 2020. An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access*, **8**, 201173-201198.

Aung, M.H., Seluka, P.T., Fuata, J.T.R., Tikoisuva, M.J., Cabealawa, M.S. and Nand, R., 2020. Random Forest Classifier for Detecting Credit Card Fraud based on Performance Metrics. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (1-6).

Bej, S., Davtyan, N., Wolfien, M., Nassar, M. and Wolkenhauer, O., 2021. LoRAS: an oversampling approach for imbalanced datasets. *Machine Learning*, **110(2)**, 279-301.

bin Alias, M.S.A., Ibrahim, N.B. and Zin, Z.B.M., 2021. Improved sampling data Workflow using Smtmk to increase the classification accuracy of imbalanced dataset. *European Journal of Molecular & Clinical Medicine*, **8(02)**, 2021.

Breiman, L., 2001. Random forests. *Machine learning*, **45(1)**, 5-32.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, **16**, 321-357.

Chen, T. and Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (785-794).

Cover, T. and Hart, P., 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, **13(1)**, 21-27.

- Çinarer, G., Emiroğlu, B.G. and Yurttakal, A.H., 2021. Predicting 1p/19q chromosomal deletion of brain tumors using machine learning. *Emerging Materials Research*, **10(2)**, 238-244.
- Efron, B., 1982. *The jackknife, the bootstrap and other resampling plans*. Society for industrial and applied mathematics.
- Garg, R., Oh, E., Naidech, A., Kording, K. and Prabhakaran, S., 2019. Automating ischemic stroke subtype classification using machine learning and natural language processing. *Journal of Stroke and Cerebrovascular Diseases*, **28(7)**, 2045-2051.
- Gulati, P., 2020. Hybrid resampling technique to tackle the imbalanced classification problem.
- Ito, F. and Singh, S., 2021. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, **13(4)**, 1503-1511.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, **30**.
- McCulloch, W.S. and Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5(4)**, 115-133.
- Mînaştireanu, E.A. and Meşniţă, G., 2020. Methods of handling unbalanced datasets in credit card fraud detection. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, **11(1)**, 131-143.
- Mochida, K., Koda, S., Inoue, K., Hirayama, T., Tanaka, S., Nishii, R. and Melgani, F., 2019. Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. *GigaScience*, **8(1)**, giy153.
- Mrozek, P., Panneerselvam, J. and Bagdasar, O., 2020, December. Efficient resampling for fraud detection during anonymised credit card transactions with unbalanced datasets. In *2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC)* (426-433). IEEE.
- Nguyen, T.T., Tahir, H., Abdelrazek, M. and Babar, A., 2020. Deep learning methods for credit card fraud detection. *arXiv preprint arXiv:2012.03754*.
- Riffi, J., Mahraz, M.A., El Yahyaouy, A. and Tairi, H., 2020. Credit card fraud detection based on multilayer perceptron and extreme learning machine architectures. In *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)* (1-5). IEEE.
- Rtayli, N. and Enneya, N., 2020. Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization. *Journal of Information Security and Applications*, **55**, 102596.
- Shah, H.B., 2020. Comparing Machine Learning Algorithms For Credit Card Fraud Detection.
- Shamsudin, H., Yusof, U.K., Jayalakshmi, A. and Khalid, M.N.A., 2020. Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset. In *2020 IEEE 16th International Conference on Control & Automation (ICCA)* (803-808). IEEE.
- Shivanna, A., Ray, S., Alshouli, K. and Agrawal, D.P., 2020. Detection of Fraudulence in Credit Card Transactions using Machine Learning on Azure ML. In *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (0268-0273). IEEE.
- Tingfei, H., Guangquan, C. and Kuihua, H., 2020. Using variational auto encoding in credit card fraud detection. *IEEE Access*, **8**, 149841-149853.
- Tran, T.C. and Dang, T.K., 2021. Machine learning for prediction of imbalanced data: Credit fraud detection. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (1-7). IEEE.
- Vapnik, V.N., The nature of static learning theory M. New York: Springer—Verlag, **1**, 995.
- Wang, J., de Moraes, R.M. and Bari, A., 2020. A predictive analytics framework to anomaly detection. In *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)* (104-108). IEEE.
- Wibowo, P. and Fatichah, C., 2021. An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, **7(1)**, 63-71.
- Wright, R.E., 1995. Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics*, 217–244. American Psychological Association.
- Zhang, D., Bhandari, B. and Black, D., 2020. Credit Card Fraud Detection Using Weighted Support Vector Machine. *Applied Mathematics*, **11(12)**, 1275.

**İnternet kaynakları**

- 1- [https://nilsonreport.com/content\\_promo.php?id\\_promo=16](https://nilsonreport.com/content_promo.php?id_promo=16), (28.01.2022)
- 2- [https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-2019/consumer\\_sentinel\\_network\\_data\\_book\\_2019.pdf](https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-2019/consumer_sentinel_network_data_book_2019.pdf), (28.01.2022)
- 3- <https://www.kaggle.com/mlg-ulb/creditcardfraud>, (28.01.2022)