

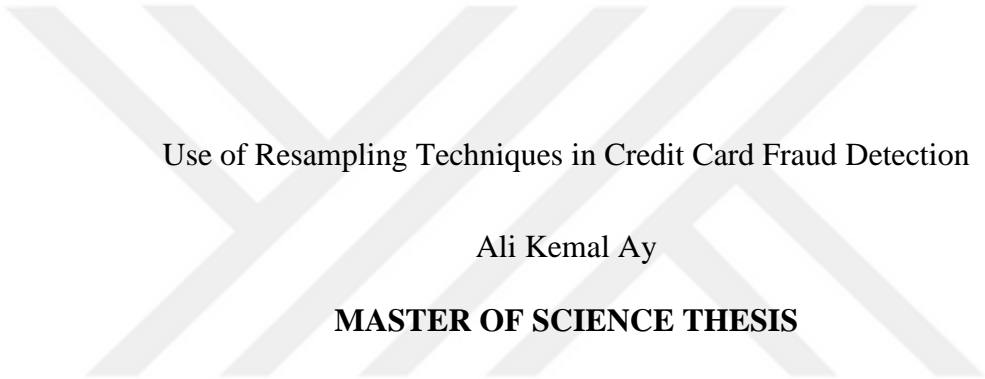
Kredi Kartı Dolandırıcılığının Tespitinde Yeniden Örnekleme Tekniklerinin Kullanımı

Ali Kemal Ay

**YÜKSEK LİSANS TEZİ**

Bilgisayar Mühendisliği Anabilim Dalı

Temmuz 2022



Use of Resampling Techniques in Credit Card Fraud Detection

Ali Kemal Ay

**MASTER OF SCIENCE THESIS**

Department of Computer Engineering

July 2022

# Kredi Kartı Dolandırıcılığının Tespitinde Yeniden Örneklemme Tekniklerinin Kullanımı

Ali Kemal Ay

Eskişehir Osmangazi Üniversitesi  
Fen Bilimleri Enstitüsü  
Lisansüstü Yönetmeliği Uyarınca  
Bilgisayar Mühendisliği Anabilim Dalı  
Bilgisayar Donanımı Bilim Dalında  
YÜKSEK LİSANS TEZİ  
Olarak Hazırlanmıştır

Danışman: Dr. Öğr. Üyesi Esra N. Yolaçan

Temmuz 2022

## ETİK BEYAN

Eskişehir Osmangazi Üniversitesi Fen Bilimleri Enstitüsü tez yazım kılavuzuna göre Dr. Öğr. Üyesi Esra N. Yolaçan danışmanlığında hazırlamış olduğum “Kredi Kartı Dolandırıcılığının Tespitinde Yeniden Örneklem Tekniklerinin Kullanımı” başlıklı Yüksek Lisans tezimin özgün bir çalışma olduğunu; tez çalışmamın tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; tezimde verdiğim bilgileri, verileri akademik ve bilimsel etik ilke ve kurallara uygun olarak elde ettiğimi; tez çalışmamda yararlandığım eserlerin tümüne atıf yaptığımı ve kaynak gösterdiğimi ve bilgi, belge ve sonuçları bilimsel etik ilke ve kurallara göre sunduğumu beyan ederim. 28/07/2022

Ali Kemal AY  
İmza

## ÖZET

Gelişmiş ve gelişmekte olan çoğu ülkede elektronik alışveriş hareketleri için kredi kartı popüler ödeme metotlarından biri haline gelmiştir. Kredi kartı ile ödeme işlemleri daha kolay ve kullanışlı bir şekilde gerçekleşmektedir. Diğer taraftan yeni dolandırıcılıklar için açıklar doğmuş ve buna bağlı olarak dolandırıcılık oranları ve kayıpları artmıştır. Oluşan bu kayıpları önleyebilmek için yapılan bir alışveriş hareketinin kötü amaçlı olup olmadığının tespitinin hızlı bir şekilde yapılabilmesi önemlidir. Ancak dolandırıcılık tespiti modelini geliştirebilmek için gerekli olan kredi kartı dolandırıcılık veri kümeleri dengesizdir. Bu çalışmada yeniden örnekleme metotları uygulandıkları aşamalara göre incelenmiş ve dengesizlik problemi için sentetik azınlık örnekleme tekniği ve rastgele az örnekleme metotları kullanılmıştır. Bununla birlikte dolandırıcılık tespit sistemlerinin hızlı bir şekilde işlem yapabilmesi gerekmektedir. Bu nedenle işlem hızını artırabilmek için öznitelik seçimi metotları kullanılmaktadır. Lasso Regresyon, Rastgele Orman, Pearson Korelasyon ve Özyinelemeli Öznitelik Eleme öznitelik seçimi metotları bu çalışmada kullanılmıştır. Öte yandan son zamanlarda dolandırıcılık tespiti için Gradyan Artırma yöntemlerinin kullanımı yaygın hale gelmiştir. Bu çalışmada kullanılan Hafif Gradyan Artırma Makinesi ve Aşırı Gradyan Artırma ile Rastgele Orman, Lojistik Regresyon ve Destek Vektör Makineleri gibi gradyan artırma yöntemlerin başarı değerlendirmesinde Duyarlılık, Kesinlik, F1-Skor, İşlem Karakteristik Eğrisi Altındaki Alan, Kesinlik-Duyarlılık Eğrisi Altındaki Alan ve işlem zamanı metriklerinden yararlanılmıştır. Elde edilen sonuçlarda yeniden örnekleme metotlarının eğitim-test ayırımından önce kullanılmasının diğer aşamalara göre daha güvenilir sonuçlar yansıttığı görülmüştür. Bununla birlikte yeniden örnekleme metotlarının en çok Hafif Gradyan Artırma Makinesi metoduna katkısı olduğu görülmüştür. Öznitelik seçimi ile ise başarı değerlendirmesinde en az kayıpla işlem zamanı da en çok azalan metot Rastgele Orman olmuştur.

**Anahtar Kelimeler:** Dengesiz Veri Kümesi, Dolandırıcılık Tespiti, Kredi Kartı, Makine Öğrenmesi, Yeniden Örnekleme, Öznitelik Seçimi

## SUMMARY

Credit card has become one of the popular payment methods for electronic shopping transactions in many developed and developing countries. Payment with credit card is easier and more convenient. On the other hand, deficits have arisen for new fraud types so that fraud rates and losses have increased. To prevent these losses, it is important to quickly determine whether a transaction is fraud or not. However, the credit card fraud datasets required to develop the credit card fraud detection model are imbalanced. In this study, resampling methods were examined according to the stages in which they were applied, and SMOTE and RUS methods were used for the imbalance problem. Due to the fact that the fraud detection system needs to be able to detect quickly, feature selection methods are used to increase the processing speed. Lasso Regression, Random Forest (RF), Pearson Correlation and Recursive Feature Elimination methods were used for feature selection in this study. On the other hand, the use of GB methods for fraud detection has become widespread recently. Recall, Precision, F1-Score, Area Under Receiver Operator Characteristics Curve, Area Under Precision Recall Curve and computing time are used in the success evaluation of gradient boosting methods such as Light Gradient Boosting Machine (LGBM) and Extreme Gradient Boosting and other machine learning methods such as RF, Logistic Regression, and Support Vector Machines. In the results obtained, it was seen that the use of resampling methods before the training-test separation revealed more reliable results than the other stages. Additionally, it was also seen that the resampling methods contributed the most to the LGBM. Along with the feature selection, RF was the method that reduced computing time the most with the least loss in success evaluation.

**Keywords:** Imbalanced Dataset, Fraud Detection, Credit Card, Machine Learning, Resampling, Feature Selection

# İÇİNDEKİLER

## Sayfa

<b>ÖZET.....</b>	<b>v</b>
<b>SUMMARY.....</b>	<b>vi</b>
<b>İÇİNDEKİLER.....</b>	<b>vii</b>
<b>ŞEKİLLER DİZİNİ .....</b>	<b>viii</b>
<b>ÇİZELGELER DİZİNİ.....</b>	<b>ix</b>
<b>SİMGELER VE KISALTMALAR DİZİNİ .....</b>	<b>x</b>
<b>1. GİRİŞ VE AMAÇ.....</b>	<b>1</b>
<b>2. LİTERATÜR ARAŞTIRMASI.....</b>	<b>4</b>
2.1. Erişilebilir Veri Kümeleri .....	4
2.2. Dengesizlik Problemi ve Yeniden örnekleme.....	7
2.2.1. Yeniden örnekleme metotları .....	7
2.2.2. Literatür incelemesi.....	9
2.3. Öznitelik Seçimi .....	11
2.3.1. Öznitelik seçimi metotları .....	11
2.3.2. Literatür incelemesi.....	13
2.4. Makine Öğrenmesi ve Değerlendirme Metrikleri .....	14
2.4.1. Makine öğrenmesi metotları ve değerlendirme metrikleri .....	15
2.4.2. Literatür incelemesi.....	19
2.5. Topluluk Öğrenmesi .....	25
2.5.1. Topluluk öğrenmesi yöntemleri .....	26
2.5.2. Literatür incelemesi.....	32
<b>3. MATERYAL VE YÖNTEM .....</b>	<b>34</b>
3.1. Materyal.....	34
3.1.1. Test ortamı .....	34
3.1.2. European Cardholders veri kümesi .....	34
3.2. Yöntem .....	35
3.2.1. Yeniden örnekleme .....	36
3.2.2. Öznitelik seçimi .....	37
3.2.3. Makine öğrenmesi ve değerlendirme metrikleri.....	40
<b>4. BULGULAR VE TARTIŞMA .....</b>	<b>42</b>
4.1. Yeniden Örnekleme .....	42
4.2. Öznitelik Seçimi .....	45
<b>5. SONUÇ VE ÖNERİLER .....</b>	<b>50</b>
<b>KAYNAKLAR DİZİNİ.....</b>	<b>51</b>

## ŞEKİLLER DİZİNİ

<b><u>Sekil</u></b>	<b><u>Sayfa</u></b>
1.1. Makine öğrenmesi tabanlı kredi kartı dolandırıcılık tespiti.....	2
2.1. SMOTE metodu ( $k = 3$ ) (Yavaş vd.'den, 2020) .....	8
2.2. En iyi öznitelik alt kümesi seçimi .....	12
2.3. Hata matrisi .....	17
2.4. ROC grafiği (Buitinck vd.'den, 2013).....	18
2.5. PR grafiği (Buitinck vd.'den, 2013) .....	19
2.6. Karar ağacı örneği .....	27
2.7. Yinelemeli rastgele ile veri kümelerinin oluşturulması.....	29
2.8. RF metodu çoğunluk oylama yapısı .....	30
3.1. Sınıf dağılımı .....	35
3.2. Öznitelik korelasyon grafiği .....	39
4.1. Metotların ROC Eğrileri.....	49
4.2. Metotların PR Eğrileri .....	49



## ÇİZELGELER DİZİNİ

<b><u>Cizelge</u></b>	<b><u>Sayfa</u></b>
1.1. 2030 yılına kadar öngörülen kart dolandırıcılığı (Nilson Report'dan, 2021).....	1
2.1. Kredi kartı dolandırıcılık türleri (Dal Pozzolo'dan, 2015).....	4
2.2. Erişilebilir kredi kartı dolandırıcılık veri kümeleri ve özellikleri.....	5
2.3. Erişilebilir kredi kartı dolandırıcılık veri kümelerinin önemli öznitelikleri.....	6
2.4. ROS örneği (örnekleme oranı: 0,5) .....	8
2.5. RUS örneği (örnekleme oranı: 0,5) .....	9
2.6. Kredi kartı dolandırıcılık tespitinde Eğitim-Test ayırımına göre yeniden örnekleme metotları .....	10
2.7. European Cardholders veri kümesi ile yapılan çalışmalarda kullanılan makine öğrenmesi metotları.....	20
2.8. European Cardholders veri kümesi ile yapılan çalışmalarda kullanılan değerlendirme metrikleri .....	23
2.9. Kredi kartı dolandırıcılık veri kümesi .....	27
2.10. Kök düğüm için karar kuralları kümesi.....	28
3.1. European Cardholders veri kümesi öznitelikleri .....	35
3.2. Kullanılan örnekleme yöntemi ve oranları .....	36
3.3. Öznitelik seçimi ile oluşturulan öznitelik alt kümeleri.....	39
4.1. Yeniden örnekleme ile veri miktarlarındaki değişim (Acc: Accuracy, Pre: Precision, Rec: Recall, F1: F1-Score, S: SMOTE, R: RUS) .....	42
4.2. Yeniden örnekleme ile veri miktarlarındaki değişim .....	43
4.3. Yeniden örnekleme metotlarının kullandığı aşamalara göre makine öğrenmesi metotlarına etkisi .....	45
4.4. Pearson Korelasyon öznitelik seçimi ile makine öğrenmesi metotları.....	46
4.5. RF öznitelik seçimi ile makine öğrenmesi metotları .....	47
4.6. Lasso Regresyon öznitelik seçimi ile makine öğrenmesi metotları .....	48
4.7. RFE ile makine öğrenmesi metotları .....	48

## SİMGELER VE KISALTMALAR DİZİNİ

<b><u>Kısaltmalar</u></b>	<b><u>Açıklama</u></b>
RF	Rastgele Orman (Random Forest)
LR	Lojistik Regresyon (Logistic Regression)
SVM	Destek Vektör Makinesi (Support Vector Machine)
KNN	K en yakın komşu (K Nearest Neighbour)
DT	Karar Ağacı (Decision Tree)
NB	Naïve Bayes
NN	Sinir Ağı (Neural Network)
MLP	Çok Katmanlı Algılayıcılar (Multilayer Perceptron)
AE	Oto Kodlayıcı (Auto Encoder)
GB	Gradyan Artırma (Gradient Boosting)
LGBM	Hafif Gradyan Artırma Makinesi (Light Gradient Boosting Machine)
XGB	Aşırı Gradyan Artırma (Extreme Gradient Boosting)
RFE	Özyinelemeli Öznitelik Eleme (Recursive Feature Elimination)
ROS	Rastgele Aşırı Örneklem (Random Oversampling)
RUS	Rastgele Az Örneklem (Random Undersampling)
SMOTE	Sentetik Azınlık Örneklem Tekniği (Synthetic Minority Oversampling Technique)
AUROC	Alıcı İşlem Karakteristik Eğrisi Altındaki Alan (Area Under Receiver Operator Characteristics Curve)
AUPRC	Kesinlik-Duyarlılık Eğrisi Altındaki Alan (Area Under Precision Recall Curve)
MCC	Matthews Korelasyon Katsayısı (Matthews Correlation Coefficient)
FPR	Yanlış Pozitif Oranı (False Positive Rate)
G-Mean	Geometrik Ortalama (Geometric Mean)
FK	Finansal Kurtarma (Financial Recovery)

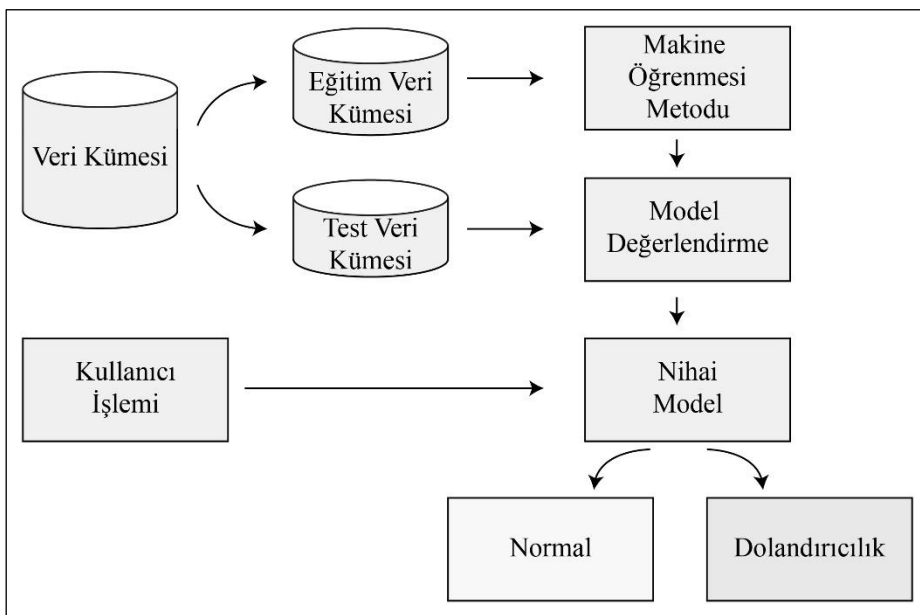
## 1. GİRİŞ VE AMAÇ

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. Nilson Report (2021) verilerine göre, 2020 yılında, dünya çapındaki dolandırıcılık kayıpları 28,6 milyar dolara ulaşmış ve bu rakamın 2030 yılına kadar 49,3 milyar dolara ulaşacağı tahmin edilmektedir (Şekil 1.1). Federal Ticaret Komisyonu'nun (Federal Trade Commission (2022)) raporladığı ödeme yöntemlerine göre başlıca dolandırıcılık türleri ve bunlara karşılık gelen kayıpları içeren grafikte, yaklaşık 2,7 milyon dolandırıcılık ve bunun yaklaşık 440 bin (%16) kadarının ödeme yöntemleri ile ilgili olduğu belirtilmiştir. Son yıllarda yükselen kripto para birimlerinin etkisiyle, listede 6. sıraya gerileyen kredi kartı dolandırıcılığı, 181 milyon dolar kayba sahip yaklaşık 88 bin bildiri ile bu listede yer almaktadır.

Çizelge 1.1. 2030 yılına kadar öngörülen kart dolandırıcılığı (Nilson Report'dan, 2021)

2030 Yılına Kadar Öngörülen Kart Dolandırıcılığı			
Yıl	Toplam Hacim	Dolandırıcılık	100 dolar
	Trilyon dolar	Milyar dolar	hacimdeki sent
2020	41,692	28,58	6,81
2021	47,229	32,20	6,82
2022	50,868	34,36	6,75
2023	54,061	36,13	6,68
2024	57,323	38,07	6,64
2025	60,583	39,89	6,58
2026	64,038	41,73	6,52
2027	67,570	43,76	6,48
2028	71,221	45,54	6,39
2029	75,111	47,50	6,32
2030	79,140	49,32	6,23

Maliyet, hız, güvenlik gibi unsurlar ve büyük veri ve makine öğrenmesi gibi teknolojilerin de gelişmesi göz önünde bulundurulduğunda kredi kartı dolandırıcılık tespiti işlemlerinin insana bağımlı olabileceğince kopararak otomatik sistemler tarafından gerçekleştirilmesinin gerekliliği fark edilmektedir. Kredi kartı ile satın alma işlemi, POS (Point of Sale) cihazlarına kart bilgilerinin çeşitli yöntemlerle (çekilerek, takılarak veya temassız) okutulması ya da internet üzerinden kullanıcının bu bilgileri el ile girmesiyle de gerçekleştirilmektedir. Yeterli bakiyenin olması, kart şifresinin doğruluğu gibi çeşitli bilgilerin sistem tarafından doğruluğu tespit edildikten sonra alışveriş başarıyla gerçekleştirilmektedir. Dolandırıcılar ise gerekli olan bu kart bilgilerini elde ederek kendi amaçları doğrultusunda harcama yapabilmektedirler. Dolayısıyla yapılan işlemin kim tarafından gerçekleştirildiğinin tespiti önemlidir. Bu noktada devreye giren makine öğrenmesi metotları şu şekilde çalışmaktadır: İlk olarak kredi kartı işlemlerinden oluşan bir veri kümesi eğitim ve test kısımlarına bölünmektedir. Sonrasında seçilen makine öğrenmesi metodu eğitim veri kümesi ile eğitilmektedir. Önceden görmediği test veri kümesi ile geliştirilen model yaptığı tahminlere göre başarısı değerlendirilmektedir. Eğer bu başarı hedeflenen düzeyde ise model kullanıma dağıtılabilmektedir. Son olarak ise gerçekleştirilen kredi kartı işlemleri modele gönderilir ve model bu işlemlerin normal ya da dolandırıcılık olduğunu tahmin eder (Şekil 1.2). Ayrıca gerçekleştirilen bir işlem zamanında doğru tespit edilmese bile mümkün olan en kısa sürede yenilenerek, sisteme geri bildirimde bulunulması sonraki dolandırıcılıkların tespiti için ileri derecede önemlidir.



Şekil 1.1. Makine öğrenmesi tabanlı kredi kartı dolandırıcılık tespiti

Makine öğrenmesi tabanlı kredi kartı dolandırıcılık tespiti sisteminin oluşturulurken karşılaşılan problemlerden birisi model geliştirebilmek için yeterli bir veri kümesinin olmamasıdır. İlgili literatür incelendiğinde çeşitli mahremiyet kaygılarından dolayı erişilebilir veri kümesi bağlamında bir yetersizlik olduğu görülmüştür. Bununla birlikte erişilebilir veri kümeleri ise dengesizdir (Sisodia vd., 2017). Dengesizlik problemi veri sayısı az olan sınıfın makine öğrenmesi metotları tarafından yeterince öğrenilmemesine neden olabilmektedir. Bir başka durum ise sistemin hangi öznitelik seçimi (feature selection), yeniden örnekleme (resampling), makine öğrenmesi ve değerlendirme metriklerini kullanacağına karar verme ihtiyacıdır.

Yürütülen bu çalışmanın katkıları şunlardır:

- Kredi kartı dolandırıcılık tespiti alanında erişilebilir veri kümeleri incelenmiş ve karşılaşılan problemler değerlendirilmiştir.
- Yeniden örnekleme metotlarının kullanıldığı öğrenme aşamalarına göre etkileri incelenmiştir.
- Öznitelik seçimi metotları ile oluşturulan 4 farklı öznitelik alt kümesinin makine öğrenmesi metotlarına etkileri paylaşılmıştır.
- Yeniden örnekleme ve öznitelik seçimi metotlarının en iyi sonucu veren kombinasyonu ile Hafif Gradyan Artırma Makinesi (Light Gradient Boosting Machine (LGBM)), Aşırı Gradyan Artırma (Extreme GB), Rastgele Orman (Random Forest (RF)), Lojistik Regresyon (Logistic Regression (LR)) ve Destek Vektör Makinesi (Support Vector Machine (SVM)) gibi öğrenme metotları eğitilip sonrasında test sonuçları paylaşılmıştır.

Çalışmanın düzeni olarak ilk kısımda kredi kartı dolandırıcılıklarının oluşturduğu kayıplar, dolandırıcılık tespitindeki problemler ve çalışmanın amacını ele alınmıştır; İkinci kısımda ise literatürde erişilebilir veri kümeleri ve yeniden örnekleme, öznitelik seçimi ve makine öğrenmesi metotlarına göre ilgili literatür detaylıca incelenmiştir ve bu metotlar hakkında arka plan bilgileri verilmiştir; üçüncü kısımda bu çalışmada kullanılan veri kümesi, topluluk öğrenmesi, değerlendirme metriklerinin neden ve nasıl kullanıldığı hakkında bilgiler paylaşılmıştır; dördüncü kısımda ise elde edilen bulgular ve tartışmalı durumlar değerlendirilmiş; son kısımda ise bu sonuçlar ve gelecek çalışmalar için öneriler sunulmuştur.

## 2. LİTERATÜR ARAŞTIRMASI

Bu bölümde öncelikle kredi kartı dolandırıcılık türleri paylaşılmaktadır. Devamında kredi kartı dolandırıcılık tespiti literatüründe erişilebilir veri kümeleri incelenmektedir. Bu inceleme sonrasında yeniden örnekleme, öznelilik seçimi gibi veri ön işlemleri ele alınmakta ve takibinde makine öğrenmesi, değerlendirme metrikleri ve topluluk öğrenmesi hakkında bilgiler ve ilgili literatür verilmektedir.

Kredi kart dolandırıcılık tespiti alanında uzmanlar tarafından belirlenen dolandırıcılık türleri Çizelge 2.1’de verilmiştir (Dal Pozzolo, 2015). Çizelge 2.1’de görüldüğü üzere, dolandırıcılıkların en çok gerçekleştiği tür olan kartsız dolandırıcılık türü e-ticaret ile gerçekleşen işlemlere denk gelmektedir. Bu tür dolandırıcılıklar, veri tabanı saldırılarında elde edilen kart numarası, son kullanım tarihi ve Kart Onay Numarası bilgileri ile gerçekleşmektedir. Birçok tüccar bu sorunun önüne 3D SECURE yöntemiyle geçmektedir (Lucas, 2019).

Çizelge 2.1. Kredi kartı dolandırıcılık türleri (Dal Pozzolo’dan, 2015)

Tür	Dolandırıcılık İşlemlerinde Sahip Olduğu Oran
Kartsız dolandırıcılıklar	>%90
Kimlik hırsızlığı, sahte kartlar	<%10
Teslim edilememiş kartlar	<%1
Kayıp ya da çalıntı kartlar	≈%1

### 2.1. Erişilebilir Veri Kümeleri

Makine öğrenmesi yöntemleri kullanılarak bir model geliştirebilmek için en temel gereksinim bir veri kümesidir. Veri kümesi ile geliştirilen model gerçekleştirilecek kredi kartı işlemlerini sınıflandırabilmektedir. Kredi kartı dolandırıcılık tespiti literatürü incelendiğinde çeşitli veri kümelerinin kullanıldığı görülmektedir ancak bunların

çoğunluğuna erişilememektedir. Sadece erişilebilir veri kümelerinin sahip olduğu özellikler ve öznitelikler ise bu bölümün devamında incelenmektedir.

Kredi kartı dolandırıcılık tespiti alanında en çok karşılaşılan erişilebilir veri kümeleri ve bu veri kümelerinin bazı özellikleri Çizelge 2.2’de gösterilmektedir. Yapılan araştırmalar sonucunda kredi kartı dolandırıcılık tespiti literatüründe ikisi gerçek diğer ikisi sentetik verilerden oluşan toplam da dört adet erişilebilir veri kümesi olduğu tespit edilmiştir. Kredi kartı dolandırıcılık veri kümelerindeki önemli öznitelikleri elde etmek için bu veri kümeleri incelenmiştir. Bu incelemelerin sonucu Çizelge 2.3’te sunulmuştur. Ayrıca burada incelenen veri kümeleri sadece kredi kartı dolandırıcılık tespiti ile ilgilidir. Örneğin Dua ve Graff (2019) gibi, kredi onayıyla ilgili veri kümeleri incelenmemiştir.

Çizelge 2.2. Erişilebilir kredi kartı dolandırıcılık veri kümeleri ve özellikleri

Özellikler	Veri Kümeleri			
	European Cardholders (Kaggle, 2017 a)	Synthetic Financial Datasets For Fraud Detection (Kaggle, 2017 b)	IEEE-CIS Fraud Detection (Kaggle, 2019)	Credit Card Transactions Fraud Detection (Kaggle, 2020)
Yıl	2017	2017	2019	2020
Veri Tipi	Gerçek	Sentetik	Gerçek	Sentetik
Örnek Sayısı	248 bin	24 milyon	1 milyon	1,8 milyon
Öznitelik Sayısı	31	11	433	23
Elde Edilme Süresi	2 gün	30 gün	30 gün	2 yıl
Sahtecilik Oranı	% 0,17	Belirtilmemiş	% 3,62	Belirtilmemiş
Sınıf Sayısı	İkili	İkili	İkili	İkili

Çizelge 2.3. Erişilebilir kredi kartı dolandırıcılık veri kümelerinin önemli öznitelikleri

Öznitelikler	Veri Kümeleri			
	European Cardholders	Synthetic Financial Datasets For Fraud Detection	IEEE-CIS Fraud Detection	Credit Card Transactions Fraud Detection Dataset
İşlem Tutarı	✓	✓	✓	✓
İşlem Tarihi	✓	✓	✓	✓
Satıcı İsmi		✓	✓	✓
Satıcı Konumu			✓	✓
Kullanıcı Adresi			✓	✓
Satıcı Kategorisi				✓
Kullanıcı Kredi Kart Numarası				✓
Diğer Kart Bilgileri			✓	

Kredi kartı dolandırıcılık alanında karşılaşılabilecek ilk problemlerden birisi bireylerin mahremiyet nedenlerinden dolayı veri kümelerinde paylaşılan öznitelik isimlerinin hatta verilerin bir şekilde gizlenmesi, maskelenmesidir. Aslında bu durum kullanıcı tarafından problem olmamakla beraber dolandırıcılık tespiti amacıyla model geliştirecek taraflar için tek taraflı problemler doğurmaktadır. Öz niteliklerin gizlenmesi probleminin önüne, özel veri kümesine sahip kurumların birlikte çalıştığı ekiplere verileri ön işlemden geçirmeden direkt vermesiyle bir miktar geçilebilir. Ancak verilerin ön işlemden geçirilmeden direkt verilmesi kayıp veya gereksiz verilerle taşan veri kümesi problemini yaratabilmektedir. Bu tip veri kümeleri ön işlemlerle daha kullanışlı hale gelebilmektedir. Karşılaşılabilecek problemlerden bir başkası erişilebilir veri kümelerinde dengesizlik problemidir. Dengesizlik probleminin önüne çeşitli örnekleme metotları ile geçilmektedir. Bir başka durum ise model tahminlerinde daha tutarlı olabilmek adına kredi kartı dolandırıcılık tespiti alanında uzman kişilerden yararlanabilmektir. Öncesinde değersiz görülen özniteliklerin uzmanlar aracılığıyla yeniden incelenmesi ve çeşitli istatistiksel özniteliklerin var olanlardan yaratılması gerekebilmektedir (Rushin, 2017; Lucas, 2019).



Sonuç olarak, bu problemler dikkate alınarak geliştirilecek yeni veri kümeleri ile daha başarılı model gelişimi sağlanması beklenmektedir.

## **2.2. Dengesizlik Problemi ve Yeniden örnekleme**

Bu kısımda dengesizlik problemi nedir, nasıl çözülebilir gibi sorular yanıtlanmıştır. Sonrasında ise dengesizlik problemini gidermek için kullanılan yeniden örnekleme metotlarına değinilmiş ve takibinde literatürde bu metotları kullanan çalışmalar incelenmiştir.

Kredi kartı dolandırıcılık tespiti alanında elde edilen erişilebilir veri kümeleri dengesizdir (Bkz. Çizelge 2.2). Dengesizlik veri kümesindeki bir sınıfın diğer sınıfı ağır bir şekilde domine etmesi durumudur. Dengeli veri kümeleri ile çalışacağı varsayılarak geliştirilen makine öğrenmesi yöntemleri bu veri kümeleri ile öğrenme problemi yaşamaktadır. Dolayısıyla eğitim sürecine geçilmeden önce bu problemin çözülmesi gerekmektedir. Bu aşamada ise yeniden örnekleme metotlarından yararlanılmaktadır.

### **2.2.1. Yeniden örnekleme metotları**

Yeniden örnekleme metotları dengesiz veri kümesindeki karakteristiği bozmadan veri miktarını azaltabilmek ya da karakteristiği daha net elde edebilmek için veri miktarını çoğaltabilmek gibi çeşitli amaçlarda kullanılır (Efron, 1982). Bölümün devamında Rastgele Aşırı Örnekleme (Random Oversampling (ROS)), Sentetik Azınlık Örnekleme Tekniği (Synthetic Minority Oversampling Technique (SMOTE)) ve Rastgele Az Örnekleme (Random Undersampling (RUS)) metotları verilmektedir.

ROS, bir sınıftaki verileri çoğaltmaya yönelik bir yöntemdir (Drummond ve Holte, 2003). Çizelge 2.4'te gösterildiği üzere, denge sağlamak amacıyla azınlık sınıfı verileri rastsal seçilerek belirlenen denge oranı elde edilene kadar devam eder. Denge oranı çoğaltılan azınlık sınıfının orijinal çoğunluk sınıfına oranıdır. Oran 1'e yaklaştıkça veri kümesi daha dengeli hale gelmektedir. Ayrıca çoğaltılmak için önceden seçilen veri, veri kümesinden çıkarılmaz dolayısıyla aynı verinin tekrar tekrar seçilme ihtimali mevcuttur.

Çizelge 2.4. ROS örneği (örnekleme oranı: 0,5)

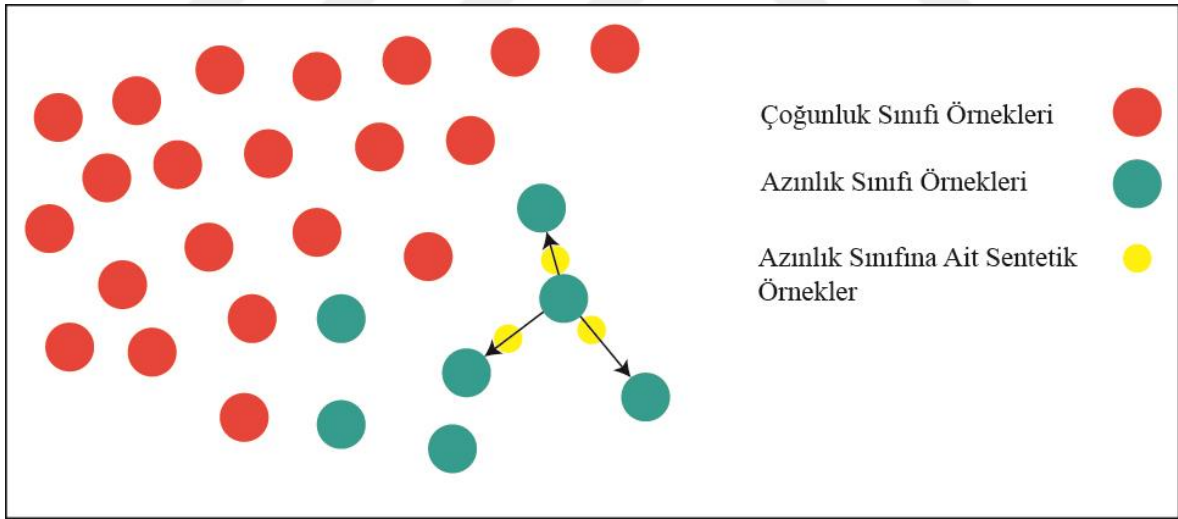
	Çoğunluk	Azınlık
Önce	9160	540
Sonra	9160	4580

SMOTE azınlık sınıfı için sentetik veri üretme tekniğidir (Chawla vd., 2002). ROS metodu ile arasındaki fark verileri kopyalamak yerine sentetik benzerlerini üretmektir. Aşağıdaki adımlar ile çalışma yapısı açıklanmaktadır (Şekil 2.1).

A (Azınlık sınıfı) içerisindeki her  $v$  verisi için:

1. A içindeki  $k$  tane yakın komşu hesaplanır.
2.  $r \leq k$  olmak şartıyla, A içerisinde  $k$  komşudan  $r$  tane rastsal seçim yinelemeli yapılır.
3.  $r$  ve  $v$  arasında rastsal bir nokta seçilir.
4. Bu sentetik veriler A içerisine eklenir.

İstenilen dengeye ulaşılan kadar 1 ile 4 arası adımlar tekrar edilir.

Şekil 2.1. SMOTE metodu ( $k = 3$ ) (Yavaş vd.'den, 2020)

RUS bir sınıftaki verileri azaltmaya yönelik bir yöntemdir (Drummond ve Holte, 2003). Denge sağlamak amacıyla çoğunluk sınıfından veriler rastsal şekilde seçilir ve seçilen bu veriler belirlenen denge oranı elde edilinceye kadar veri kümesinden atılır (Çizelge 2.5). Buradaki denge oranı orijinal azınlık sınıfının azaltılmış çoğunluk sınıfına oranıdır. Oran 1'e yaklaştıkça eşitliğe doğru gidilir.

Çizelge 2.5. RUS örneği (örnekleme oranı: 0,5)

	Çoğunluk	Azınlık
Önce	9160	540
Sonra	1080	540

### 2.2.2. Literatür incelemesi

Bu kısımda European Cardholders veri kümesi ile yapılmış çalışmalar yeniden örnekleme metotlarının kullanıldıkları aşamalara göre Çizelge 2.6'da incelenmiştir. Toplam 20 çalışmanın 4 tanesinde aşırı örnekleme metotları, eğitim, test veri kümesi ayırımından önce uygulanmasıyla beraber diğer 6 çalışmada ne zaman, nasıl uygulandığı bilinmemektedir. Kalan 10 tane çalışma da ise aşırı örnekleme metotları, eğitim, test ayırımından sonra kullanılmaktadır. Örnekleme metotlarının ne zaman uygulandığı bilinmeyen çalışmalar, bilinenlerin yüzdeleri dikkate alınarak dağıtıldığında çalışmaların yaklaşık %30'unda aşırı örnekleme, eğitim test ayırımından önce yapılmıştır denebilir. Aşırı örnekleme metotlarının tüm veri kümesi üzerine uygulanması veya eğitim, test veri kümeleri ayırımı yapıldıktan sonra ayrı ayrı uygulanması durumlarında farklı sonuçlar yaratabilmektedir. Bununla birlikte Çizelge 2.6'da en çok kullanılan az örnekleme metodunun RUS olduğu ve aşırı örnekleme için ise SMOTE olduğu tespit edilmiştir. ROS metodunun SMOTE metoduna kıyasla daha az tercih edilmesinin nedeni direkt kopya veriler üretmesinin sentetik verilere kıyasla verimsiz olmasıdır. Kullanılan diğer yeniden örnekleme metotları ise şunlardır: Uyarlanabilir Sentetik (Adaptive Synthetic (ADASYN)), Sınır Çizgisi-1 SMOTE (Borderline-1 SMOTE), Yerelleştirilmiş Rastgele Afın Gölge Örnekleme (Localized Random Affine Shadowsampling (LoRAS)), Yakın-Kayıp (NearMiss), SVM, Çekişmeli Üretken Ağ (Generative Adversarial Network (GAN)), Mini-Yığın (Mini-Batch) K En Yakın Komşu (K Nearest Neighbour (KNN)). İncelenen çalışmalar dikkate alındığında bu çalışmada da SMOTE ve RUS metotları hibrit bir şekilde kullanılmıştır.

Çizelge 2.6. Kredi kartı dolandırıcılık tespitinde Eğitim-Test ayırımına göre yeniden örnekleme metotları

<b>Makaleler</b>	<b>Metot (Yeniden Örnekleme)</b>	<b>Aşırı örnekleme metotları kullanımı Eğitim-Test ayırımından önce mi?</b>
Alam vd., 2020	ADASYN, RUS, SMOTE	Hayır
Aung vd., 2020	SMOTE	Hayır
Bej vd., 2020	ADASYN, Borderline-1 SMOTE, LoRAS, SMOTE	Hayır
Gulati, 2020	Neighborhood Cleaning Rule, SMOTE,	Hayır
Janbandhu vd., 2020	ADASYN, SMOTE	Hayır
Mînaştireanu ve Meşniță, 2020	ROS, RUS	Hayır
Mrozek vd., 2020	RUS, SMOTE	Belirtilmemiş
Nguyen vd., 2020	NearMiss, RUS, SMOTE	Belirtilmemiş
Riffi vd., 2020	SMOTE	Evet
Rtayli ve Enneya, 2020	SMOTE	Belirtilmemiş
Shah vd., 2020	ROS, RUS, SMOTE	Belirtilmemiş
Shamsudin vd., 2020	ADASYN, Borderline, RUS, SVM-SMOTE	Belirtilmemiş
Shivanna vd., 2020	SMOTE	Evet
Tingfei vd., 2020	GAN, SMOTE	Hayır
Wang vd., 2020	Mini-Batch Undersampling, Oversampling	Hayır
Zhang vd., 2020	KNN, NearMiss, RUS	Hayır
bin Alias vd., 2021	SMOTE, TOMER	Evet
Isabella vd., 2021	ROS	Evet
Tran ve Dang, 2021	ADASYN, SMOTE	Belirtilmemiş
Wibowo ve Fatichah, 2021	ADASYN, Borderline-SMOTE, ROS, SMOTE	Hayır

### 2.3. Öznitelik Seçimi

Bu kısımda öznitelik seçiminin ne olduğu, kullanım nedenleri açıklanmıştır. Sonrasında ise kullanılan öznitelik seçimi yöntemlerine değinilmiş ve takibinde literatürde bu metotları kullanan çalışmalar incelenmiştir.

Geliştirilecek makine öğrenmesi modelinin yukarıda bahsedildiği üzere bir veri kümesine ihtiyacı vardır. Bununla birlikte çeşitli özniteliklere sahip bu veri kümesinde hangi özniteliğin modele katkı sağlayıp sağlamadığını tespit edebilmenin eğitim süresini azaltma, modelin doğruluğunu geliştirebilme, aşırı uyum problemini ve boyut lanetini önleme gibi kazanımları vardır. Öznitelik seçimi metotları üçe ayrılmaktadır. Bunlar:

- Filtreleme
- Sarmalayıcı
- Gömülü metotlardır.

#### 2.3.1. Öznitelik seçimi metotları

Filtreleme metotları özniteliklerin önemini öznitelik ile hedef değişken arasındaki ilişkiye göre hesaplar. Hesaplama sonucunda veri setinde filtreleme yapılır ve ilgili öznitelikler seçilerek bir alt küme oluşturulur. Bu işlemler makine öğrenmesi metotlarından önce yapıldığı için öğrenme aşamasından bağımsızdır ve bu nedenle istenilen öğrenme metoduyla birleştirilebilir (Blum ve Langley, 1997). Ayrıca diğer iki öznitelik seçimi metodu arasında zaman karmaşası bakımından en hızlı metottur. Bu çalışma kapsamında filtreleme metodu olarak en yaygın kullanılan metotlardan biri olan Pearson Korelasyonundan yararlanılmıştır.

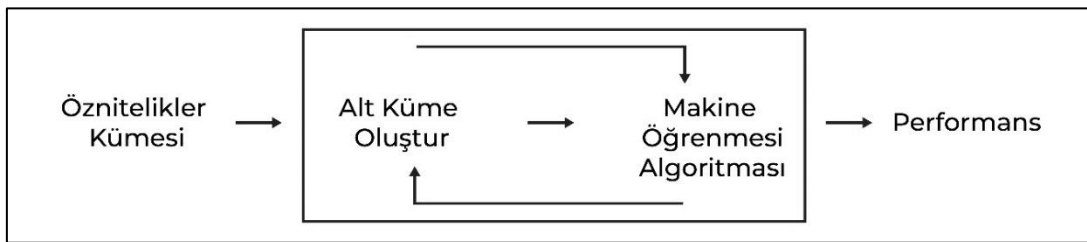
Pearson Korelasyon iki öznitelik arasındaki ilişkiyi gösteren -1 ve 1 arasındaki bir değerdir (Denklem 2.1). Bu değer 0'a yaklaştıkça zayıf korelasyon, 1'e yaklaşıırken pozitif bir korelasyon ve -1'e yaklaşıırken ters bir korelasyon olduğu anlamına gelmektedir.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (2.1)$$

Formülde yer alan parametreler:

- $r$  : korelasyon katsayısı
- $x_i$  : x değişkeni değerleri
- $\bar{x}$  : x değişken değerlerinin ortalaması
- $y_i$  : y değişkeni değerleri
- $\bar{y}$  : y değişken değerlerinin ortalaması

Sarmalayıcı metotlar genel olarak belirli bir makine öğrenmesi metodu kullanarak en iyi öznitelik alt kümesini aramaya (Şekil 2.2) dayanan metotlardır (Blum ve Langley, 1997). İleri ve geri yönlü olarak ikiye ayrılır. İleri yönlü arama metodu, boş bir seçilenler alt kümesi ile başlar ve her bir adımda model tespit işlemine en fazla katkı sağlayan özniteligi seçilenler kümesine ekleyip, orijinal kümeden silerek belirlenen öznitelik sayısına ulaşılan kadar devam eder. Geri yönlü ya da özyinelemeli olarak geçen metot ise tüm özniteliklerden her bir adımda model tespit işlemine en az katkı sağlayan öznitelik silinerek belirlenen öznitelik sayısına ulaşılan kadar devam eder. Bu stratejiler nedeniyle sahip sarmalayıcı metotların hesaplama maliyetleri yüksektir.



Şekil 2.2. En iyi öznitelik alt kümesi seçimi

Gömülü metotlarda öznitelik seçme işlemi modelin öğrenme süreci içerisinde gerçekleşir. Sarmalayıcı metotlara kıyasla daha az işlem maliyeti gerektirir. Bununla birlikte gömülü olarak öznitelik seçimine sahip olan metotlar (RF, LGBM ve XGB) aşırı uyum problemine karşı daha güçlüdür. Bu metotların detaylıca açıklandığı 2.5 Topluluk Öğrenmesi başlığı altında sahip olduğu öznitelik seçimi metotları da açıklanmaktadır.

RF sınıflandırıcısına dayalı bir öznitelik seçimi metodu, elde edilmesi nispeten ucuz ve yüksek boyutlu verilere başarıyla uygulanan çok değişkenli öznitelik önem puanlarını sağlamaktadır (Breiman (2001)). RF rastsal DT'lere dayalı bir topluluk öğrenmesi metodudur ve çeşitli öznitelik önemlilik ölçümleri sağlar. Bunlardan birisi istatistiksel permütasyon testlerinden diğeri modelin eğitiminden türemektedir. RF metodu devasa boyutlu veri kümelerinde üstün bir performansa neden olan gömülü öznitelik seçimi uygulamaktadır. Gini safsızlığı öznitelik önemlilik düzeyinin genel bir göstergesi olarak kullanılabilir. RF modeli kapsamında  $T$ , DT içerisindeki her bir  $n$  düğümünde Gini safsızlığı kullanılarak en iyi bölünme aranır. Her bir öznitelik yerine RF'ın getirdiği kısıtlama ile rastgele seçilen öznitelik alt kümesi ve bunlar için gerekli eşit değeri kapsamlı bir şekilde aranır. Bulunan her bir bölünme için Gini safsızlığındaki azalma kaydedilir ve her DT'deki her bir düğüm için toplanır (Denklem 2.2). Bu sayı bir  $\theta$  özniteliğinin bölünmeler için ne sıklıkta seçildiğini ve ne kadar ayırt edici olduğunu gösterir.

$$I_G(\theta) = \sum_T \sum_n \Delta i_\theta(n, T) \quad (2.2)$$

Tibshirani (1996), tarafından ilk defa önerilen Lasso Regresyon metodu regresyon katsayılarına cezalandırma işlemi için L1 düzenleme fonksiyonu uygulayan cezalı en küçük kareler yöntemidir. L1 fonksiyonu doğası gereği hem küçültme hem de öznitelik seçimi işlemini eş zamanlı gerçekleştirir. Bununla birlikte modern veri analizinde önemi artan öznitelik seçimi metodları arasında Lasso Regresyon metodu seyrek temsili nedeniyle daha tercih edilmektedir.

### 2.3.2. Literatür incelemesi

Dhankhad vd. (2018), kredi kartı dolandırıcılık tespiti için denetimli makine öğrenmesi yöntemleri üzerine kapsamlı bir çalışma yapmışlardır. RF, Karar Ağacı (Decision Tree (DT)), GB, XGB ve LR gibi farklı algoritmalar ile en çok kullanılan 5 öznitelik elde edilmiştir. Eğitim ve test işlemleri de bulunan bu öznitelikler üzerinden gerçekleştirilmiştir. Ayrıca tüm algoritmalar  $V14$ 'ü bu sıralamaya dahil ettiği için temel öznitelik olarak varsayılmıştır. Bununla birlikte  $V4$  özniteliği de biri hariç diğer dört algoritma tarafından önemli görülmüştür. Varmedja vd. (2019), öznitelik seçimi için Will Koehrsen tarafından

geliştirilen bir aracı kullanmışlardır. %95 kümülatif öneme katkıda bulunmayan öznitelikler silinmiş ve toplam 31 öznitelikten 27 tane kalmıştır. Ek olarak öznitelik seçimi ve veri kümesinin dengelenmesi işlemlerinin başarılı sonuçlara ulaşmak için dikkat edilmesi gereken unsurlar olduğunu paylaşmıştır. Shah (2020), yaptığı çalışmada modelin başarısına bir fayda sağlamayan *Time* özneliğini silmiştir. *Amount* özneliği içinse normalize işlemi uygulamıştır. Sharma (2020), yoğunluk grafikleri (density plot) kullanarak öznitelikler hakkında kaliteli bilgi elde etmiştir. Bu grafikler neticesinde V4 ve V11 özniteliklerinin net bir şekilde ayırık; V12, V14 ve V18 özniteliklerinin kısmen ayırık; V1, V2, V3 ve V10 özniteliklerinin oldukça farklı; V25, V26 ve V28 özniteliklerinin ise iki sınıf için benzer dağılımlara sahip olduğunu paylaşmıştır. Ayrıca RF, AdaBoost, XGB algoritmaları ile önemli öznitelikleri elde etmiş ve 3 metodun ortak olarak bulduğu öznitelikler: V14, V10, V4, V12 ve V17'dir. Rtayli ve Enneya (2020 b), kredi kartı dolandırıcılık tespiti için hibrit bir şekilde destek vektör makinaları ile Özyinelemeli Öznitelik Eleme (Recursive Feature Elimination (RFE)) yöntemlerini kullanmıştır. Ileberi vd. (2022), yaptıkları çalışmada öznitelik seçimi için genetik algoritma kullanmışlardır. Genetik algoritma uygunluk metodu olarak RF tercih edilmiştir. Çalışma bu algoritma ile 5 farklı öznitelik vektörü elde etmiş ve testlerini gerçekleştirmiştir. Sonuçlara bakıldığında aralarındaki ve önceki çalışmalara kıyasla en yüksek doğruluk değerine 5. öznitelik vektörü ile ulaşmıştır. Bu vektörde bulunan öznitelikler ise: *Time*, V1, V7, V8, V9, V11, V12, V14, V15, V22, V27, V28 ve *Amount* öznitelikleridir.

Literatür incelemesi kapsamında çoğunlukla öznitelik seçimi için öznitelik önemlilik metodundan yararlanılmıştır. Bu çalışmada ise öznitelik seçimi için Pearson Korelasyon, RF, Lasso Regresyon ve RFE metotlarından yararlanılmaktadır. Bununla birlikte incelenen çalışmalar dahilinde en çok tercih edilen öznitelik ise V14 olarak tespit edilmiştir.

## 2.4. Makine Öğrenmesi ve Değerlendirme Metrikleri

Makine öğrenmesi, yapay zekanın bir alt dalı olan, geçmiş deneyimlerden insan müdahalesi olmadan öğrenme yeteneğine sahip, gelecek çıktıları tahmin edebilen metotlardır. Hangi durumda ne yapacağı geleneksel programlama yönteminde olduğu gibi açıkça bildirilmeden, kendisinin verilerdeki örüntü, karakteristik aracılığıyla öğrendiği yöntemlerdir. Kullanım alanları oldukça geniştir. Örnek olarak:



- Spam filtreleme
- Hava tahmini
- Ev fiyat tahmini
- Dolandırıcılık tespiti gibi birçok alanda kullanılmaktadır.

Kredi kartı dolandırıcılık tespitinde makine öğrenmesi yöntemleri gerçekleştirilen kart ile satın alma işleminin dolandırıcılık olup olmadığını tahmin etmesi üzerinedir. Her bir işlemin doğru tespiti hem kart sahibi hem banka hem de makine öğrenmesi modelinin sürdürülebilirliği açısından oldukça önemlidir. Bu bağlamda bir modelin güvenilirliğinin kanıtlanabilmesi için model test aşamasından geçer. Test aşamasında eğitim aşamasında kullanılmayan, önceden ayrılmış verilerden oluşan bir alt küme ile bu model test edilir. Bu çalışmada olduğu gibi dengesiz bir veri kümesi ile çalışılıyorsa test aşamasında hangi metriğin tercih edileceği dikkate alınması gereken bir unsurdur. Çünkü Doğruluk gibi bazı metrik değerleri modellerin tespit başarısı yeterli olmamasına rağmen yüksek olabilmekte, yanıltabilmektedir. Bu kısmın devamında çalışmada kullanılan makine öğrenmesi metotları, değerlendirme metrikleri ve ilgili literatür çalışmaları verilmiştir.

#### 2.4.1. Makine öğrenmesi metotları ve değerlendirme metrikleri

LR, bağımsız değişkenleri dikkate alarak bağımlı değişkeni tahmin etmek için kullanılan bir sınıflandırma metodudur (Wright (1995)). Bağımlı değişkenler sınıfların etiketleri olabilmekte iken bağımsız değişkenler ise öznelilikleri temsil etmektedir. LR, lineer regresyonun (Denklem 2.3) modifiye halidir. Lineer regresyon Sigmoid fonksiyonu (Denklem 2.4) ile düzenlenerek bu modifikasyon (Denklem 2.5) elde edilmiştir. Böylece fonksiyon çıktı değeri 0 ile 1 arasına dönüştürülerek ikili sınıflandırma işlemine uygun hale getirilmiştir. Denklem 2.5'te  $\beta_0$  değeri eğilim,  $\beta_1$  ise ağırlık katsayılarını,  $X$  girdi değerini ve  $y$  ise tahmin değerini göstermektedir.

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (2.3)$$

$$g(X) = \beta_0 + \beta_1 X \quad (2.4)$$

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \quad (2.5)$$

SVM ilk olarak C. Cortes ve V. Vapnik tarafından sınıflandırma ve regresyon problemlerini çözmek için tanıtıldı (Cortes ve Vapnik (1995)). Bu metot için temel fikir, iki sınıf arasındaki marjı maksimize eden optimal bir hiper düzlem elde etmektir. SVM metodunun özelliklerinden birisi, verileri doğrusal olmayan bir  $\emptyset$  fonksiyonu aracılığıyla daha yüksek boyutlu bir uzaya yansıtarak doğrusal olmayan bir karar sınırı bulabilmesidir. Bu, orijinal girdi uzayında düz bir çizgi ile ayrılamayan veri noktalarının, bir sınıfın veri noktalarını diğerinden ayıran doğrusal bir hiper düzlemin olabileceği bir özellik uzayı  $F$ 'e kaldırıldığı anlamına gelir. Bu hiper düzlem  $I$  girdi uzayına geri yansıtıldığında, doğrusal olmayan bir eğri biçimine sahip olacaktır. SVM, optimizasyon probleminden (Denklem 2.6) Denklem 2.7'ye bağlı olarak formülleştirilmiştir. Burada  $\emptyset$  çekirdek fonksiyonu,  $x_i$  eğitim noktalarını girdi uzayından daha yüksek boyutlu bir özellik uzayına eşler. Düzenleme parametresi  $C$ , eğitim verilerinde düşük bir hata elde etmek ile ağırlıkların normunu en aza indirmek arasındaki dengeyi kontrol eder.

$$\text{Min } \Phi(w) = \frac{1}{2} \omega^T \omega + C \sum_{i=1}^n \xi_i \quad (2.6)$$

$$\begin{aligned} y_i ((\omega, \emptyset(x_i)) + b) &\geq 1 - \xi_i, \quad i=1, \dots, n \\ \xi_i &\geq 0, \quad i=1, \dots, n \end{aligned} \quad (2.7)$$

Hata matrisi kolay anlaşılması ve diğer temel metrikleri (kesinlik, duyarlılık, doğruluk gibi) hesaplamada kullanılması sebebiyle en yaygın değerlendirme metriklerinden birisidir. Modelin genel performansını açıklayan  $N \times N$  ( $N$ , sınıf sayısı) boyutlu bir matristir. Dolayısıyla ikili sınıflandırma problemlerinde  $2 \times 2$  boyutunda hata matrisi kullanılmaktadır (Şekil 2.3). Bu matriste kullanılan terimler şöyle açıklanabilir:

- Doğru-pozitif (True Positive (TP)), asıl sınıfı pozitif (dolandırıcılık) olan bir verinin pozitif olarak tahmin edilmesi durumudur.
- Yanlış-pozitif (False Positive (FP)), asıl sınıfı negatif (normal) olan bir verinin pozitif olarak tahmin edilmesi durumudur.

- Doğru-negatif (True Negative (TN)), asıl sınıfı negatif (normal) olan bir verinin negatif olarak tahmin edilmesi durumudur.
- Yanlış-negatif (False Negative (FN)), asıl sınıfı pozitif (dolandırıcılık) olan bir verinin negatif olarak tahmin edilmesi durumudur.

Gerçek	Pozitif (1)	Doğru-pozitif	Yanlış-negatif
	Negatif (0)	Yanlış-pozitif	Doğru-negatif
		Pozitif (1)	Negatif(0)
		Tahmin	

Şekil 2.3. Hata matrisi

Duyarlılık, bazı yerlerde Hassaslık veya Doğru-pozitif Oranı (True Positive Rate (TPR)) olarak da geçen, Denklem 2.8’de gösterilen doğru pozitifin tüm pozitiflere oranıdır.

$$Duyarlılık = \frac{TP}{TP + FN} \quad (2.8)$$

Kesinlik, Denklem 2.9’da gösterildiği gibi doğru pozitiflerin, pozitif tahmin edilen tüm verilere oranıdır.

$$Kesinlik = \frac{TP}{TP + FP} \quad (2.9)$$

F1-Skor, Kesinlik ve Duyarlılık metriklerinin harmonik ortalamasıdır (Denklem 2.10).

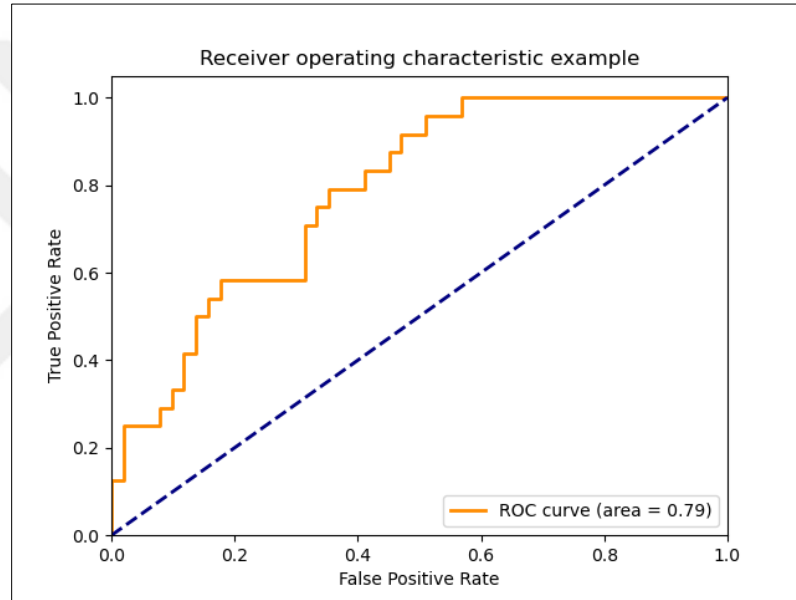
$$F1-Skor = \frac{2 * (Kesinlik * Duyarlılık)}{Kesinlik + Duyarlılık} \quad (2.10)$$

Doğruluk doğru tahmin edilen pozitif ve negatif sınıf verilerinin tüm verilere oranıdır (Denklem 2.11).

$$Doğruluk = \frac{TP + TN}{TP + TN + FN + FP} \quad (2.11)$$

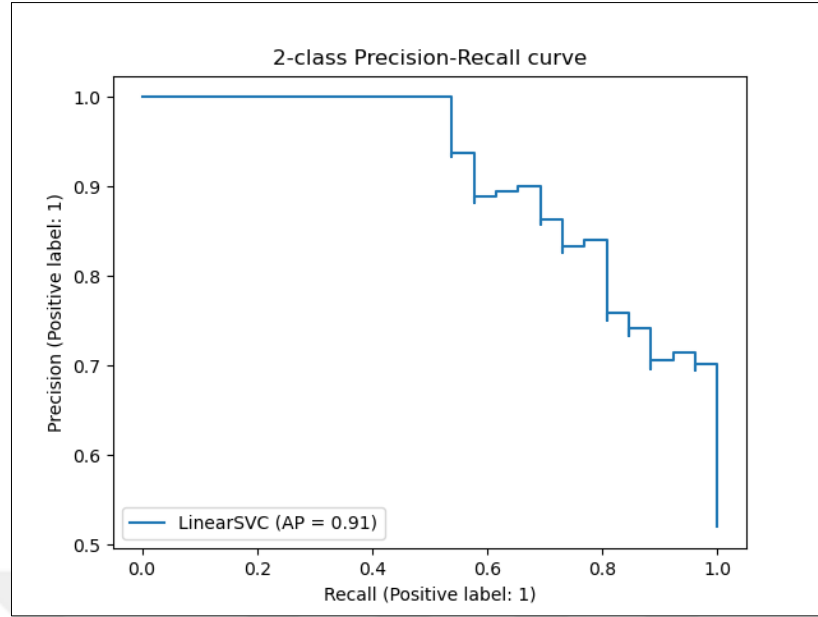
AUROC tahmin analizinde en yaygın kullanılan değerlendirme metriklerinden biridir. Modelin farklı olasılık eşik değerlerinde kullanımında elde edilen sonucu yansıtmaktadır. X ekseninde özgüllük (Denklem 2.12), y ekseninde duyarlılık ile (Bkz. Denklem 2.8) oluşan bir grafikdir (Şekil 2.4). Burada siyah kesikli çizgi ortalama rastgele bir modelin elde edebileceği yani 1 üzerinden 0,5 değerini, mavi çizgi ise 0,79 değerini göstermektedir.

$$\text{Özgüllük} = \frac{TN}{TN + FP} \quad (2.12)$$



Şekil 2.4. ROC grafiği (Buitinck vd.'den, 2013)

AUPRC, ROC eğrisinin düzenlenmiş halidir (Şekil 2.5). Dengesizlik problemine sahip veri kümelerinde her iki sınıf dikkate alındığında yanlış bir algı yaratıldığı için sadece azınlık olan sınıfı (pozitif, dolandırıcılık) kapsayan metrikler tercih edilmiş ve özgüllük yerine burada kesinlik kullanılmaktadır.



Şekil 2.5. PR grafiği (Buitinck vd.’den, 2013)

#### 2.4.2. Literatür incelemesi

Bu kısımda hangi metotların ne sıklıkta tercih edildiğini görebilmek için European Cardholders veri kümesi ile 2018 ve sonrası yıllarda yapılan 63 çalışma, makine öğrenmesi ve değerlendirme metriklerine göre detaylıca Çizelge 2.7 ve Çizelge 2.8’de incelenmektedir. Çizelge 2.7’de ele alınan makine öğrenmesi metotları şunlardır: RF, LR, SVM, KNN, DT, Naïve Bayes (NB), XGB, Sinir Ağları (Neural Network (NN)), Çok Katmanlı Algılayıcılar (Multilayer Perceptron (MLP)), Oto Kodlayıcı (Auto Encoder (AE)), AdaBoost, Bagging, GB, LightGBM. Çizelge 2.8’de ele alınan değerlendirme metrikleri ise şunlardır: Duyarlılık (Recall), Kesinlik (Precision), Doğruluk (Accuracy), F1-Skor (F1-Score (F1)), Alıcı İşlem Karakteristik Eğrisi Altındaki Alan (Area Under Receiver Operating Characteristics Curve (AUROC)), Özgüllük (Specificity), Kesinlik-Duyarlılık Eğrisi Altındaki Alan (Area Under Precision-Recall Curve (AUPRC)), Matthews Korelasyon Katsayısı (Matthews Correlation Coefficient (MCC)), Yanlış Pozitif Oranı (False Positive Rate (FPR)), Geometrik Ortalama (Geometric Mean (G-Mean)), Yanlış Sınıflandırma Oranı (Misclassification Rate (MR)), Finansal Kurtarma (Financial Recovery (FR)). Bununla birlikte çalışmada kullanılan öğrenme metotları iki bölümde anlatılmıştır. LR ve SVM bu bölümün devamında iken, topluluk öğrenmesi metotları RF, LGBM ve XGB metotları Bölüm 2.6’da verilmiştir. Çalışmada kullanılan değerlendirme metrikleri ise bu makine öğrenmesi metotlarının devamında açıklanmıştır: Hata Matrisi (Confusion Matrix), Duyarlılık, Kesinlik, F1-Skoru,



Çizelge 2.7. European Cardholders veri kümesi ile yapılan çalışmalarda kullanılan makine öğrenmesi metotları (devam)

Makaleler	Metotlar												
	RF	LR	SVM	KNN	DT	NB	XGB	NN	MLP	AE	AdaBoost	GB	LightGBM
Alanezi vd., 2020		✓			✓	✓							
Ata ve Hazım, 2020	✓		✓	✓		✓							
Aung vd., 2020	✓	✓	✓	✓		✓		✓					
Babu ve Pratap, 2020								✓					
Bej vd., 2020	✓	✓		✓									
El Hajjami vd., 2020	✓	✓	✓		✓	✓							
Gulati, 2020		✓											
Husejinović, 2020						✓							
Ito vd., 2020		✓		✓		✓							
Janbandhu vd., 2020	✓		✓	✓									
Khatri vd., 2020	✓	✓		✓	✓	✓							
Kittidachanan vd., 2020						✓							
Lin ve Jiang, 2020	✓									✓			
Mînaştireanu ve Meşniță, 2020	✓				✓	✓							
Mrozek vd., 2020	✓	✓		✓									
Muter ve Molood, 2020			✓										
Nguyen vd., 2020	✓		✓										
Novakovic ve Markovic, 2020						✓				✓			
Riffi vd., 2020									✓				
Rtayli ve Enneya, 2020 a	✓		✓		✓								
Rtayli ve Enneya, 2020 b										✓			
Saheed vd., 2020	✓		✓			✓							
Shah, 2020	✓	✓		✓		✓							
Shamsudin vd., 2020	✓			✓		✓							

Çizelge 2.7. European Cardholders veri kümesi ile yapılan çalışmalarda kullanılan makine öğrenmesi metotları (devam)

Makaleler	Metotlar												
	RF	LR	SVM	KNN	DT	NB	XGB	NN	MLP	AE	AdaBoost	GB	LightGBM
Sharma, 2020	✓						✓				✓		✓
Shivanna vd., 2020									✓				
Tingfei vd., 2020										✓			
Trivedi vd., 2020	✓	✓	✓	✓	✓	✓							
Wang vd., 2020	✓	✓	✓		✓	✓		✓					
Zhang vd., 2020	✓	✓	✓										
bin Alias vd., 2021	✓	✓					✓						
Cynthia ve George, 2021		✓	✓										
Isabella vd., 2021				✓							✓	✓	
Tran ve Dang, 2021	✓	✓		✓	✓								
Wibowo ve Fatichah, 2021										✓			
Alfaiz ve Fati, 2022	✓	✓		✓	✓		✓					✓	✓
Budianto vd., 2022		✓											
Ileberi vd., 2022	✓	✓			✓	✓		✓					
Liang vd., 2022	✓	✓	✓	✓	✓				✓		✓	✓	
Mathew vd., 2022	✓	✓		✓	✓								
<b>TOPLAM</b>	<b>37</b>	<b>33</b>	<b>23</b>	<b>21</b>	<b>21</b>	<b>20</b>	<b>8</b>	<b>7</b>	<b>7</b>	<b>6</b>	<b>4</b>	<b>4</b>	<b>2</b>





Çizelge 2.8. European Cardholders veri kümesi ile yapılan çalışmalarda kullanılan değerlendirme metrikleri (devam)

Makaleler	Duyarlılık	Kesinlik	Doğruluk	F1-Skor	AUROC	Özgüllük	AUPRC	MCC	FPR	G-Mean	MR	FR
Aung vd., 2020		✓	✓									
Babu ve Pratap, 2020			✓									
Bej vd., 2020			✓	✓								
El Hajjami vd., 2020			✓		✓							
Gulati, 2020	✓	✓	✓	✓	✓							
Husejinović, 2020	✓	✓					✓					
Ito vd., 2020	✓	✓	✓		✓	✓						
Janbandhu vd., 2020	✓	✓										
Khatri vd., 2020	✓	✓										
Kittidachanan vd., 2020	✓			✓	✓	✓						
Lin ve Jiang, 2020	✓		✓			✓		✓				
Mînaştireanu ve Meşniță, 2020	✓	✓			✓							
Mrozek vd., 2020				✓	✓							
Muter ve Molood, 2020			✓									
Nguyen vd., 2020	✓	✓	✓	✓								
Novakovic ve Markovic, 2020	✓	✓		✓								
Riffi vd., 2020	✓	✓	✓						✓			
Rtayli ve Enneya, 2020 a			✓		✓							
Rtayli ve Enneya, 2020 b	✓	✓	✓	✓	✓	✓						
Saheed vd., 2020	✓	✓	✓			✓						
Shah, 2020	✓	✓	✓	✓								
Shamsudin vd., 2020	✓	✓		✓								
Sharma, 2020					✓							
Shivanna vd., 2020	✓	✓	✓	✓	✓							
Tingfei vd., 2020	✓	✓	✓	✓		✓						

Çizelge 2.8. European Cardholders veri kümesi ile yapılan çalışmalarda kullanılan değerlendirme metrikleri (devam)

Makaleler	Duyarluluk	Kesinlik	Doğruluk	F1-Skor	AUROC	Özgüllük	AUPRC	MCC	FPR	G-Mean	MR	FR
Trivedi vd., 2020	✓	✓	✓	✓					✓			
Wang vd., 2020	✓	✓		✓								
Zhang vd., 2020												✓
bin Alias vd., 2021	✓	✓	✓					✓				
Cynthia ve George, 2021	✓	✓	✓	✓								
Isabella vd., 2021			✓									
Tran ve Dang, 2021	✓	✓	✓			✓						
Wibowo ve Fatichah, 2021	✓	✓	✓	✓	✓		✓					
Alfaiz ve Fati, 2022	✓	✓	✓	✓	✓							
Budianto vd., 2022	✓	✓	✓	✓	✓							
Ileberi vd., 2022	✓	✓	✓	✓								
Liang vd., 2022	✓	✓	✓	✓	✓							
Mathew vd., 2022	✓	✓	✓	✓								
<b>TOPLAM</b>	<b>43</b>	<b>42</b>	<b>42</b>	<b>26</b>	<b>26</b>	<b>14</b>	<b>6</b>	<b>4</b>	<b>4</b>	<b>2</b>	<b>1</b>	<b>1</b>

Bu incelemeler neticesinde öğrenme yöntemleri olarak yürütülen bu çalışmada RF, LGBM, XGB, LR ve SVM metotları kullanılmıştır. İlk üç metot topluluk öğrenmesi ve diğer iki metot tekli öğrenici amacıyla tercih edilmiştir. Karar ağacına dayanan artırma metotları birçok uygulamada ayırt edici konumdadır (Friedman, 2001; Li, 2012; Sharma, 2020; bin Alias vd., 2021; Alfaiz ve Fati, 2022). Yine literatür çalışmaları aracılığıyla en yaygın kullanılan Duyarlılık, Kesinlik, Doğruluk, F1-Skor, AUROC ve AUPRC değerlendirme metrikleri bu çalışmada kullanılmıştır.

## 2.5. Topluluk Öğrenmesi

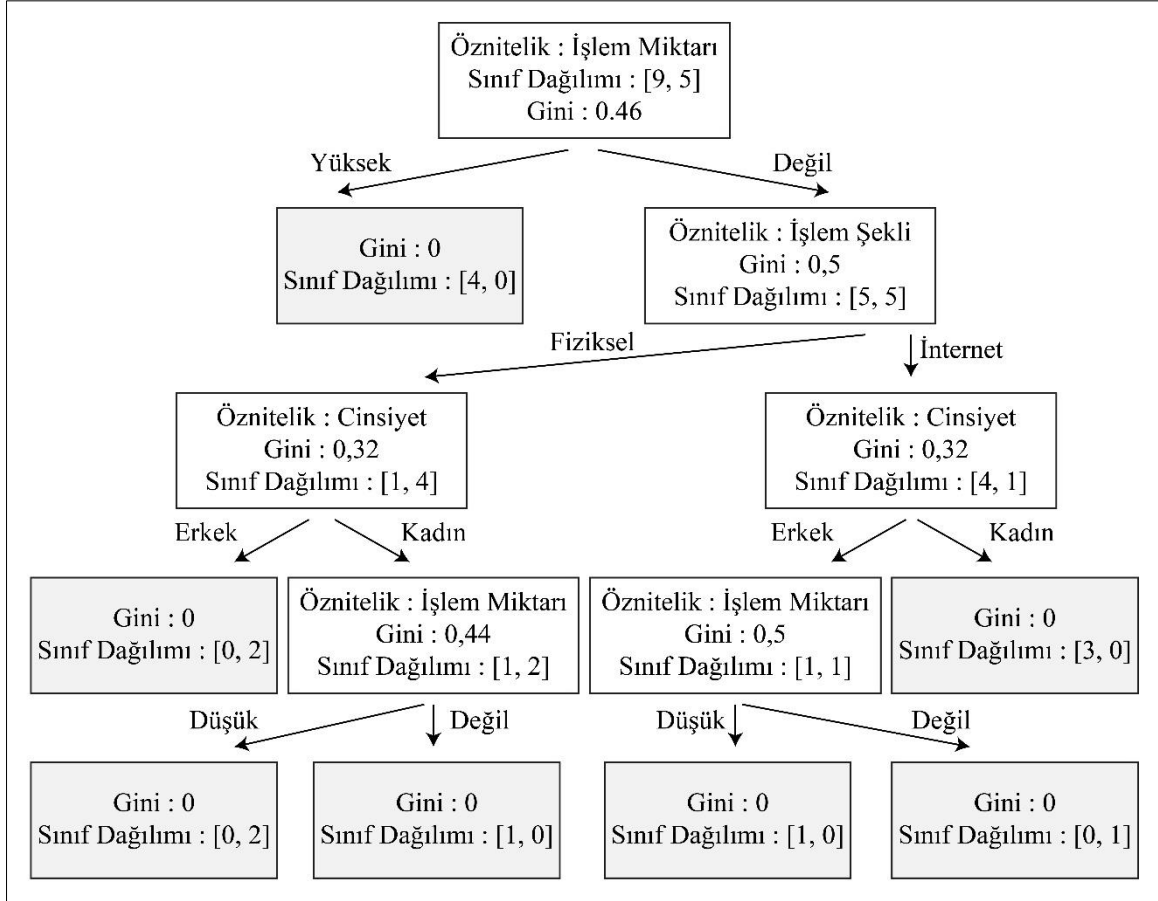
Bu kısımda ilk olarak çalışmada kullanılan topluluk öğrenmesi yöntemleri verilmiştir. Bu yöntemlerin devamında ise ilgili literatür çalışmaları incelenmiştir.

### 2.5.1. Topluluk öğrenmesi yöntemleri

Topluluk öğrenmesi temelde bir zayıf öğrencinin farklı şekillerde eğitilerek her birinin nihai tahminini çeşitli yöntemlerle birleştirilerek güçlü bir model oluşturma yöntemine dayanmaktadır. Zayıf öğrenci modelinin ne olması gerektiği önemli bir durumdur. Homojen ve heterojen olarak türleri vardır. Eğer zayıf öğrenciler sadece bir model üzerinden türetiliyorsa, bu model homojen direkt birden çok farklı model kullanılıyorsa bu durumda heterojen olmaktadır. Homojen zayıf öğrenci tabanlı topluluk öğrenmesi yöntemleri torbalama ve artırma iken heterojen zayıf öğrenci tabanlı ise yığma yöntemidir. Bu yöntemler şu şekillerde çalışırlar: Torbalama yönteminde homojen zayıf öğrenciler birbirinden bağımsız paralel bir şekilde öğrenme işlemlerini gerçekleştirir ve nihai olarak bazı deterministik ortalama işlemleri kullanılarak bu öğrenciler birleştirilir; Artırma yönteminde homojen zayıf öğrenciler ardışık ve bir önceki bağlı şekilde öğrenme işlemlerini gerçekleştirir; Yığma yönteminde ise heterojen olan zayıf öğrenciler paralel bir şekilde öğrenme işlemini gerçekleştirir. Aşağıda bu çalışmada kullanılan topluluk öğrenmesi metotları ve dayandıkları zayıf öğrenci metotları verilmiştir.

İlk olarak DT metodu bir topluluk öğrenmesi metodu değildir; ancak birçok topluluk öğrenmesi metodunun zayıf öğrencisi olarak seçildiği için bu başlık altında açıklanmıştır. Karar ağacı, regresyon ve sınıflandırma problemleri için kullanılan ağaç yapısına sahip bir denetimli öğrenme metodudur. Bir ağaç yapısında öncelikle bir kök düğüm gereklidir ve bu kök düğüm dallara ayrılarak nihai yaprak düğümlerinde sadece bir sınıfa ait veriler kalana kadar bu işlem devam etmektedir. Bu metot bir düğümde dallanma işlemi için gerekli olan karar kuralı elde edebilmek amacıyla her özneliği ve bu özneliklere ait verileri detaylıca analiz etmektedir. Şekil 2.6’de kök düğümde dallanma için gerekli öznelik, Çizelge 2.9’da paylaşılan veri kümesi dikkate alınarak işlem miktarı karar olarak ise yüksek olup olmaması belirlenmiştir. Çizelge 2.10’da görüldüğü üzere kök düğüm için toplamda 5 farklı karar kuralı vardır ve bu kurallar arasında en fazla bilgi kazanımı veren karar kuralı seçilmiştir. Burada belirtilen kazanımları hesaplamak için *Gini* (Denklem 2.7) ve *Entropi* (Denklem 2.8) en yaygın kullanılan yöntemler iken bu örnekte *Gini* kullanılmıştır. Öznelik ve karar kuralı seçimi ile dallanma sağlandıktan sonra tüm veriler buna göre sınıflandırılır, işlem şekli özneliği yüksek olanlar sol düğüm altında, yüksek olmayanlar sağ düğüm altında toplanır. Görüldüğü üzere sol düğüm altında sadece bir sınıfa ait veriler bulunduğu için burada

tamamen saf durum meydana gelmiş ( $Gini=0$ ) ve burada artık bir ayırım yapmaya ihtiyaç yoktur. İşlem miktarı yüksek olmayanların ayrıldığı düğümde ise her iki sınıftan eşit miktarda veri olduğu için safsızlık maksimum düzeyde ( $Gini=0,5$ ) ve dallanma işlemi gereklidir. Kalan veriler arasından tekrar en fazla kazanımı verecek öznelilik ve karar kuralı belirlenip saf düğümler elde edilerek tüm ağaç yapısı bu şekilde oluşturulur.



Şekil 2.6. Karar ağacı örneği

Çizelge 2.9. Kredi kartı dolandırıcılık veri kümesi

İşlem Miktarı	İşlem Şekli	Cinsiyet	Dolandırıcılık
Düşük	Fiziksel	Kadın	Hayır
Düşük	Fiziksel	Erkek	Hayır
Yüksek	Fiziksel	Kadın	Evet
Orta	Fiziksel	Kadın	Evet
Orta	İnternet	Kadın	Evet

Çizelge 2.9. Kredi kartı dolandırıcılık veri kümesi (devam)

İşlem Miktarı	İşlem Şekli	Cinsiyet	Dolandırıcılık
Orta	İnternet	Erkek	Hayır
Yüksek	İnternet	Erkek	Evet
Düşük	Fiziksel	Kadın	Hayır
Düşük	İnternet	Kadın	Evet
Orta	İnternet	Kadın	Evet
Düşük	İnternet	Erkek	Evet
Yüksek	Fiziksel	Erkek	Evet
Yüksek	İnternet	Kadın	Evet
Orta	Fiziksel	Erkek	Hayır

Çizelge 2.10. Kök düğüm için karar kuralları kümesi

Kök Düğüm için Olası Karar Kuralları Kümesi		
Öznitelik	Kural	Bilgi Kazanım
İşlem Miktarı	Düşük mü?	0,06
İşlem Miktarı	Yüksek mi?	0,14
İşlem Miktarı	Orta mı?	0,04
İşlem Şekli	Fiziksel/İnternet?	0,10
Cinsiyet	Erkek/Kadın?	0,03

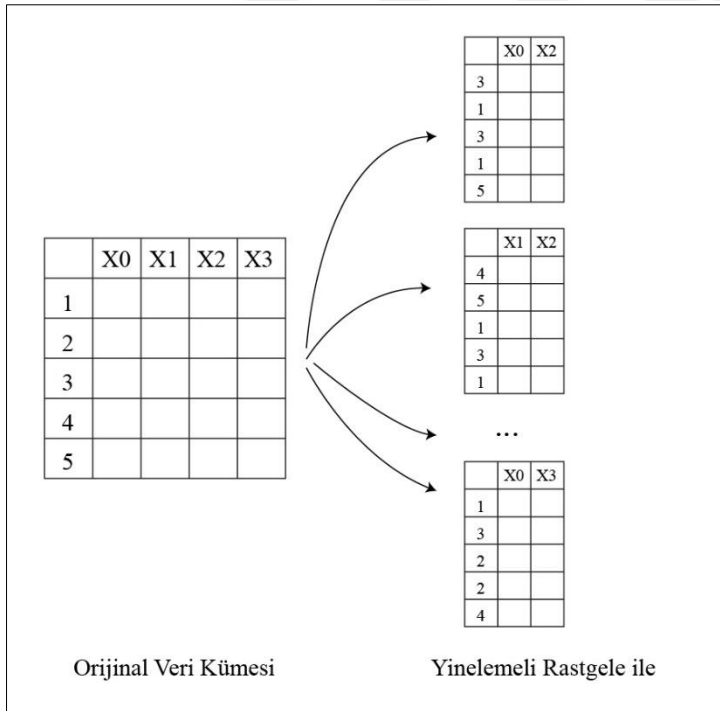
$$Gini = 1 - \sum_{i=1}^n p^2(c_i) \quad (2.7)$$

$$Entropi = \sum_{i=1}^n -p(c_i) \log_2(p(c_i)) \quad (2.8)$$

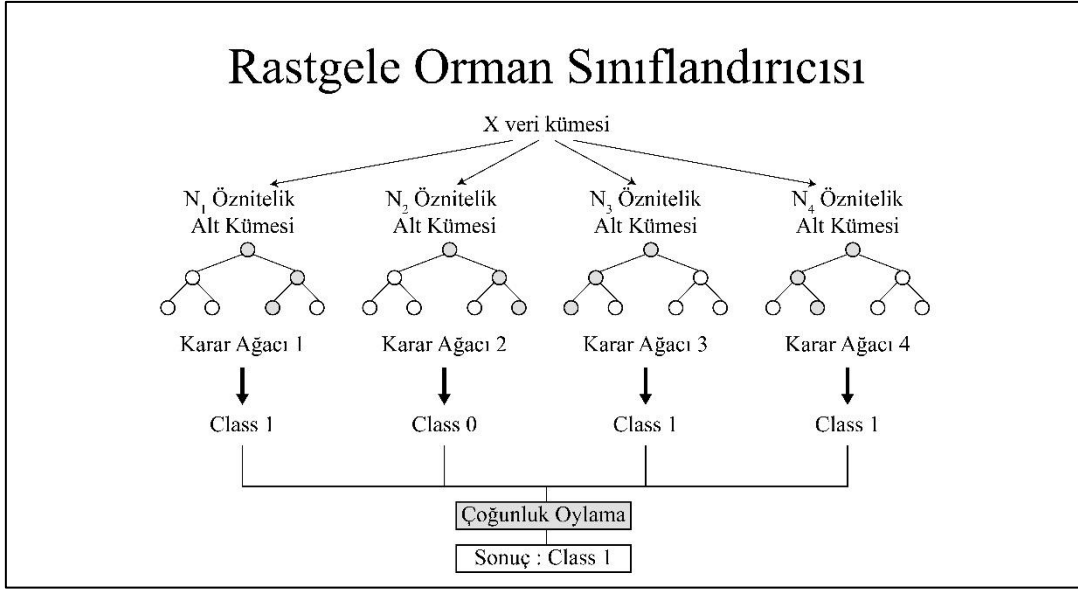
DT'ler geçerli düğümde en iyi ayırım noktasını seçtiği için daha önceki ayırımlardan sonra gelişebilecek daha iyi ayırımları fark edemeyen ağgözlü bir yaklaşıma sahiptir. Bu sebeple yüksek varyansa sahip olma eğilimindedir ve eğitim veri kümesinde iyi ancak test veri kümesinde kötü performans gösterebilme (aşırı uyum) olasılığı vardır. Bu problem nedeniyle oluşturulan ağaç yapısında budama işlemine gidilebilmektedir. Budama işlemi ile halen dallara ayrılan bir düğüm nihai yaprak düğüme dönüştürülür. Böylece DT keskin

kararlarından ziyade daha genel kararlara sahip olarak aşırı uyum problemi nispeten çözülmesi hedeflenmiştir.

RF, Şekil 2.7’de gösterildiği üzere, orijinal veri kümesi ile aynı dağılıma sahip veri alt kümelerini, orijinal veri kümesinden rastgele öznitelik seçimi ve yinelemeli seçim yaparak oluşturan bir torbalama metodudur (Breiman (2001)). Yukarıda anlatıldığı üzere her torbalama yönteminde birden çok homojen zayıf öğrenciler vardır. Bu metot kapsamında zayıf öğrenciler için DT metodu kullanılmaktadır. Probleme göre DT sayısı değişiklik gösterebilmektedir. Her bir DT metodu için gerekli olan veri kümesi ise yukarıda belirtilen alt veri kümelerinden seçilmektedir. Bu ağaçlar birbirinden bağımsız eğitilir ve en çok tahmin edilen sınıf ana modelin tahmin ettiği sınıf olarak seçilir (Şekil 2.8). Bu eklenmiş rastgelelik ile DT metodunun sahip olduğu yüksek varyans veya verideki değişime bağlı hassaslık ve dolayısıyla aşırı uyum probleminin çözülmesi hedeflenmiştir. Dolayısıyla her bir DT’de ayrıca bir budama işlemi yapılmamaktadır.



Şekil 2.7. Yinelemeli rastgele ile veri kümelerinin oluşturulması



Şekil 2.8. RF metodu çoğunluk oylama yapısı

GB bir önceki modelin hatasını azaltmaya çalışan ardışık modeller geliştirmeye dayanan bir yöntemdir (Friedman (2001)). Özellikle büyük ve karmaşık veri kümelerinde tahmin hızı ve doğruluğu ile öne çıkan bir metottur. Makine öğrenmesi kapsamında sapma varyans dengesi (bias-variance tradeoff) önemli bir durumdur. Bu bağlamda GB metodunda amaç modelin sapma hatasını en aza indirmektir. Metodun çalışması bağlamında öncelikle tüm verilere sabit bir başlangıç tahmin değeri atanır (Denklem 2.13). Bu değer başlangıç yaprağı olarak adlandırılır. Bu değeri sınıflandırma probleminde kullanabilmek için olasılık değerine dönüştürülür (Denklem 2.14). Bu tahmin değeri ile hedef değer arasındaki hataya (Denklem 2.15) göre yaprak sayısı genelde sekiz ile otuz iki arasında olan basit bir DT oluşturulur. Yaratılan bu DT'nin her bir yaprağı ilgili veri ya da verilerin hata değerlerine sahiptir. Tahmin değerinin güncellenmesi için hata değerinin tahmin değeri ile işleme girebilmesi gerekmektedir ancak hata değerleri olasılıksal bir değer olduğu için dönüşüm ihtiyacı vardır. En yaygın kullanılan dönüşüm formülü Denklem 2.16'da gösterilmiştir. Her bir yaprak için dönüşüm uygulanır. Elde edilen dönüştürülmüş değeri ile güncel tahmin değerini elde edebilmek için Denklem 2.17'den yararlanılır. Tahmin değeri sınıflandırma problemi kapsamında 0, 1 arasında olması gerektiği için Denklem 2.14'ten geçer nihai tahmin değeri elde edilir. Tekrar hata değeri hesaplanır bu hata değerine göre yeni bir DT oluşturulur. Bu şekilde belirlenen ağaç sayısına ya da elde edilen hata miktarı eşik değerinin altında kalana kadar işlemler yinelenmeli şekilde devam eder. Çizelge 2.11'de GB metodu kapsamında kullanılan formüller açıklamaları ile paylaşılmıştır.



$$\log(odds) = \log\left(\frac{c_{Pozitif}}{c_{Negatif}}\right) \quad (2.13)$$

$$p = \frac{e^{\log(odds)}}{1 + e^{\log(odds)}} \quad (2.14)$$

$$e = o - p \quad (2.15)$$

$$t = \frac{\sum e}{\sum [p_{\text{önceki}} \times (1 - p_{\text{önceki}})]} \quad (2.16)$$

$$p_{\text{yeni}} = i + l \times t \quad (2.17)$$

Yukarıdaki beş denklemde (Bkz. Denklem 2.13 – 2.17) bulunan parametreler şöyledir:

$c_{Pozitif}$ : pozitif sınıftaki verilerin sayısı

$c_{Negatif}$ : negatif sınıftaki verilerin sayısı

$odds$ : bir durumun gerçekleşmesinin gerçekleşmemesine oranıdır.

$o$ : hedef değişken

$e$ : hata

$t$ : dönüşüm

$p_{\text{yeni}}$ : güncel tahmin değeri

GB metodunda DT oluşturabilmek için olası tüm veri ve öznitelikleri dikkate alması gerekmektedir. Bu durum özellikle büyük veri ile çalışırken oldukça zaman tüketen bir duruma dönüşmektedir. LightGBM metodu iki farklı teknikle çalışarak neredeyse normal bir gradyan artırma metodunun elde ettiği aynı başarı seviyesine 20 kat daha hızlı ulaşabilmektedir (Ke vd. (2017)). Bu tekniklerden ilki gradyan tabanlı tek taraflı örneklemedir. Bu teknik tüm örneklemler yerine daha yüksek gradyan veren örneklemler üzerinde çalışarak aynı başarı seviyesi ve hızlı işlem yapabilmeyi hedefler. Diğer teknik ise özel öznitelik paketlemedir. Bu teknik seyrek öznitelikleri birleştirmeyi hedefleyerek daha az öznitelikle çalışmaktadır. Bu sebeple işlem süresinin dahada azalması beklenmektedir.

XGB metodu çeşitli sistem optimizasyonları ve algoritmik geliştirmeler ile popüler olan algoritmalarından 10 kat daha hızlı çalışabilen bir gradyan artırma türevi metottur (Chen ve Guestrin (2016)). Bu optimizasyonlar ve algoritmik geliştirmeler şunlardır:

- Parallelleştirme ile DT oluşturulmasında tüm işlemci çekirdekleri kullanılır.
- Dağıtık hesaplama ile devasa veri kümelerinin eğitimi için farklı makineler kullanılır.
- Çekirdek dışı hesaplama ile bilgisayar hafızasına sığmayan büyük veri kümeleri işlenebilir.
- Önbellek optimizasyonu ile donanımın en verimli kullanımını sağlar.
- Seyreklik farkındalığı ile kayıp verileri otomatik ele alır.
- Blok yapısı paralel ağaç oluşturmaya destekler.
- Sürekli eğitim ile hali hazırda eğitilmiş bir modelin yeni veriler üzerinde eğitimine devam etmesine imkân verir.

XGB, gradyan artırma optimizasyon adımının ortalama kare hatasını azaltmaya çalıştığı ve ikili sınıflandırma için standart log kaybının kullanıldığı regresyon ağaçlarına dayanır. Çok sınıflı bir sınıflandırma problemi için amaç fonksiyonu çapraz entropi kaybını optimize etmektir. Birleştirilen kayıp fonksiyonunu ile düzenleme terimi amaç fonksiyonuna ulaşır. Düzenleme terimi karmaşıklığı kontrol eder ve aşırı uyum riskini azaltır. XGB,  $n$  boyutlu bir düzlemde model bulunmaya çalışırken negatif gradyanı takip ederek her optimizasyon adımında tahminsel doğruluğu iyileştirmek için optimizasyon için gradyan inişini kullanır. Modelde kullanılan fonksiyonlar kümesini öğrenmek için XGB, Denklem 2.18'i en aza indirir. Denklem 2.1'de  $\Theta$  öğrenilmiş parametre kümesidir;  $l$  hedef ve tahmin değişkeni arasındaki farkı ölçebilen türevlenebilir dışbükey kayıp fonksiyonudur;  $\Omega$  ise düzenleme terimidir.

$$L(\Theta) = \sum_i l(y_i, \hat{y}_i) + \Omega(\Theta) \quad (2.18)$$

### 2.5.2. Literatür incelemesi

Dhankhad vd. (2018), yaptıkları çalışmada topluluk öğrenmesi olarak GB, XGB, RF ve yığma sınıflandırıcısı ve diğer makine öğrenmesi metotlarından bazılarını kullanmışlardır. F1-Skor metriği dikkate alındığında %95 ile aralarındaki en yüksek değere

RF, XGB ve yığma sınıflandırıcısı ile ulaşmışlardır. Varmedja vd. (2019), kredi kartı dolandırıcılık tespiti için yeniden örnekleme, öznitelik seçim aracı ve çeşitli makine öğrenmesi yöntemlerini kullanmışlardır. Kesinlik metriğinde LR, NB, RF ve MLP metotları sırasıyla %59, %16, %96 ve %79, Duyarlılık metriğinde ise %92, %83, %82 ve %82 değerlerini almışlardır. Rtayli ve Enneya (2020 a), öznitelik seçimi için RF tabanlı SVM metodu önermektedir. Duyarlılık metriğinde RF tabanlı SVM, izolasyon ormanı, yerel aykırı faktör ve DT metotları sırasıyla %87, %34, %5, %0 ve AUROC metriğinde ise %91, %67, %52 ve %50 değerlerini elde etmişlerdir. Isabella vd. (2021), öznitelik seçimi için CART metodu ile birlikte çalışmada BCE-GBHEC (Binary Cross Entropy – Gradient Boost Hybrid Ensemble Classifier) metodu önerilmiştir. Doğruluk metriği dikkate alındığında, AdaBoost+SVM, AdaBoost, KNN, BCE-GBHEC metotları sırasıyla %96, %97, %97 ve %97 değerlerini elde etmişlerdir. Ileberi vd. (2022), öznitelik seçimi için genetik algoritmaya dayanan öznitelik seçimi ile RF, DT, ANN, NB ve LR öğrenme metotlarını önermiştir. Çalışma genetik algoritma ile en uygun 5 öznitelik vektörünü yaratmıştır. RF ile AUROC metriğinde %96 ve F1-Skor metriğinde ise %82 ile çalışmanın en yüksek değerine ulaşmıştır.

İncelenen çalışmalar ile kredi kartı dolandırıcılık tespiti amacıyla kullanılan topluluk öğrenmesi yöntemlerinin başarılı sonuçlar verdikleri görülmüştür. Bu bağlamda özellikle GB yöntemlerinin de son yıllarda yaygınlığı artmıştır (Bkz. Çizelge 2.7). Yürütülen bu çalışmada ise topluluk öğrenmesi metotları olarak RF, LGBM ve XGB metotları kullanılmıştır.

### 3. MATERYAL VE YÖNTEM

Bu tez çalışmasında ele alınan kredi kartı dolandırıcılık tespiti topluluk öğrenmesi yöntemleri ile denenmesi için European Cardholders veri kümesi ve dolayısıyla gerekli önışlemler için örnekleme, öznitelik seçimi gibi bazı yöntemler kullanılmaktadır. Bu başlık altında öncelikle veri kümesi sonrasında çalışmada kullanılan bu yöntemler açıklanmaktadır.

#### 3.1. Materyal

Bu kısımda çalışmada kullanılan yazılım dili, kullanılan bilgisayar ve European Cardholders veri kümesi ile ilgili açıklamalar verilmektedir.

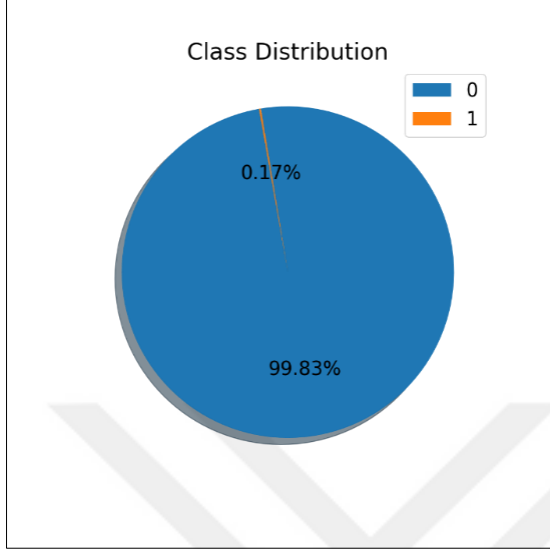
##### 3.1.1. Test ortamı

Bu çalışma kapsamında kullanılan öznitelik seçimi ve örnekleme metotları, eğitim ve test süreci için makine öğrenmesi ve değerlendirme metrikleri için Python yazılım dili kullanılmıştır. Bununla birlikte çalışmanın gerçekleştirildiği bilgisayarın marka, model, işlemci, ram, disk ve ekran kartı bilgileri şöyledir: Monster Tulpar T7 V20.4, Intel Core i7 11800H, 16GB RAM, 500GB M.2 SSD, RTX3060 6GB.

##### 3.1.2. European Cardholders veri kümesi

Bu çalışmada kullanılan European Cardholders veri kümesi, 2013 yılı Eylül ayında 2 gün boyunca Avrupalı kullanıcıların yaptığı transfer işlemlerinden oluşmaktadır (Kaggle, 2013). Veri kümesindeki toplam 284.807 adet verinin sadece 492 tanesi sahtecilik etiketlidir (Şekil 3.1). Dolayısıyla veri kümesi dengesiz ve dengesizlik oranı %0,17'dir. European Carholders veri kümesi toplamda 31 özniteliğe sahiptir. Bu özniteliklerden bazıları hassas finansal bilgiler olduğu için verinin anonim halde tutulması önemlidir (Varmedja vd., 2019). Bu sebeple PCA dönüşüm tekniği uygulanmıştır. Bu özniteliklerden sadece 3 tanesi açık bir şekilde paylaşılmıştır: *Time*, *Amount*, *Class*. *Time* özniteliği veri kümesindeki ilk kredi kartı işlemi ile ilgili kredi kartı işlem arasındaki süreyi göstermektedir. *Amount* özniteliği ise kredi

kartı ile yapılan işlem tutarını göstermektedir. *Class* özneliği de işlemin etiketini gösteren 0 ve 1 değerlerini alır. Kalan 28 tane öznelik isimleri *V1*, *V2*, ..., *V28* şeklinde kodlanmıştır.



Şekil 3.1. Sınıf dağılımı

Çizelge 3.1. European Cardholders veri kümesi öznelikleri

Öznelik	Türü	Açıklama
Time	Tam sayı	İlk işlem ile ilgili işlem arasındaki süre (saniye)
Amount	Reel sayı	İlgili işlem miktarı
normAmount	Reel sayı	Amount özneliğinin normalize edilmiş hali
Class	Tam sayı	İlgili işlemin sınıfı (0 : Normal, 1: Dolandırıcılık )
V1	Reel sayı	Birinci Temel Bileşen
V2	Reel sayı	İkinci Temel Bileşen
...	...	...
V28	Reel sayı	Son Temel Bileşen

### 3.2. Yöntem

Bu kısımda çalışmada kullanılan yeniden örnekleme, öznelik seçimi, makine ve topluluk öğrenmesi metotları ve değerlendirme metrikleri ile ilgili bilgiler verilmektedir.

### 3.2.1. Yeniden örnekleme

Önceden gösterildiği üzere veri kümesi dengesizdir (Bkz. Şekil 3.1). Ağırlıklı olarak normal etiketli verilerden oluştuğu için, eğitilen model normal veriler üzerine eğilimli olmakta ve dolandırıcılık verilerinde başarısız kalmaktadır. Bu problemi çözebilmek için yeniden örnekleme metotları kullanılmaktadır. Bahsedildiği üzere örnekleme metotları aşırı ve az örnekleme şeklinde ikiye ayrılmaktadır (Bkz. Bölüm 2.3.2). Aşırı örnekleme veri kümesindeki ilgili sınıf verilerinin yetersiz olması durumunda veri miktarını çoğaltabilmek için kullanılmaktadır. Az örnekleme ise veri kümesini istenilen denge seviyesine indirebilmek amacıyla kullanılmaktadır. Yüksek oranlarda kullanıldığında verilerin sahip olduğu önemli bilgilerin kaybına neden olabilmektedir.

Yürütülen bu çalışmada aşırı örnekleme olarak SMOTE ve az örnekleme olarak RUS metodu kullanılmıştır. Bu metotlar farklı oranlarda test edilerek en yüksek başarı seviyesine ulaşılmaya çalışılmıştır. Kullanılan oranlar Çizelge 3.2’de verilmiştir. Burada her bir aşırı örnekleme yöntem ve oranı her bir az örnekleme yöntem ve oranıyla eşleştirilerek 6 farklı durum içinde 6 farklı eşleştirme oluşturulmuş ve eğitim veri kümesine uygulanmıştır.

Aşırı örnekleme kapsamında Çizelge 3.2’deki oranlar seçilirken azınlık veri miktarının yarısı ve kendisi kadar ayrıca çoğunluk veri miktarının %1’i, %5’i, %10’u ve kendisi kadar üretebilecek oranlar olmasına dikkat edilmiştir. Böylece orijinal azınlık verilerinden üretilen sentetik veri miktarı ile performans seviyesindeki değişim arasındaki ilişki test edilmiştir. Ayrıca az örnekleme için çoğunluk veri miktarı ne kadar azalırsa aynı oranda bilgi kaybı olmadığı da gözlemlenmiştir.

Çizelge 3.2. Kullanılan örnekleme yöntemi ve oranları

Yöntem ve Oran						
Aşırı Örnekleme	SMOTE (0,0025)	SMOTE (0,0034)	SMOTE (0,01)	SMOTE (0,05)	SMOTE (0,1)	SMOTE (1)

Az Örnekleme	-	RUS(0,005)	RUS(0,01)	RUS(0,1)	RUS(0,5)	RUS(1)
--------------	---	------------	-----------	----------	----------	--------

Makine öğrenmesi birtakım süreçlerden oluşmaktadır. Genel olarak öncelikle verinin işlenmesi, eğitim ve test veri kümelerine ayrılması sonra eğitim ve daha sonrasında test işlemlerinden geçerek modelin üretilmesi olarak ifade edilebilir. Yeniden örnekleme metotlarının ise kullanıldıkları aşamalar çalışmadan çalışmaya değişiklik gösterebilmektedir (Bkz. Çizelge 2.6). Bu sebeple örnekleme metotlarının kullanıldıkları aşamaların tespit başarılarına ne tür etkilerinin olduğunu değerlendirmek için 4 farklı deney ortamı oluşturulmuştur:

1. Dengesiz veri kümesinde modellerin sonuçlarını görebilmek için ilk ortamda herhangi bir örnekleme metodu kullanılmamıştır.
2. Dengeli bir test ortamında modellerin sonuçlarını görebilmek için ikinci ortamda örnekleme metotları sadece test veri kümesine uygulanarak dengeli hale gelmiştir.
3. Dengeli bir eğitim kümesinin etkilerini görebilmek için üçüncü ortamda örnekleme metotları eğitim ve test veri kümelerine ayrı ayrı uygulanarak her ikisi de dengeli hale getirilmiştir.
4. Örnekleme metotlarının eğitim test ayırımına etkisini görebilmek için son ortamda tüm veri kümesine örnekleme metotları uygulanıp veri kümesi dengeli hale geldikten sonra eğitim, test veri kümesi ayırımı yapılmaktadır.

### 3.2.2. Öznitelik seçimi

Öznitelik seçimi bir model geliştirirken tüm özniteliklerden bir alt küme oluşturma işlemidir. Çeşitli yöntemler kullanılarak bu alt kümede bulunacak öznitelikler belirlenir. Böylece işlem maliyeti önemli oranlarda azalabilmektedir. Kredi kartı dolandırıcılık tespit sistemlerinde kredi kartı işlem tespit süreleri oldukça önemlidir. Ağır işleyen bir tespit sistemi özellikle kullanıcı bazında bir memnuniyetsizliğe neden olabilmektedir. Öznitelik seçimi yaparak aşırı uyum problemini kısmen çözebilmek ve modelin tespit başarısını geliştirebilmek de mümkündür. Dolayısıyla özellikle dolandırıcılık işlemlerindeki tespit

başarı seviyesindeki artış maddi kayıplarda azalma anlamına gelmektedir. Bu çalışma kapsamında 4 farklı öznitelik metodu kullanılarak oluşturulan 4 farklı öznitelik alt kümesi ile bu hedeflere ulaşılmaya çalışılmıştır.

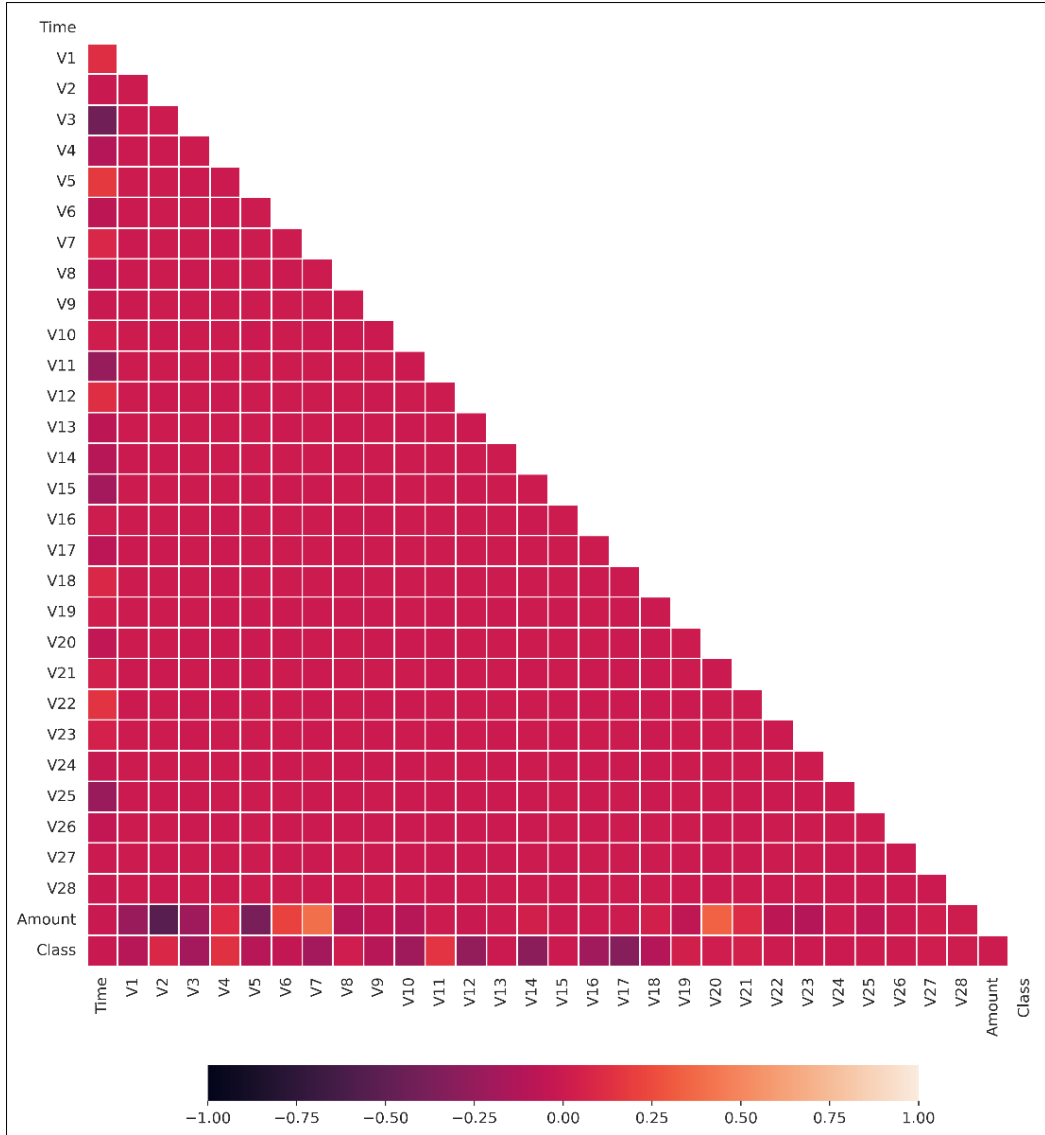
Pearson Korelasyonu bu çalışmada kullanılan öznitelik seçimi yöntemlerinden ilkidir. Bu yöntem öznitelikler arasındaki ilişkiyi anlamak için kullanılmaktadır. Kredi kartı dolandırıcılığı kapsamında bir kredi kartı işleminin normal ya da dolandırıcılık tespiti için hangi öznitelik geliştirilen modele daha çok katkı sağlayacağı bu yöntemle görülebilmektedir. Ayrıca hedef değişkenle ilişkili birden çok özneliğin olup olmadığı da bu şekilde tespit edilebilmektedir. Pearson Korelasyon yöntemi ile 3 farklı ilişki elde edilebilmektedir:

- Direkt ilişki: Bir öznitelikle bir başka özneliğin aynı durumlarda artması ya da azalması durumudur.
- Ters ilişki: Bir öznitelik artarken diğer özneliğin azaldığı ve aynı şekilde diğer öznitelik artarken ilk özneliğin azalması durumudur.
- İlişki yok: Bir öznitelik ile bir başka özneliğe bağlı olarak artma ya da azalma göstermemesi yani bir ilişkisinin olmamasıdır.

European Cardholders veri kümesi öznitelik korelasyon grafiği Şekil 3.2’de gösterilmiştir. Burada renkler beyaz tonuna giderken direkt ilişki, kırmızı tonunda korelasyon yok, siyah tonuna doğru ise ters ilişki anlamına gelmektedir.

Öznitelik önemliliği yöntemi öznitelikleri önemliliğine göre sıralayan bir öznitelik seçimi metodudur. Bu yöntem kapsamında en önemli  $n$  adet öznitelik seçilip makine öğrenmesi yöntemine girdi olarak verilir. Öznitelik önemlilik metotları olarak bu çalışmada RF ve Lasso Regresyon yöntemleri ve sarmalayıcı yöntem olarak RFE kullanılmıştır. Bu yöntemler ile elde edilen öznitelik sıralamaları Çizelge 3.3’te paylaşılmıştır. Her bir makine öğrenmesi metodu bu öznitelik sıralamalarından ilk 5’i, 10’u, 15’i, 20’si, 25’i ve 29’u (tamamı) seçerek karşılaştırılmıştır.





Şekil 3.2. Öznitelik korelasyon grafiği

Çizelge 3.3. Öznitelik seçimi ile oluşturulan öznitelik alt kümeleri

Yöntem	İsim	Sıralama
Pearson Korelasyon	altv <sub>1</sub>	V17, V14, V12, V10, V16, V3, V7, V18, V1, V9, V5, V6, V24, V13, V15, V23, V22, V25, V26, normAmount, V28, V27, V8, V20, V19, V21, V2, V4, V11, Class
RF	altv <sub>2</sub>	V14, V11, V12, V10, V3, V17, V4, V7, V16, V9, V2, V6, V21, V18, V1, normAmount, V8, V20, V5, V13, V28, V19, V26, V23, V27, V15, V25, V24, V22, Class
Lasso Regresyonu	altv <sub>3</sub>	V14, V4, Class

RFE	altv <sub>4</sub>	V10, V14, V4, V11, V12, V17, V3, V8, V20, V16, V7, V18, normAmount, V19, V13, V2, V9, V1, V5, V21, V6, V26, V23, V25, V28, V15, V24, V22, V27, Class
-----	-------------------	--

### 3.2.3. Makine öğrenmesi ve değerlendirme metrikleri

Makine öğrenmesi verilerdeki karakteristiği çeşitli yöntemler aracılığıyla kendi kendine öğrenebilen ve sonrasında önceden görülmemiş verileri sınıflayabilen metotlardır. Makine öğrenmesi metotları için girdi bu çalışmada European Cardholders veri kümesidir. Bkz. Çizelge 3.2’de belirtilen çeşitli örnekleme oranları dengesiz olan European Cardholders veri kümesine uygulanmıştır ve sonrasında makine öğrenmesi metotları (RF, LGBM, XGB, LR ve SVM) 5 Katlamalı Çapraz Doğrulama (5 Fold Cross Validation) ile eğitilmiştir. Sonuç olarak test ile elde edilen sonuçlar arasında en iyi örnekleme oranı bulunmuştur. En iyi örnekleme oranı seçilerek, örnekleme metotlarının kullanıldığı aşamaları karşılaştırabilmek için Bkz. Bölüm 3.2.1’de belirtilen 4 farklı deney ortamında makine öğrenmesi yöntemleri uygulanmış ve elde edilen sonuçlar karşılaştırılmıştır. Bu çalışmalar neticesinde en iyi örnekleme oranı ve güvenilirliği diğerlerine kıyasla yüksek olan örnekleme metotlarının kullanım aşaması belirlenmiştir. Sonrasında ise Çizelge 3.3’te verilen öznitelik alt kümeleri bu örnekleme oranı ve kullanım aşaması ile kullanılmıştır. Bölüm 3.2.2’de de belirtildiği üzere Lasso regresyon harici diğer öznitelik alt kümelerinin (altv<sub>1</sub>, altv<sub>2</sub> ve altv<sub>4</sub>) ilk 5, 10, 15, 20, 25 ve 29 (tamamı) özniteliği seçilerek karşılaştırılmıştır.

Son adım olarak değerlendirme metrikleri geliştirilen modelin tespit başarısını ölçmek amacıyla kullanılmaktadır. Normal verilerinin tespitini kapsayan bazı metrikler dengesiz veri kümesi problemi nedeniyle yanlış sonuçlar yansıtabilmektedir. Örneğin 100 işleme sahip veri kümesinde 4 dolandırıcılık 96 normal işlem verisi olduğu varsayalım. Normal verilerin hepsini doğru, dolandırıcılık verilerini ise yanlış tespit ettiğinde model doğruluk metriği ile %96 başarı elde ettiği görülecektir ancak dolandırıcılık tespiti sistemi herhangi bir dolandırıcılık tespit edememiştir. Bunun gibi dengesizlik probleminin metrikler üzerindeki etkileri Bulgular ve Tartışma bölümü Yeniden Örnekleme kısmı altında incelenmektedir.

Bu çalışma dolandırıcılık tespiti üzerinde başarılı olan model gelişimine odaklandığı için özellikle dolandırıcılık verilerine yoğunlaşan metriklerden (Duyarlılık, Kesinlik) faydalanılmıştır. Dolayısıyla modelin mümkün olabildiğince yüksek Duyarlılık değeri elde etmesi ilk hedeftir ki bu değer dolandırıcılık verilerini ne kadar doğru tespit ettiğini gösterir. Göz ardı edilemeyecek Kesinlik metriği ise aslında normal işlem verilerini dolandırıcılık olarak tespit etme oranını gösterir. F1-Skor metriği ise Kesinlik ve Duyarlılık metriklerinin harmonik ortalaması ile hesaplandığı için bu çalışmada geliştirilen modellerin değerlendirilmesi açısından önemlidir.



#### 4. BULGULAR VE TARTIŞMA

Bu bölümde yapılan denemeler sonucunda öncelikle yeniden örnekleme metotlarının öğrenme yöntemlerine etkileri incelenmiştir. Sonrasında örnekleme metotlarının kullanıldıkları aşamalara göre kıyaslamaları yapılmıştır. Bu kıyaslamaların takibinde öznitelik seçimi metotlarının makine öğrenmesi metotlarına etkileri incelenmiştir.

##### 4.1. Yeniden Örnekleme

European Cardholders veri kümesindeki dengesizlik problemini ele alırken en yüksek başarı seviyesine ulaşmak için 6 farklı deney ortamı oluşturulmuştur (Bkz. Bölüm 3.2.1). Bu deney ortamlarından ikincisi içerisindeki eğitim veri kümesi için SMOTE(0,0034) ile RUS(1) kombinasyonu en iyi sonucu verdiği gözlemlenmiştir (Çizelge 4.1). Çizelge 4.2'deki veri miktarlarındaki değişime bakıldığında aşırı örnekleme metodu ile azınlık veri miktarının kendisi kadar veri üretebilmesi yeterli olduğu bulunmuştur. Aynı şekilde çoğunluk veri miktarına bakıldığında ise az örnekleme için yüksek bir oran kullanılarak yaklaşık %99'u silinmiş ancak en iyi sonucu vermiştir.

Çizelge 4.1. Yeniden örnekleme ile veri miktarlarındaki değişim (Acc: Accuracy, Pre: Precision, Rec: Recall, F1: F1-Score, S: SMOTE, R: RUS)

Makine Öğrenmesi	Yeniden Örnekleme		Değerlendirme Metrikleri (%)					
	Eğitim	Test	Acc	Pre	Rec	F1	AUROC	AUPRC
XGB	-	R(1)	89,7	99,7	79,3	88,3	89,7	94,8
	S(0,0034)	R(1)	90,9	99,8	81,8	89,9	90,9	95,6
	S(0,0034), R(0,005)	R(1)	90,9	99,7	82,8	90,5	90,9	94,8
	S(0,0034), R(0,01)	R(1)	90,9	99,7	82,9	90,5	90,9	95,9
	S(0,0034), R(0,1)	R(1)	93,1	99,7	85,7	92,2	93,1	95,8
	S(0,0034), R(0,5)	R(1)	93,2	98,4	88,3	93,1	93,2	95,9
	S(0,0034), R(1)	R(1)	93,8	98,4	90,1	94,1	93,8	96,2

Çizelge 4.2. Yeniden örnekleme ile veri miktarlarındaki değişim

	<b>Yeniden Örnekleme</b>	<b>Azınlık Veri Miktarı</b>	<b>Çoğunluk Veri Miktarı</b>	<b>Toplam</b>
Başlangıç	-	492	284.315	284.807
1. Adım	S(0,0034)	966	284.315	285.281
2. Adım	R(1)	966	966	1932

Karmaşık bir sonuç tablosu olmaması amacıyla sadece SMOTE metodunun en iyi sonuç veren parametresi ile RUS metodu için kullanılan örnekleme oranı değerlerinin tamamı XGB metodu aracılığıyla verilmiştir (Bkz. Çizelge 4.1). İlk satırda gösterilen yeniden örneklemenin kullanılmadığı eğitim veri kümesine kıyasla ikinci satırda gösterilen SMOTE(0,0034) yönteminin kullanıldığı eğitim kümesi ile Duyarlılık metriğinde %3 artış elde edilmiştir. Benzer şekilde RUS metodu 0,1 ve 1 arası örnekleme oranları ile çoğunluk veri miktarındaki azalmanın dolandırıcılık tespitine olumlu katkı yaptığı görülmesine rağmen Kesinlik metriği %2 oranında azalmıştır.

Yeniden örnekleme metotlarının kullanıldıkları aşamaların tespit başarılarına ne tür etkilerinin olduğunu değerlendirmek için 4 farklı deney ortamı üzerindeki sonuçlar Çizelge 4.3'te paylaşılmıştır. İlk deney ortamında, metotların elde ettiği Doğruluk, Kesinlik ve AUPRC değerleri dengesizlik probleminden etkilenmiştir. Test kümesi dengeli hale geldiğinde (ikinci deney ortamı) bu metrik değerleri olması gereken değerlere dönüşmüştür çünkü bu metrikler dengeli veri kümelerine uygun tasarlanmıştır. Doğruluk metriği, pozitif ve negatif verilere dayalı hesaplanmaktadır (Bkz. Denklem 2.11). Bu problemde olduğu gibi, negatif veriler pozitif verilerden sayıca fazla olabilir ve pozitif sınıfın (dolandırıcılık) da tespiti negatif sınıfa (normal) göre daha önemlidir. Fakat dengesizlik nedeniyle, pozitif sınıfın düşük seviyede tespiti negatif sınıfın yüksek seviyede tespitinin altında kalmaktadır ve Doğruluk metriği, her ne kadar dolandırıcılık verilerinin tespitinde başarısız olsa da böyle bir senaryoda yüksek değer vermektedir. Bununla birlikte AUROC ve AUPRC metrikleri de dengesiz veri kümelerinde dikkat edilmesi gereken metriklerdendir. FP değerindeki yüksek derecedeki değişiklik AUROC'te kullanılan FPR değerini çok fazla etkilememekte iken Kesinlik metriğini etkilemektedir. Bu sebeple dengesiz bir ortamda Kesinlik ve buna bağlı AUPRC metrikleri dengeli ortama kıyasla daha düşük çıkabilmektedir. Sonuç olarak ilk iki deney ortamı arasındaki fark dengesizlik probleminin metrikler üzerindeki yanıltıcı etkisini

göstermektedir. İkinci ve üçüncü deney ortamları arasındaki karşılaştırma ise eğitim kümesine yeniden örnekleme metotları uygulamanın modelin tespit yeteneğine olumlu etkisini göstermektedir. Eğitim kümesinde örnekleme metotları kullanılarak eğitilen tüm modeller, Kesinlik hariç diğer tüm metriklerde artış sağlamıştır. Kesinlik metriğindeki azalmanın sebebi ise az örnekleme nedeniyle oluşan normal işlem verilerindeki bilgi kaybıdır. Bununla birlikte eğitim veri kümesine uygulanan örnekleme metotlarının en çok katkı sağladığı öğrenme metodu LGBM olarak bulunmuştur. Bu deneyler arasında en önemlisi, üçüncü ve dördüncü deney ortamlarının karşılaştırılması ise örnekleme metotlarını eğitim-test ayırımından önce veya sonra uygulamanın doğuracağı bir başka yanıltıcı artışı ifade etmektedir. Dördüncü deney ortamında eğitim-test ayırımından önce aşırı örnekleme metotları tüm veri kümesine uygulanır ve sonrasında eğitim-test ayrımı yapılırken, orijinal bir verinin eğitim kümesine, bu verinin örnekleme metodu ile yaratılan kopyasının ise test veri kümesine dağılma ihtimali vardır. Bu ihtimal gerçekleştiğinde öğrenme yöntemi benzer veriler ile eğitim ve test işlemini yapmış olmaktadır ancak test veri kümesinin en önemli özelliği önceden görülmemiş, eğitim kümesinden farklı verileri bulundurması gerektiğidir. Dolayısıyla, dördüncü deney ortamında, öğrenme metodunun eğitildiği verilerin benzeriyle test yapılması, bu metodun başarısını yükseltmektir. Birinci deney ortamında Doğruluk metriği değerleri gibi dördüncü deney ortamında da metrik değerleri yanlış algıya yaratmaktadır. Sonuç olarak dört deney ortamı arasında en güvenilir değerlerin üçüncü deney ortamından elde edildiği çıkarımı yapılmıştır.

Makine öğrenmesi yöntemleri bakımından karşılaştırma yapıldığında, üçüncü satırlar dikkate alınarak Duyarlılık metriğinde LR, XGB, LGBM, RF ve SVM metotlarının sonuçları sırasıyla şöyledir: %91, %90, %90, %88 ve %88. Kesinlik metriğinde ise RF, XGB, LGBM, SVM ve LR metotlarının sonuçları şöyledir: %99, %98, %98, %98, %97. Doğruluk metriğinde XGB, LGBM, LR, RF ve SVM metotlarının sonuçları şöyledir: %94, %94, %94, %93, %93.

Çizelge 4.3. Yeniden örnekleme metotlarının kullandığı aşamalara göre makine öğrenmesi metotlarına etkisi

Makine Öğrenmesi	Yeniden Örnekleme	Değerlendirme Metrikleri (%)					
		Acc	Pre	Rec	F1	AUROC	AUPRC
XGB	1. deney ortamı	<b>99,8</b>	94,2	78,8	85,8	89,7	87,3
	2. deney ortamı	90,1	<b>99,8</b>	78,8	88,1	89,6	95,2
	3. deney ortamı	94,2	98,3	90,1	94	94	96,4
	4. deney ortamı	96,8	99,4	94,7	97	96,9	<b>98,2</b>
LGBM	1. deney ortamı	99,7	46,8	38,2	42,1	68,6	42,1
	2. deney ortamı	68,3	99,7	38,2	55,2	68,6	83,8
	3. deney ortamı	93,6	97,8	89,8	93,6	93,7	95,7
	4. deney ortamı	97,2	98,3	<b>96,2</b>	<b>97,2</b>	<b>97,1</b>	98,1
RF	1. deney ortamı	99,6	94,8	77,6	85,3	89,3	87,1
	2. deney ortamı	89,1	99,6	79,1	88,2	89,3	94,7
	3. deney ortamı	92,8	98,7	87,6	92,8	93,2	96,2
	4. deney ortamı	95,9	98,7	94,3	96,4	95,6	97,7
LR	1. deney ortamı	99,7	87,2	62,1	72,5	80,8	75,2
	2. deney ortamı	80,8	99,7	62,1	76,5	80,8	90,8
	3. deney ortamı	94,8	96,6	91,3	93,9	94,1	96,3
	4. deney ortamı	95	96,9	91,8	94,3	94,8	96,9
SVM	1. deney ortamı	99,6	95,4	68,1	79,5	84,3	81,1
	2. deney ortamı	84,3	99,6	68,1	80,9	84,4	91,8
	3. deney ortamı	93,2	97,8	87,7	92,5	93,2	96,4
	4. deney ortamı	94,7	97,8	91,3	94,4	94,8	96,7

#### 4.2.Öznitelik Seçimi

Makine öğrenmesi metotları eğilirken ilgili işlem maliyetini azaltmak ve performansı aynı seviyede tutabilmek için 4 farklı öznitelik alt kümesi oluşturuldu (Bkz. Çizelge 3.3). Her bir öznitelik alt kümesi için sonuçlar sırasıyla Çizelge 4.4, Çizelge 4.5 ve Çizelge 4.6 ve Çizelge 4.7’de verilmiştir. Lasso regresyon metodu (Çizelge 4.6) hariç diğer tüm öznitelik alt kümelerinden ilk 5, 10, 20 ve 29 öznitelik seçilerek makine öğrenme yöntemleri uygulanmıştır. Sonuç olarak öznitelik sayıları azalırken, eğitim-test (işlem)

süreleri ve değerlendirme metrikleri bakımından metotlar şu şekilde sonuçlar vermiştir: XGB algoritmasında işlem süresi azalırken, metriklerde en fazla %3'e kadar düşüş gözlemlenmiştir; LGBM algoritmasında ise işlem süresinde artma görülmüş ve metrikler bazında en fazla %2'ye kadar düşüş gözlemlenmiştir; RF algoritmasında da işlem sürelerinde önemli derecede azalmalar görülmüş ve bu algorithmada da başarı seviyesinde en fazla %2'e kadar düşüş gözlemlenmiştir. LR algoritmasında işlem süresi azalırken, değerlendirme metriklerinde %4 ile diğer algoritmalara kıyasla en fazla düşüşe sahip olmuştur. Son olarak SVM metodunda işlem süresi azalırken, Kesinlik metriğinde bir artış ve diğer metriklerde en fazla %3'e kadar düşüş görülmüştür. Bununla birlikte, en az öznitelik ile öğrenme yöntemlerinin AUROC ve AUPRC değerleri incelendiğinde, RFE'nin, diğer öznitelik seçimi metotları arasından öne çıktığı görülmektedir. Öğrenme yöntemlerinin AUROC ve AUPRC grafikleri Şekil 4.1 ve Şekil 4.2'de sırayla paylaşılmıştır.

Çizelge 4.4. Pearson Korelasyon öznitelik seçimi ile makine öğrenmesi metotları

Makine Öğrenmesi	Öznitelik Sayısı	Eğitim-Test Süreleri (ms)	Değerlendirme Metrikleri (%)					
			Acc	Pre	Rec	F1	AUROC	AUPRC
XGB	29	398 – 10	<b>93,9</b>	98,4	89,8	<b>93,9</b>	<b>94,2</b>	95,7
	20	403 – 10	93,7	98,4	89,4	93,7	93,7	96,1
	10	342 – 11	93,1	97,1	88,2	92,4	93,1	96,2
	5	353 – 13	91,7	96,7	88,1	92,2	91,7	95,1
LGBM	29	995 – 5	93,6	97,8	89,7	93,6	93,8	95,6
	20	1080 – 7	92,7	97,9	89,2	93,3	93,6	95,7
	10	1128 – 6	92,6	97,8	87,6	92,4	93,2	95,9
	5	1228 – 8	92,6	97,3	89,4	93,2	92,6	95,6
RF	29	1229 – 30	93,7	98,6	89,3	93,7	93,8	<b>97,2</b>
	20	1061 – 31	92,6	98,7	87,6	92,8	92,9	95,9
	10	866 – 32	92,6	97,6	88,1	92,6	92,6	96,2
	5	690 – 30	92,8	98,2	88,2	92,9	92,8	95,8
LR	29	96 – 0	93,6	96,7	<b>90,7</b>	93,6	93,6	96,2
	20	59 – 0	93,1	96,6	88,4	92,3	92,8	95,8
	10	36 – 1	92,8	97,7	86,7	91,9	93,1	96,3



	5	14 – 0	92,9	97,8	86,8	92	93,2	95,9
SVM	29	82 – 21	92,6	98,3	87,8	92,8	93	95,7
	20	85 – 25	92,7	98,9	87,1	92,6	93,1	95,6
	10	64 – 19	92,1	99,1	85,7	91,9	91,8	95,7
	5	60 – 17	92	<b>99,8</b>	84,6	91,6	91,9	96,2

Çizelge 4.5. RF öznitelik seçimi ile makine öğrenmesi metotları

Makine Öğrenmesi	Öznitelik Sayısı	Eğitim-Test Süreleri (ms)	Değerlendirme Metrikleri (%)					
			Acc	Pre	Rec	F1	AUROC	AUPRC
XGB	29	398 – 12	93,7	97,8	90,8	<b>94,2</b>	93,7	96,7
	20	383 – 11	93,6	97,2	<b>90,9</b>	93,9	93,7	96,4
	10	330 – 10	93,8	97,1	90,2	93,5	93,8	95,9
	5	353 – 11	93,2	97,1	87,7	92,2	93,1	95,8
LGBM	29	982 – 6	93,6	98,2	89,6	93,7	93,8	96,2
	20	877 – 4	94,1	97,1	89,7	93,3	94	95,9
	10	1027 – 8	93,8	97,6	89,6	93,4	93,9	96,2
	5	1198 – 10	93,2	97,8	88,4	92,9	93,1	95,7
RF	29	1241 – 30	93,6	98,7	88,7	93,4	93,8	<b>97,1</b>
	20	1017 – 31	94,2	98,1	88,6	93,1	94,1	97
	10	852 – 31	93,6	97,8	89,6	93,5	93,7	<b>97,1</b>
	5	699 – 30	92,8	98,1	87,9	92,7	93,2	96,4
LR	29	90 – 1	94,1	96,7	90,6	93,6	93,9	96,1
	20	65 – 0	<b>94,7</b>	97,7	90,6	94	<b>94,9</b>	96,8
	10	44 – 0	94,2	98,1	90,2	94	94,2	95,6
	5	13 – 0	93,4	97,8	88,1	92,7	93,4	95,7
SVM	29	77 – 22	93,1	97,7	88,3	92,8	93,2	96,1
	20	63 – 20	93,1	97,6	88	92,6	93	96,2
	10	52 – 17	92,8	98,9	88,1	93,2	92,9	96,1
	5	60 – 16	92,3	<b>99,8</b>	84,6	91,6	92,4	95,8

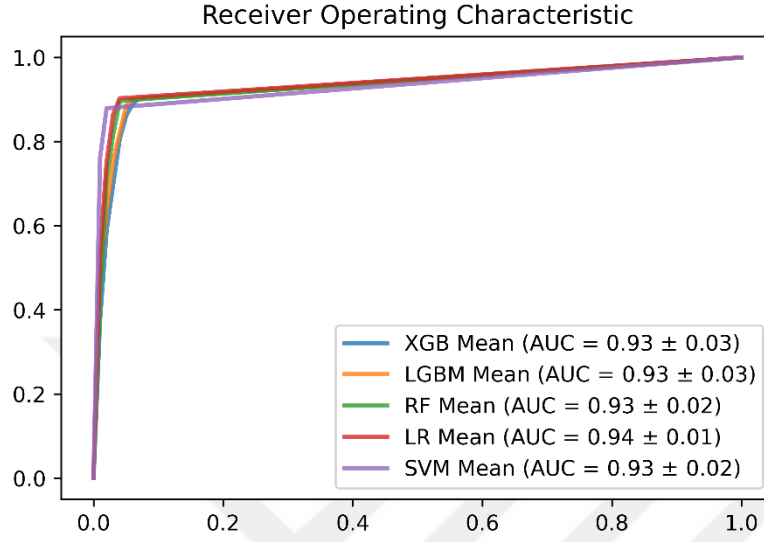
Çizelge 4.6. Lasso Regresyon öznitelik seçimi ile makine öğrenmesi metotları

Makine Öğrenmesi	Öznitelik Sayısı	Eğitim-Test Süreleri (ms)	Değerlendirme Metrikleri (%)					
			Acc	Pre	Rec	F1	AUROC	AUPRC
XGB	2	366 – 14	92,1	93,7	89,1	91,3	92	93,8
LGBM	2	1732 – 12	91,8	94,6	89	91,7	91,8	95,1
RF	2	516 – 32	91,9	95,2	89	92	91,8	94,9
LR	2	14 – 0	<b>92,7</b>	95,8	<b>89,8</b>	<b>92,7</b>	92,8	94,8
SVM	2	48 – 16	92,6	<b>97,8</b>	86,6	91,9	<b>92,9</b>	<b>96,1</b>

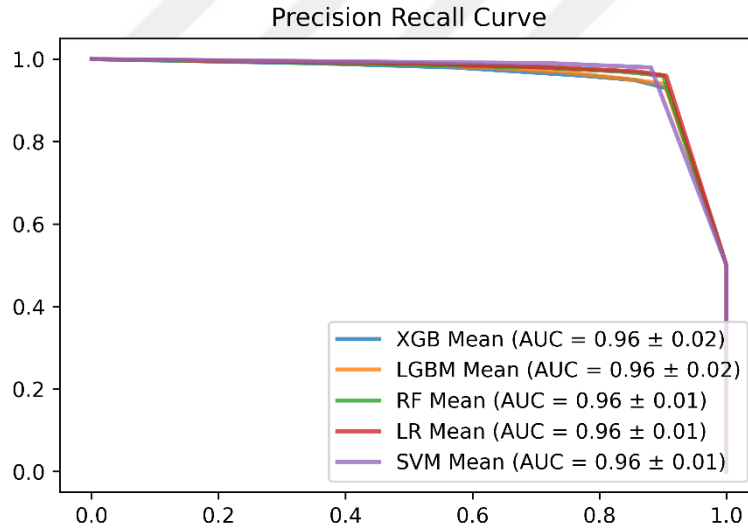
Çizelge 4.7. RFE ile makine öğrenmesi metotları

Makine Öğrenmesi	Öznitelik Sayısı	Eğitim-Test Süreleri (ms)	Değerlendirme Metrikleri (%)					
			Acc	Pre	Rec	F1	AUROC	AUPRC
XGB	29	408 – 11	93,8	98,2	90,3	94,1	93,8	<b>97,4</b>
	20	371 – 11	93,7	98,1	90,2	94	93,7	97,3
	10	329 – 11	93,7	97,9	90,3	93,9	93,7	97,1
	5	375 – 12	93,4	96,3	89,8	92,9	93,4	96,4
LGBM	29	982 – 6	93,9	98,4	90,2	94,1	93,8	96,4
	20	931 – 7	93,7	98,3	90,2	94,1	93,7	96,7
	10	974 – 8	93,6	97,8	90	93,7	93,7	96,7
	5	1113 – 9	93,4	97,3	89,6	93,3	93,4	96,4
RF	29	1241 – 30	94,3	<b>99,2</b>	89,2	93,9	94,4	96,8
	20	1032 – 30	94,2	98,4	89,1	93,5	94,2	96,4
	10	856 – 33	93,8	99,1	89,8	<b>94,2</b>	93,7	96,8
	5	669 – 30	93,2	97,4	89,3	93,2	93,1	96,2
LR	29	88 – 0	93,8	97,3	<b>91,1</b>	94,1	93,7	95,9
	20	64 – 0	<b>94,9</b>	97,6	90,9	94,1	<b>94,8</b>	96,6
	10	43 – 1	94,7	97,6	90,8	94,1	94,7	96,7
	5	15 – 1	93,8	97,9	90,4	94	94,1	96,3
SVM	29	76 – 23	92,9	98,2	88,1	92,9	92,8	96,2

	20	67 – 19	92,7	98,1	88,2	92,9	92,7	96,3
	10	55 – 16	93,7	97,8	88,6	93	93,6	95,8
	5	52 – 15	93,1	98,7	88,4	93,3	93	95,6



Şekil 4.1. Metotların ROC Eğrileri



Şekil 4.2. Metotların PR Eğrileri

## 5. SONUÇ VE ÖNERİLER

Günümüz bilgisayarları yüksek boyutlu veri kümelerini hızlı bir şekilde işleyebildiği için büyük veri, veri madenciliği ve makine öğrenmesi gibi kavramlar daha yaygın hale gelmiş ve kredi kartı dolandırıcılık tespiti için yararlanılan yöntemler arasına girmiştir. Bununla birlikte makine öğrenmesi yöntemleriyle geliştirilen bir dolandırıcılık tespit sisteminin hızlı ve güvenilir çalışması gerekmektedir. Bu nedenle makine öğrenmesi yöntemlerinin en önemli ihtiyacı olan verilerin doğru ön işlemlerden geçirilmesi ihtiyacı ki bu amaç doğrultusunda yeniden örnekleme ve öznitelik seçimi metotları kullanılmıştır.

Yürütülen bu çalışma kapsamında makine öğrenmesi kullanarak yapılan kredi kartı dolandırıcılık tespitinde temel olarak dengesizlik problemi, örnekleme metotlarının kullanım aşamaları ve değerlendirme metrikleri üzerine odaklanılmıştır. Örnekleme metotları ve değerlendirme metriklerinin hangi durumlarda yanıltıcı sonuçlar üretebildiği görülmüştür: Test veri kümesinin dengesiz olması; Dengesiz test veri kümesinde Doğruluk, AUROC metriklerinin kullanımı; Örnekleme metotlarının tüm veri kümesine uygulanması. Ayrıca SMOTE ile dolandırıcılık işlemlerinin tespitinde önemli derecede başarı artışı elde edilmiştir. Bununla birlikte dolandırıcılık tespiti sisteminin daha hızlı çalışabilmesi adına öznitelik seçimi metotları kullanılmış ve bunun en belirgin örneği RF metodunda görülmekte iken XGB, LR ve SVM gibi metotlarda nispeten daha az görülmüştür. Diğerlerinin aksine LGBM metodunda ise işlem süresi artmıştır.

Kredi kartı dolandırıcılık tespiti için veri kümeleri oldukça önemlidir. Ancak literatür incelendiğinde erişilebilir veri kümeleri bağlamında yetersizlik olduğu görülmüştür. Kredi kartı işlemlerini mevsimler ve yıllar süresince kapsayan veri kümeleri erişilebilir hale geldiğinde, kişilere özgü satın alma davranışları üzerinden daha gelişmiş modeller üretilir. Bununla birlikte kredi kartı dolandırıcılık tespitine özel yeniden örnekleme metotları ve öznitelik seçimi metotları geliştirilebilir.

## KAYNAKLAR DİZİNİ

- Abdulsalami, B. A., Kolawole, A.A., Ogunrinde, M.A., Lawal, M., Azeez, R.A., vd., 2019, Comparative Analysis of Back-propagation Neural Network and K-Means Clustering Algorithm in Fraud Detection in Online Credit Card Transactions, Fountain Journal of Natural and Applied Sciences, 8(1).
- Alam, T. M., Shaukat, K., Hameed, I.A., Luo, S., Sarwar, M.U. vd., 2020, An investigation of credit card default prediction in the imbalanced datasets. IEEE Access, 8, p.201173-201198.
- Alanezi, M. A., Homeed, M.T., Mohamed, Z.S., Zeki, A.M., 2020, Comparing Naïve BaEvet, Decision Tree and Logistic Regression Methods in Fraudulent Credit Card Transactions, In 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), p.1-5, IEEE.
- Alfaiz, N. S., Fati, S. M., 2022, Enhanced Credit Card Fraud Detection Model Using Machine Learning, Electronics, 11(4), p.662.
- Al-Shabi, 2019, M.A., 2019, Credit card fraud detection using autoencoder model in unbalanced datasets, Journal of Advances in Mathematics and Computer Science, 33(5), p.1-16.
- Ata, O., Hazim, L., 2020, Comparative analysis of different distributions dataset by using data mining techniques on credit card fraud detection, Tehnički vjesnik, 27(2), p.618-626.
- Aung, M. H., Seluka, P.T., Fuata, J.T.R., Tikoisuva, M.J., Cabealawa, M.S., vd., 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), p.1-6, IEEE.
- Babu, A.M., Pratap, A., 2020, Credit Card Fraud Detection Using Deep Learning, In 2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS), p.32-36, IEEE.
- Bej, S., Davtyan, N., Wolfien, M., Nassar, M., Wolkenhauer, O., 2021, LoRAS: An oversampling approach for imbalanced datasets. Machine Learning, 110(2), p.279-301.
- bin Alias, M. S. A., Ibrahim, N. B., 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.
- Blum, A.L., Langley, P., 1997, Selection of relevant features and examples in machine learning, Artificial intelligence, 97(1-2), pp.245-271.

### KAYNAKLAR DİZİNİ (devam)

- Budianto, I. R., Azaria, R. K., Gunawan, A. A., 2022, Machine Learning-based Approach on Dealing with Binary Classification Problem in Imbalanced Financial Data, In 2021 International Seminar on Machine Learning, Optimization, and Data Science, p.152-156, IEEE.
- Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J. and Layton, R., 2013, API design for machine learning software: experiences from the scikit-learn project, arXiv preprint arXiv:1309.0238.
- Caroline Cynthia, P., Thomas George, S., 2021, An outlier detection approach on credit card fraud detection using machine learning: a comparative analysis on supervised and unsupervised learning, In Intelligence in Big Data Technologies—Beyond the Hype, p.125-135, Springer, Singapore.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2002, SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, p.321-357.
- Cortes, C., Vapnik, V., 1995, Support vector machine, Machine learning, 20(3), pp.273-297.
- Dal Pozzolo, A. 2015, Adaptive machine learning for credit card fraud detection.
- Dhankhad, S., Mohammed, E., Far, B., 2018, Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study, In 2018 IEEE international conference on information reuse and integration (IRI), p.122-125.
- Dighe, D., Patil, S., Kokate, S., 2018, Detection of credit card fraud transactions using machine learning algorithms and neural networks: A comparative study, In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), p.1-6, IEEE.
- Dornadula, V.N., Geetha, S., 2019, Credit card fraud detection using machine learning algorithms, Procedia computer science, 165, p.631-641.
- Drummond, C., Holte, R. C., 2003, C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, In Workshop on learning from imbalanced datasets II (Vol. 11, pp. 1-8).
- Dua, D., Graff, C., 2019, UCI Machine Learning Repository Credit. Approval Dataset, <https://archive.ics.uci.edu/ml/datasets/Credit+Approval>, erişim tarihi: 18.06.2022
- Efron, B., 1982, The jackknife, the bootstrap and other resampling plans, Society for industrial and applied mathematics.

## KAYNAKLAR DİZİNİ (devam)

- El Hajjami, S., Malki, J., Berrada, M., Fourka, B., 2020, Machine learning for anomaly detection, performance study considering anomaly distribution in an imbalanced dataset, In 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), pp. 1-8, IEEE.
- Federal Trade Commission, 2022, Data Book 2021, [https://www.ftc.gov/system/files/ftc\\_gov/pdf/CSN%20Annual%20Data%20Book%202021%20Final%20PDF.pdf](https://www.ftc.gov/system/files/ftc_gov/pdf/CSN%20Annual%20Data%20Book%202021%20Final%20PDF.pdf), erişim tarihi: 16.04.2022
- FreeCodeCamp, Receiver Operating Characteristics Plot, <https://www.freecodecamp.org/news/content/images/2020/08/how-random-forest-classifier-work.PNG>, erişim tarihi: 16.04.2022
- Friedman, J. H., 2001, Greedy function approximation: a gradient boosting machine, Annals of statistics, p.1189-1232.
- Georgieva, S., Markova, M., Pavlov, V., 2019, Using neural network for credit card fraud detection, In AIP Conference Proceedings, 2159(1), p.030013, AIP Publishing LLC.
- Gulati, P., 2020, Hybrid resampling technique to tackle the imbalanced classification problem.
- Hordri, N.F., Yuhaniz, S.S., Azmi, N.F.M., Shamsuddin, S.M., 2018, Handling class imbalance in credit card fraud using resampling methods, Int. J. Adv. Comput. Sci. Appl, 9(11), p.390-396.
- Husejinovic, A., 2020, Credit card fraud detection using naive BaEvetian and c4. 5 decision tree classifiers, Husejinovic, A.(2020), Credit card fraud detection using naive BaEvetian and C, 4, p.1-5.
- Ileberi, E., Sun, Y., Wang, Z., 2022, A machine learning based credit card fraud detection using the GA algorithm for feature selection, Journal of Big Data, 9(1), 1-17.
- Isabella, S. J., Srinivasan, S., Suseendran, G., 2021, A Framework Using Binary Cross Entropy-Gradient Boost Hybrid Ensemble Classifier for Imbalanced Data Classification, Webology, 18(1).
- Itoo, F., Singh, S., 2021, Comparison and analysis of logistic regression, Naïve BaEvet and KNN machine learning algorithms for credit card fraud detection, International Journal of Information Technology, 13(4), p.1503-1511.
- Janbandhu, R., Begum, S., Ramasubramanian, N., 2020, Credit card fraud detection, In Computing in Engineering and Technology, p. 225-238, Springer, Singapore.

### KAYNAKLAR DİZİNİ (devam)

- Jhangiani, R., Bein, D., Verma, A., 2019, Machine learning pipeline for fraud detection and prevention in e-commerce transactions, In 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), p. 0135-0140, IEEE.
- Kaggle, 2017 a, European Cardholders Dataset, <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>, erişim tarihi: 16.04.2022
- Kaggle, 2017 b, Synthetic Financial Datasets For Fraud Detection, <https://www.kaggle.com/datasets/ealaxi/paysim1>, erişim tarihi: 16.04.2022
- Kaggle, 2019, IEEE-CIS Fraud Detection, <https://www.kaggle.com/c/ieee-fraud-detection>, erişim tarihi: 16.04.2022
- Kaggle, 2020, Credit Card Transactions Fraud Detection Dataset, <https://www.kaggle.com/datasets/kartik2112/fraud-detection?select=fraudTrain.csv>, erişim tarihi: 16.04.2022
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y., 2017, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems, p.30.
- Khare, N., Sait, S.Y., 2018, Credit card fraud detection using machine learning models and collating machine learning models, International Journal of Pure and Applied Mathematics, 118(20), p.825-838.
- Khatri, S., Arora, A., Agrawal, A. P., 2020, Supervised machine learning algorithms for credit card fraud detection: a comparison, In 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), p. 680-683. IEEE.
- Kittidachanan, K., Minsan, W., Pornnopparath, D., Taninpong, P., 2020, Anomaly Detection based on GS-OCSVM Classification, In 2020 12th International Conference on Knowledge and Smart Technology (KST), p.64-69, IEEE.
- Liang, X., Gao, Y., Xu, S., 2022, ASE: Anomaly Scoring Based Ensemble Learning for Imbalanced Datasets, arXiv preprint arXiv:2203.10769.
- Li, P., 2012, Robust logitboost and adaptive base class (abc) logitboost.
- Lin, T.H., Jiang, J.R., 2020, Anomaly Detection with Autoencoder and Random Forest, In 2020 International Computer Symposium (ICS), p.96-99, IEEE.
- Lucas, Yvan, 2019, Credit card fraud detection using machine learning with integration of contextual knowledge.



### KAYNAKLAR DİZİNİ (devam)

- Mathew, J. C., Nithya, B., Vishwanatha, C. R., Shetty, P., Priya, vd., 2022, An Analysis on Fraud Detection in Credit Card Transactions using Machine Learning Techniques, In 2022 Second International Conference on Artificial Intelligence and Smart Energy, p. 265-272.
- Mînaştireanu, E.A., Meşniţă, G., 2020, Methods of handling unbalanced datasets in credit card fraud detection, BRAIN, Broad Research in Artificial Intelligence and Neuroscience, 11(1), p.131-143.
- Mrozek, P., Panneerselvam, J., Bagdasar, O., 2020, Efficient resampling for fraud detection during anonymised credit card transactions with unbalanced datasets, In 2020 ACM 13th International Conference on Utility and Cloud Computing, p.426-433.
- Muter, Z.K., Molood, A.T., 2020, Design The Modified Multi Practical Swarm Optimization To Enhance Fraud Detection, Ibn AL-Haitham Journal For Pure and Applied Sciences, 33(2), p.156-166.
- Nadim, A.H., Sayem, I.M., Mutsuddy, A., Chowdhury, M.S., 2019, Analysis of machine learning techniques for credit card fraud detection, In 2019 International Conference on Machine Learning and Data Engineering (iCMLDE), p. 42-47, IEEE.
- Nguyen, T.T., Tahir, H., Abdelrazek, M., Babar, A., 2020, Deep learning methods for credit card fraud detection, arXiv preprint arXiv:2012.03754.
- Nilson Report, 2021, Annual Fraud Report, [https://nilsonreport.com/content\\_promo.php?id\\_promo=16](https://nilsonreport.com/content_promo.php?id_promo=16), erişim tarihi: 16.04.2022
- Niu, X., Wang, L., Yang, X., 2019, A comparison study of credit card fraud detection: Supervised versus unsupervised, arXiv preprint arXiv:1904.10604.
- Novakovic, J., Markovic, S., 2020, Performance of Support Vector Machine in Imbalanced Data Set, In 2020 19th International Symposium INFOTEH-JAHORINA, p. 1-5.
- Parreno-Centeno, M., Ali, M.A., Guan, Y., Moorsel, A.V., 2019, Unsupervised Machine Learning for Card Payment Fraud Detection, In International Conference on Risks and Security of Internet and Systems, p. 247-262. Springer, Cham.
- Prakash, B., Murthy, G.V.M., Ashok, P., Prithvi, B.P., Kira, S.S.H., 2018, ATM Card Fraud Detection System Using Machine Learning Techniques, International Journal for Research in Applied Science Engineering Technology, 6(4), p.5124-5129.
- Puh, M., Brkić, L., 2019, Detecting credit card fraud using selected machine learning algorithms, In 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), p.1250-1255, IEEE.

### KAYNAKLAR DİZİNİ (devam)

- Pumsirirat, A., Yan, L., 2018, Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine, *International Journal of advanced computer science and applications*, 9(1), p.18-25.
- Riffi, J., Mahraz, M.A., El Yahyaouy, A., 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, p.1-5.
- Rtayli, N., 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, p.102596.
- Rtayli, N., Enneya, N., 2020, Selection features and support vector machine for credit card risk identification, *Procedia Manufacturing*, 46, p.941-948.
- Rtayli, N., Enneya, N., 2019, Credit Card Risk Detection based on Feature-Filter and Fraud Identification, In 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), p.1-8, IEEE.
- Rushin, Gabriel, vd., 2017, Horse race analysis in credit card fraud—deep learning, logistic regression, and Gradient Boosted Tree, 2017 systems and information engineering design symposium (SIEDS), IEEE.
- Saheed, Y.K., Hambali, M.A., Arowolo, M.O., Olasupo, Y.A., 2020, Application of GA feature selection on Naive BaEvet, Random Forest and SVM for credit card fraud detection, In 2020 International Conference on Decision Aid Sciences and Application (DASA), p.1091-1097, IEEE.
- Sailusha, R., Gnaneswar, V., Ramesh, R., Rao, G.R., 2020, Credit card fraud detection using machine learning, In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), p.1264-1270, IEEE.
- Shah, H.B., 2020, Comparing Machine Learning Algorithms For Credit Card Fraud Detection.
- Shamsudin, H., Yusof, U.K., Jayalakshmi, A., 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), p.803-808, IEEE.
- Sharma, N., 2020, Credit Card Fraud Detection Predictive Modeling.
- Shiguihara-Juarez, P., Murrugarra-Llerena, N., 2019, Reducing Dimensionality of Variables for a Classification Problem: Fraud Detection, In 2019 IEEE Sciences and Humanities International Research Conference (SHIRCON), p.1-4, IEEE.

### KAYNAKLAR DİZİNİ (devam)

- Shiguihara-Juarez, P., Murrugarra-Llerena, N., 2019, Reducing Dimensionality of Variables for a Classification Problem: Fraud Detection, In 2019 IEEE Sciences and Humanities International Research Conference (SHIRCON), p.1-4, IEEE.
- Shivanna, A., Ray, S., Alshouiliy, K., 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), p.0268-0273, IEEE.
- Soh, W.W., Yusuf, R.M., 2019, Predicting credit card fraud on a imbalanced data, International Journal of Data Science and Advanced Analytics (ISSN 2563-4429), 1(1), p.12-17.
- Sohony, I., Pratap, R., Nambiar, U., 2018, Ensemble learning for credit card fraud detection, In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, p.289-294.
- Taneja, S., Suri, B., Kothari, C., 2019, Application of balancing techniques with ensemble approach for credit card fraud detection, In 2019 International Conference on Computing, Power and Communication Technologies (GUCON), p.753-758, IEEE.
- Tibshirani, R., 1996, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp.267-288.
- Tingfei, H., Guangquan, C., Kuihua, H., 2020, Using variational auto encoding in credit card fraud detection, IEEE Access, 8, p.149841-149853.
- Tran, T. C., 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), p.1-7, IEEE.
- Trivedi, N.K., Simaiya, S., Lilhore, U.K., Sharma, S.K., 2020, An efficient credit card fraud detection model based on machine learning methods, International Journal of Advanced Science and Technology, 29(5), p.3414-3424.
- Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., Anderla, A., 2019, Credit card fraud detection-machine learning methods. In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), p.1-5, IEEE.
- Wang, J., de Moraes, R.M., Bari, A., 2020, A Predictive Analytics Framework to Anomaly Detection, In 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService), p.104-108, IEEE.

**KAYNAKLAR DİZİNİ (devam)**

- Wibowo, P., Fatichah, C., 2021, An in-depth performance analysis of the oversampling techniques for Yüksek-class imbalanced dataset. Register: Jurnal Ilmiah Teknologi Sistem Informasi, 7(1), p.63-71.
- Wright, Raymond E., 1995, Logistic regression.
- Yang, W., Zhang, Y., Ye, K., Li, L., Xu, C.Z., 2019, June. Ffd: A federated learning based method for credit card fraud detection, In International conference on big data, p.18-32, Springer, Cham.
- Yang, Y., Liu, C., Liu, N., 2019, Credit card fraud detection based on CSat-related AdaBoost, In Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition, p.420-425.
- Yavaş, M., Güran, A., Uysal, M., 2020, Covid-19 Veri kümesinin Smote tabanlı örnekleme yöntemi uygulanarak sınıflandırılması, Avrupa Bilim ve Teknoloji Dergisi, p.258-264.
- Zhang, D., Bhandari, B., Black, D., 2020, Credit Card Fraud Detection Using Weighted Support Vector Machine, Applied Mathematics, 11(12), p.1275.