

**ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

YÜKSEK LİSANS TEZİ

KREDİ KARTI SAHTE İŞLEM TESPİTİ

Kazım SOYLU

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

**ANKARA
2018**

Her hakkı saklıdır

TEZ ONAYI

Kazım SOYLU tarafından hazırlanan “**Kredi Kartı Sahte İşlem Tespiti**” adlı tez çalışması 20/06/2018 tarihinde aşağıdaki jüri tarafından oy birliği ile Ankara Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı’nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.



Danışman: Prof. Dr. Şahin EMRAH

Ankara Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı

Jüri Üyeleri:



Başkan: Doç. Dr. Süleyman TOSUN

Hacettepe Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı



Üye : Prof. Dr. Şahin EMRAH

Ankara Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı



Üye : Dr. Öğr. Üyesi Bülent TUĞRUL

Ankara Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı

Yukarıdaki sonucu onaylarım.

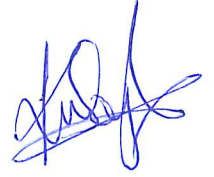
Prof. Dr. Atila YETİŞEMİYEN

Enstitü Müdürü

ETİK

Ankara Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez içindeki bütün bilgilerin doğru ve tam olduğunu, bilgilerin üretilmesi aşamasında bilimsel etiğe uygun davrandığımı, yararlandığım bütün kaynakları atıf yaparak belirttiğimi beyan ederim.

20/06/2018



Kazım SOYLU

ÖZET

Yüksek Lisans Tezi

KREDİ KARTI SAHTE İŞLEM TESPİTİ

Kazım SOYLU

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Şahin EMRAH

Günümüzde kredi kartları ve banka kartlarının kullanımı oldukça yaygınlaşmıştır ve bu kartların internet üzerinden yapılan alışverişlerde kullanımı da artmıştır. Artan bu kart kullanımı ile beraber kredi kartı dolandırıcılığı sorunu da ortaya çıkmıştır. İnsanlar güvensiz ağlarda veya internet sayfalarında alışveriş yaparken kart bilgileri dolandırıcıların eline geçebilmektedir. Hatta mağazalarda yapılan alışverişlerde pos cihazlarına yerleştirilen düzenekler ile kart bilgileri kopyalanabilmektedir. Bunların sonucunda dolandırıcıların kart bilgilerini kullanarak maddi zararlara yol açması sonucu ortaya çıkmaktadır. Bankalar müşterilerin bu dolandırıcılıklardan en az şekilde etkilenmesi için bu sahte işlemlerin tespitine önem vermektedirler. Bankalardan günde çok sayıda işlem yapıldığı için sahte işlemlerin insan gözüyle tespit edilebilmesi çok zordur. Bu yüzden sahte işlemlerin tespit edilmesi için otomasyon sistemleri kullanılmalı ve bu sistemlerin sahte işlemleri tespit etme oranı mümkün olduğunca yüksek; gerçek işlemleri sahte olarak belirlemesi oranı da mümkün olduğunca düşük olmalıdır. Bu amaçla makine öğrenmesi yöntemleri kullanılarak kredi kartı işlemleri sınıflandırılmıştır. Veri kümesi olarak, Eylül 2013'te Avrupalı kart sahiplerinin kredi kartı işlemlerinin bulunduğu veriler kullanılmıştır. Bu veri kümesi üzerinde derin öğrenme, Rastgele Orman ve sınıflandırıcı yığını yöntemleri kullanılmış ve bu yöntemler karşılaştırılmıştır. Bu veri kümesinin dengesiz olması sebebiyle örnekleme yöntemlerinden de faydalanılmıştır. Çalışma sonucunda derin öğrenme ile 0.963, rastgele orman yöntemi ile 0.956, sınıflandırıcı yığını ile 0.979 AUC değeri elde edilmiştir.

Haziran 2018, 48 sayfa

Anahtar Kelimeler: kredi kartı, sahte işlem, dolandırıcılık, makine öğrenmesi, derin öğrenme, rastgele orman, sınıflandırıcı yığını, örnekleme

ABSTRACT

Master Thesis

CREDIT CARD FRAUD DETECTION

Kazım SOYLU

Ankara University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Prof.Dr.Şahin EMRAH

Nowadays, the usage of credit and debit cards increased considerably. Usage of these cards on online shopping also has been increased. With increasing card usage, credit card fraud problem emerged. While people do shopping on unsecured networks and websites, their card information can be stolen by fraudsters. Also when they are in a shop and paying with a physical card, their cards can be copied. After the fraudsters get the card information, they try to use card and cause a financial loss if they are successful. Financial compaines have an effort to prevent these frauds in order to have customer satisfaction and reduce their financial loss. It seems impossible for frauds to be detected with humans since many transactions done from the banks per day. Therefore, automation systems should be used in order to detect fake transactions. Accuracy of detecting frauds should be as high as possible and the rate at which legitimate transactions are classified as frauds should be as low as possible. For the sake of this purpose, credit card transactions are classified into legitimate or fraud using machine learning methods. Deep learning, random forest and stacking ensemble methods were used on the dataset which includes credit card transactions made by European cardholders in September, 2013 and results of these methods were compared. As it is a highly unbalanced dataset, sampling methods were utilized. As a result of this work, we have AUC values of 0.963, 0.956 and 0.979 for the deep learning model, random forest model and the stacking ensemble model, respectively.

June 2018, 48 pages

Key Words: credit card, fraud, fake transaction, machine learning, deep learning, random forest, stacking ensemble, sampling

TEŞEKKÜR

Çalışmalarımı destekleyen, bilgi ve tecrübeleriyle yönlendiren, bu çalışmanın ortaya çıkarılmasında çok fazla emeği olan hocam sayın Prof. Dr. Şahin EMRAH'a (Ankara Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı); çalışmalarım süresince her türlü fedakarlığı esirgemeyen değerli eşim Büşra EMEK SOYLU'ya ve değerli çalışma arkadaşlarıma en derin duygularla teşekkür ederim.

Kazım SOYLU

Ankara, Haziran 2018

İÇİNDEKİLER

TEZ ONAY SAYFASI

ETİK.....	i
ÖZET.....	ii
ABSTRACT	iii
TEŞEKKÜR	iv
KISALTMALAR DİZİNİ	vii
ŞEKİLLER DİZİNİ	viii
ÇİZELGELER DİZİNİ	ix
1. GİRİŞ	1
1.1 Kredi Kartı Sahte İşlem Tespitinde Karşılaşılan Problemler.....	1
1.2 Önerilen Sınıflandırma Yöntemi	2
2. KURAMSAL TEMELLER VE KAYNAK ÖZETLERİ	4
2.1 Makine Öğrenmesi.....	4
2.1.1 Sınıflandırıcı topluluğu.....	5
2.1.2 Gizli Markov Modeli.....	11
2.1.3 Derin öğrenme	12
2.1.4 Yapay bağışıklık sistemi	16
2.2 Örneklem Yöntemleri	17
2.2.1 SMOTE	17
2.3 Performans Ölçümü.....	17
2.4 Özellik Seçme.....	22
2.5 Sahte İşlem Tespiti Üzerine Yapılan Çalışmalar	24
3. MATERYAL VE YÖNTEM.....	28
3.1 Ön İşlemler	28
3.1.1 Özellik seçme	28
3.1.2 Kategorik değerler	29
3.1.3 Numerik işlemler	29
4. ARAŞTIRMA BULGULARI	33
4.1 Veri Kümesi	33
4.2 Uygulama Altyapısı.....	33
4.3 Modellerin Karşılaştırılması	35

5. SONUÇ.....	42
KAYNAKLAR	46
ÖZGEÇMİŞ.....	48



KISALTMALAR DİZİNİ

AUC	Area Under The Curve
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-sampling Technique
DVM	Destek vektör makinesi
SVM	Support vector machine



ŞEKİLLER DİZİNİ

Şekil 2.1 Rastgele orman yöntemi	7
Şekil 2.2 Adaboost algoritması	9
Şekil 2.3 Yığınlama yöntemi	11
Şekil 2.4 Biyolojik nöronun yapısı	12
Şekil 2.5 Yapay nöronun yapısı	13
Şekil 2.6 Klasik yapay sinir ağı mimarisi	15
Şekil 2.7 Derin sinir ağı mimarisi	15
Şekil 2.8 Örnek bir alıcı işletim karakteristiği eğrisi	21
Şekil 2.9 Düşmanca öğrenme oyunu.....	26
Şekil 4.1 Gartner sihirli çeyrekler	34
Şekil 4.2 Derin öğrenme modeli ROC eğrisi	36
Şekil 4.3 Rastgele orman ROC grafiği.....	37
Şekil 4.4 Sınıflandırıcı yığını ROC eğrisi	38
Şekil 4.5 Modellerin AUC bazında karşılaştırması.....	39
Şekil 4.6 Modellerin doğruluk karşılaştırılması.....	40
Şekil 4.7 Modellerin kesinlik karşılaştırması.....	40
Şekil 4.8 Sahte işlem tespiti oranı bazında modellerin kıyaslanması	41
Şekil 4.9 Sahte olarak sınıflandırılan gerçek işlem sayıları	41
Şekil 5.1 Çeşitli yöntemlerin çalışma zamanları.....	43
Şekil 5.2 Sahte işlem tespiti oranı bazında karşılaştırma.....	44
Şekil 5.3 Farklı çalışmaların sonuçlarının kıyaslanması.....	44

ÇİZELGELER DİZİNİ

Çizelge 2.1 Hata matrisi.....	18
Çizelge 2.2 AUC değeri aralıkları.....	21
Çizelge 2.3 Pearson korelasyon katsayısı açıklaması	23
Çizelge 3.1 Özellik seçme işlemi sonucu.....	28
Çizelge 3.2 Derin öğrenme modeli parametreleri	31
Çizelge 3.3 Rastgele orman modeli parametreleri	32
Çizelge 3.4 Sınıflandırıcı yığını parametreleri.....	32
Çizelge 4.1 Bilgisayar konfigürasyonu	35
Çizelge 4.2 Derin öğrenme modeli metrikleri.....	36
Çizelge 4.3 Derin öğrenme modeli hata matrisi.....	36
Çizelge 4.4 Rastgele orman modeli metrikleri.....	37
Çizelge 4.5 Rastgele orman modeli hata matrisi.....	37
Çizelge 4.6 Sınıflandırıcı yığını modeli metrikleri	38
Çizelge 4.7 Sınıflandırıcı yığını hata matrisi	38

1. GİRİŞ

Günümüzde kredi kartları ve debit kartlarının kullanımı oldukça yaygınlaşmıştır. Bankalararası Kart Merkezi (BKM) tarafından yayınlanan istatistiklerde, Türkiye’de Eylül 2017’de 61.251.618 adet kredi kartı; 127.300.550 adet bankamatik kartı mevcuttur. Türkiye’de 2017 yılında internet üzerinden kart kullanılarak toplam 236.503.586 işlem gerçekleştirilmiştir. Kart kullanımının yaygınlaşması beraberinde kredi kartı dolandırıcılığını da getirmektedir. Güvensiz internet sitelerinde kullanılan kredi kartı bilgileri dolandırıcılar tarafından çalınabilmektedir. POS cihazları ve ATM’lerde de kart kopyalama işlemi yapılabilmektedirler. Dolandırıcılar bu tür yöntemlerle elde ettikleri kredi kartı bilgileri ile işlem yapma girişiminde bulunmaktadırlar. Genellikle çok küçük tutarlar için işlem denemektedirler ve başarılı oldukları zaman büyük tutarlarda işlemler yapmaktadırlar.

Kredi kartı sahteciliğini tespit etmek için çeşitli yöntemler geliştirilmiştir. Bu yöntemlerden bazıları kart sahiplerinin alışveriş alışkanlıklarını çıkarmaktadırlar. Dolandırıcıların yaptıkları işlemler kart sahiplerinin alışveriş alışkanlıklarından genellikle farklı olduğu için sahte işlemler tespit edilebilmektedir. Ancak kart sahiplerinin alışveriş alışkanlıkları zamanla değiştiği gibi dolandırıcıların yöntemleri de zamanla değişmektedir.

1.1 Kredi Kartı Sahte İşlem Tespitinde Karşılaşılan Problemler

Kredi kartı işlem veri kümelerinde sahte işlemlerin oranı çok düşüktür. Bu çalışma kapsamında kullanılan veri kümesinde 284.807 işlem bulunmaktadır ancak sadece 492 adedi sahte işlemdir. Bu şekilde bir sınıfın diğerine oranının çok düşük olduğu veri kümelerine dengesiz veri kümesi denir. Dengesiz veri kümelerinde sınıflandırma işlemi dengeli veri kümelerine göre zor bir işlemdir. Klasik sınıflandırma algoritmaları, dengesiz veri kümeleri üzerinde çalıştırıldığında çoğunluğa sahip olan sınıfa doğru eğilim gösterir. Bu da her zaman yüksek doğruluk oranı (accuracy) verir. Örneğin; bu çalışmada kullanılan veri kümesinde çoğunluğa sahip olan sınıfın; yani gerçek

işlemlerin oranı % 99,8'dir. Bu durumda sınıflandırma yaparken bütün işlemleri gerçek olarak sınıflanırsak bile % 99,8 başarıya ulaşılacaktır. Bu sonuç; dengesiz veri kümelerinde doğruluk oranının iyi bir performans ölçüm kriteri olmadığını göstermektedir. Bu yüzden dengesiz veri kümeleri üzerinde yapılan sınıflandırma işlemlerinde ROC eğrisi analizi ve ROC eğrisi altında kalan alan hesabı gibi farklı yöntemler kullanılmaktadır.

Anomali, sahte işlem ve istenmeyen e-posta tespiti gibi dengesiz veri kümeleri üzerinde çalışma gerektiren durumlarda azınlık sınıfa ait verilerin çoğaltılması veya çoğunluk sınıfa ait olan verilerin bir kısmının elenmesi yöntemleri kullanılmaktadır. Alt örnekleme yöntemi ile çoğunluk sınıfa ait veriler elenerek azaltılabilir. Üst örnekleme yöntemi ile de azınlık sınıftaki veriler kullanılarak yeni veriler türetilir. Bu iki yöntemin amacı da veri kümesini dengeli bir hale getirmektir. Dengeli bir hale getirilen veri kümesi ile klasik sınıflandırma algoritmaları çalıştırılabilmektedir.

Karşılaşılan zorluklardan bir diğeri de insanların alışveriş alışkanlıklarının zamanla değişmesidir. Kredi kartı sahiplerinin alışveriş alışkanlıklarının örüntüsü çıkarılıp, sahte işlem tespitinde kullanılabilir ancak bu örüntünün belirli aralıklarla güncellenmesi gerekmektedir. Kart sahiplerinin alışkanlıkları değişeceği gibi kötü niyetli kişilerin de dolandırıcılık yöntemleri değişmektedir. Bu yüzden tasarlanan model, yeni dolandırıcılık yöntemlerini algılayabilmeli ve öğrenebilmelidir. Yani sistem sürekli yeni verilerle eğitilmelidir.

1.2 Önerilen Sınıflandırma Yöntemi

Kredi kartı ve banka kartlarının yaygınlaşması ve internet üzerinden yapılan alışverişlerin artması ile kredi kartı sahteciliği de artmıştır. Bankalara her gün çok sayıda kredi kartı işlemi gelmekte ve bu işlemlerin banka çalışanları tarafından gözle kontrolü çok zor olmaktadır. Dolandırıcıların başarılı olması durumunda bankalar ciddi maddi kayıplar yaşayabilmektedir. Bu yüzden dolandırıcılık tespiti yapabilen bir otomasyona ihtiyaç olmaktadır. Otomasyon ile hedef maddi kaybı minimuma indirmek ve müşteri memnuniyetini de olabildiğince en üst seviyede tutabilmektir. Başarılı olan

her bir sahte işlem bankalar için maddi kayıp; sahte işlem olarak sınıflandırılan her bir gerçek işlem de müşteri memnuniyetini azaltabilecek bir faktör olabilmektedir. Gerçek bir işlemin sahte olarak sınıflandırılması sonucunda bankalar müşterinin kredi kartını kullanıma kapatmakta ve müşterilere telefon yoluyla ulaşarak bilgilendirme yapmaktadırlar. Bu durumun çok yaşanması müşteri memnuniyetinde düşüşe sebep olmaktadır. Bu yüzden kullanılacak otomasyon sisteminin sahte işlemleri tespit etme oranı yüksek, gerçek işlemleri sahte olarak sınıflandırma oranı düşük olmalıdır.

Sınıflandırma makine öğrenmesi yöntemleri ile yapılmıştır. Veri kümesi olarak 2013 yılı Eylül ayında Avrupalı kart sahipleri tarafından yapılan kredi kartı işlemleri kullanılmıştır. Bu veri kümesi üzerinde derin öğrenme ve rastgele orman yöntemleri çalıştırılmış ve karşılaştırılmıştır. Ayrıca bu iki yöntemin bir arada olduğu bir süper öğrenici oluşturulmuş ve sonuçların değişimi gözlemlenmiştir. Bu yöntemlerin uygulanmasında H2O Makine Öğrenmesi Platformu ve Python programlama dili kullanılmıştır.

Kredi kartı işlemleri veri kümesi dengesiz bir küme olduğu için aşırı örnekleme yöntemi kullanılarak dengeli hale getirilmiştir.

2. KURAMSAL TEMELLER VE KAYNAK ÖZETLERİ

Bu bölümde kredi kartları ile yapılan sahte işlemlerin tespiti için bilgisayar bilimleri literatüründe yapılan çalışmalar ve kullanılan metotlar incelenmiştir. Kredi kartı işlem veri kümelerinde gerçek işlem sayısının sahte işlem sayısına oranı çok büyüktür. Bu yüzden kredi kartı işlem veri kümeleri dengesiz kümelerdir. Örneğin; bu tez çalışmasında kullanılan veri kümesinde 284.807 işlem bulunmaktadır ve bu işlemlerin sadece 492 adedi sahte işlemdir. Dengesiz veri kümeleri üzerinde geleneksel sınıflandırma algoritmalarının kullanılabilmesi için veri kümesinde dengeleme yapılması gerekmektedir. Bu amaç için kullanılan yöntemlerden olan alt örnekleme yöntemleri bu bölümde incelenmiştir.

Kredi kartı sahte işlem tespiti üzerine yapılan çalışmalar incelendiğinde yapay sinir ağları, yapay bağışıklık sistemi, Lojistik Regresyon, Gizli Markov Modeli, Rastgele Orman (Random Forest), Karar Ağaçları gibi yöntemlerin kullanıldığı görülmüştür. Dengesiz veri kümeleri üzerinde yapılan sınıflandırma işlemlerinde performans değerlendirmesi, dengeli veri kümelerindekinden farklıdır. AUC (Area Under the Curve), PR (Precision Recall), ROC (Receiver Operating Characteristic) gibi yöntemler dengesiz veri kümelerinin sınıflandırma performanslarının ölçümü için kullanılmaktadır. Bu bölümde bu yöntemler yapılan çalışmalarla birlikte ele alınmıştır.

Sınıflandırma için veri kümesi üzerinde özellik seçme işlemi uygulanabilir. Veri kümesindeki bütün özellikler sonuca ulaşmada etkili değildir. Bazı özelliklerin elenmesi hem sonucu değiştirmeyebilir hem de algoritmanın çalışma zamanının kısılmasına fayda sağlayabilir. Bu çalışmada özellik seçmenin de etkisi incelenmiştir.

2.1 Makine Öğrenmesi

Makine öğrenmesi, bilgisayarları programlayarak girdi olarak verilen verilerden istenilen çıkarımları yapmasını sağlamaktır. Öğrenen sistemler geliştirmek, bu sistemleri iyileştirmek, insanların yerine sınıflandırma ve kümeleme yapabilecek

algoritmalar geliřtirmek makine öğrenmesinin birer amacıdır. Makine öğrenmesi bu amaçları ile birlikte veri madencilięi, büyük veri, istatistik, matematik, doğal dil işleme gibi alanlar ile birlikte yapay zekaya doğru ilerlemektedir. Yapay zeka, bilgisayarların erişeceği nihai hedeftir.

Makine öğrenmesi genel olarak gözetimli ve gözetimsiz olmak üzere iki şekilde yapılmaktadır.

❖ **Gözetimli öğrenme:** Öğrenme için kullanılan verilerde belirli sayıda sınıf ya da etiket bulunuyor ve kayıtların hangi sınıfa ait olduęu belirli ise gözetimli öğrenme yapılmaktadır. Örneęin; makine öğrenmesine giriş seviyesinde sıklıkla kullanılan Iris veri kümesinde üç çiçeęe ait yaprak ölçüleri ve bu ölçülerin hangi çiçeęe ait olduęu belirtilmiştir. Öğrenme algoritmasına ölçüler ve ait olduęu çiçek bilgisi verilerek eğitildięi zaman gözetimli öğrenme yapılmış olmaktadır.

❖ **Gözetimsiz öğrenme:** Bu öğrenme yönteminde verilerin hangi sınıfa ait olduęu bilgisi bulunmamaktadır.

Kredi kartı sahtecilięinin tespit edilmesinde bilgisayarların rolü büyüktür. Bir bankada, operasyonel ekipler tarafından bütün kredi kartı işlemlerinin incelenip sahte işlemlerin bulunması mümkün değildir. Bu yüzden sahte işlemleri tespit eden sistemlere ihtiyaç vardır. Bu sistemler insan faktörünü sıfıra indirmese de büyük oranda azalmaktadır. Literatürde yapılan çalışmalar incelendięi zaman sınıflandırıcı topluluęu (Rastgele Orman, Gradyan Arttırma Makinesi), Gizli Markov Modeli, Lojistik Regresyon ve Derin Öğrenme gibi makine öğrenmesi yöntemleri ile sahtecilięin tespit edilmesine yönelik çalışmalar yapıldıęı görülmüştür.

2.1.1 Sınıflandırıcı topluluęu

Topluluk yöntemi, birçok sınıflandırıcının daha iyi sonuçlar elde etmek için bir arada kullanılmasıdır. Her bir sınıflandırıcı tek başına zayıf öğrenici olarak adlandırılır. Bu

zayıf öğreniciler tek başına iyi sonuçlar vermeyebilir ancak hepsi bir araya geldiği zaman tek bir sınıflandırıcıya göre daha iyi sonuçlar verebilmektedir. Birlikte kullanıldığında ya da sınıflandırıcı topluluğu oluşturmak için üç yöntem bulunmaktadır.

❖ Önyükleyici Birleştirme (Bagging – Bootstrap aggregation)

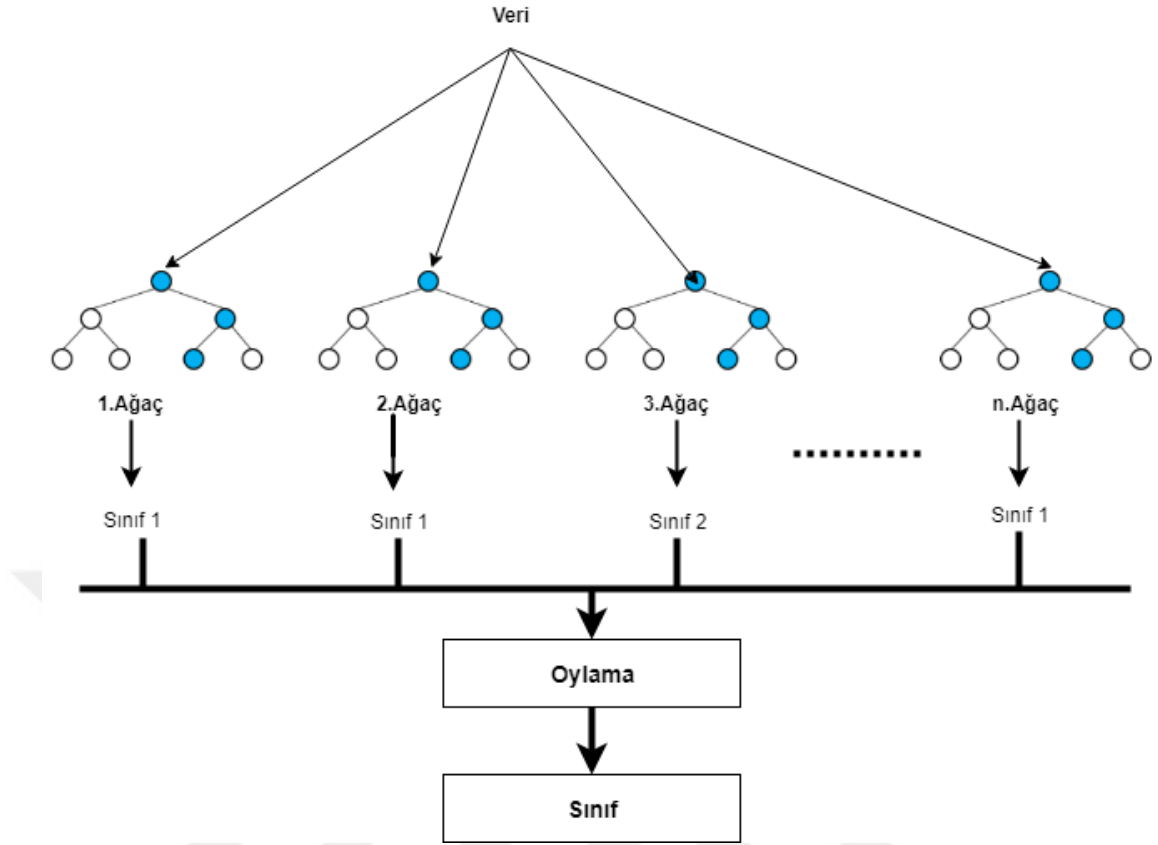
Önyükleyici birleştirme yöntemi 1994 yılında Leo Breiman tarafından ortaya atılmıştır. Amacı varyansı azaltmaktır. Bu yöntemde her bir zayıf öğrenici için eğitim veri kümesinden rastgele alt kümeler oluşturulur ve bu kümeler ile eğitilir. Veri kümesindeki bir kayıt bir alt kümede birden fazla kez yer alabilir. Çünkü rastgele alt küme oluşturulurken alınan rastgele bir kayıt elenmez; tekrar seçim için tutulur. Aynı zamanda aynı kayıt birden fazla alt kümede de yer alabilir.

Test aşamasında her bir sınıflandırıcı test edilen örnek için bir sınıf belirler. Sınıflandırma problemlerinde her bir sınıflandırıcının bulduğu sonuçla oylanarak en çok oyu alan sınıf modelin tahmini olarak kabul edilir. Regresyon problemlerinde ise bu sonuçların ortalaması alınacak sonuca erişilir.

Bu kategoriye giren en popüler algoritmalar Rastgele Orman ve En Yakın k-Komşu yöntemleridir.

• Rastgele Orman (Random Forest)

Rastgele orman yöntemi birçok karar ağacının (Decision Tree) bir araya gelerek oluşturduğu topluluk yöntemidir ve gözetimli öğrenme modelidir. 2001 yılında Leo Breiman tarafından yayınlanmıştır. Rastgele orman, tek bir veri kümesinden rastgele seçilen örnekler üzerinde birçok karar ağacı oluşturup; bu ağaçların yaptığı tahminlerin birleştirilmesidir. Bu tahminleri kullanmada farklı yöntemler bulunmaktadır. Sınıflandırma problemlerinde en çok oyu alan sınıfın seçilmesi; regresyon için ise ortalamalarının alınmasıdır.



Şekil 2.1 Rastgele orman yöntemi

Rastgele orman algoritmasında ağaç sayısı arttıkça daha iyi başarı elde edilmesi beklenmektedir. Ancak ağaç sayısı arttıkça algoritmanın çalışma zamanı da artmaktadır. Rastgele orman yönteminde ağaç sayısı ne kadar artarsa artsın ezberleme (overfitting) durumu ile karşılaşmamaktadır. Algoritma, kategorik değerler ile çalışabilmektedir.

- **En Yakın k-Komşu Algoritması (KNN)**

En yakın k-komşu algoritması makine öğrenmesinde sıklıkla kullanılan bir algoritmadır. KNN parametrik olmayan bir algoritmadır. Klasik sınıflandırma yöntemlerinde öncelikle sınıflandırma modeli eğitilir ve daha sonra test edilir. En yakın k-komşu algoritmasında eğitime işlemi yoktur. Bir örneğin ait olduğu sınıf bulunurken, o örneğin bütün veri kümesindeki en yakın k komşusuna bakılır. Hangi komşuların kullanılacağı kullanılan uzaklık hesabına bağlıdır. En çok kullanılan uzaklık ölçüleri Öklid, Manhattan ve Minkowski uzaklıklarıdır.

$$d_{\text{öklid}} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2.1)$$

$$d_{\text{Manhattan}} = \sum_{i=1}^n |x_i - y_i| \quad (2.2)$$

$$d_{\text{Minkowski}} = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}} \quad (2.3)$$

Uzaklık formüllerinde yer alan x_i ve y_i aralarındaki uzaklık hesaplanan iki örnektir.

❖ Arttırma (Boosting)

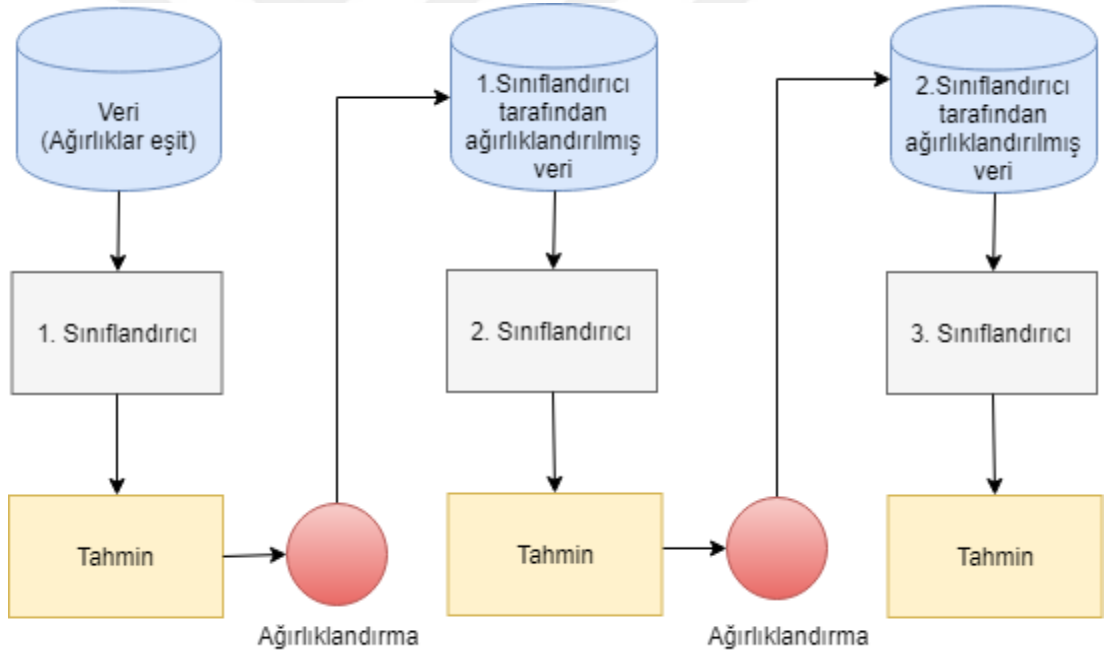
Önyükleyici birleştirme yöntemi ile benzer bir yöntemdir. Zayıf öğreniciler bir araya getirilerek oluşturulmaktadır. Ezberleme (overfitting) ve yetersiz kalma (underfitting) durumlarını önlemek için kullanılır. Çalışma mantığı önyükleyici birleştirme ile benzerdir. Algoritmanın adımları şu şekildedir:

- Eğitim kümesinden rastgele bir küme alınır.
- Seçilen küme ile sınıflandırma modeli eğitilir ve tüm eğitim kümesi ile test edilir.
- Her bir örnek için sınıflandırma hatası hesaplanır. Yanlış sınıflandırma yapılmış ise, o örnek için ağırlık arttırılır.
- Yeni bir rastgele küme oluşturulur.
- Yukarıdaki adımlar yüksek doğruluk elde edilene kadar devam edilir.

Bu kategoriye giren popüler algoritmalar Adaboost ve Gradyan Arttırma Makinesidir.

- **Adaboost**

Adaboost algoritması, arttırma yöntemini kullanan bir birliktelik yöntemidir. 1995 yılında Yoav Freund ve Robert E. Schapire tarafından ortaya atılmıştır. Birçok zayıf öğrenici bir araya gelerek bir sınıflandırıcı topluluğu oluşturmaktadır. Adaboost yönteminin amacı bu sınıflandırıcı topluluğunun başarısını yükseltmektir. Bu yöntemde eğitim kümesi ile eğitim ve tahmin yapılır. Yanlış tahmin edilen veriler için ağırlık arttırılır ve sonraki sınıflandırıcının yeni ağırlıklar ile eğitilmesi sağlanır. Bir sınıflandırıcının, kendisinden önceki sınıflandırıcının hatalı olarak tahmin ettiği veriler üzerine daha fazla dikkat etmesi sağlanır. Kısacası, sınıflandırıcı topluluğundaki bir sınıflandırıcının çıktısı diğer bir sınıflandırıcının girdisi olmaktadır.



Şekil 2.2 Adaboost algoritması

- **Gradyan arttırma makinesi**

Gradyan arttırma makinesi 2001 yılında Jerome H. Friedman tarafından yayınlanmıştır. Yöntemin amacı diğer arttırma yöntemlerinde olduğu gibi sınıflandırma başarısını arttırmaktır. Gradyan arttırma makinesinin temel olarak üç elemanı vardır.

- ✓ Optimize edilecek kayıp fonksiyonu
- ✓ Tahmin yapacak zayıf öğreniciler
- ✓ Kayıp fonksiyonun değerini küçültecek yeni zayıf öğreniciler eklenmesi

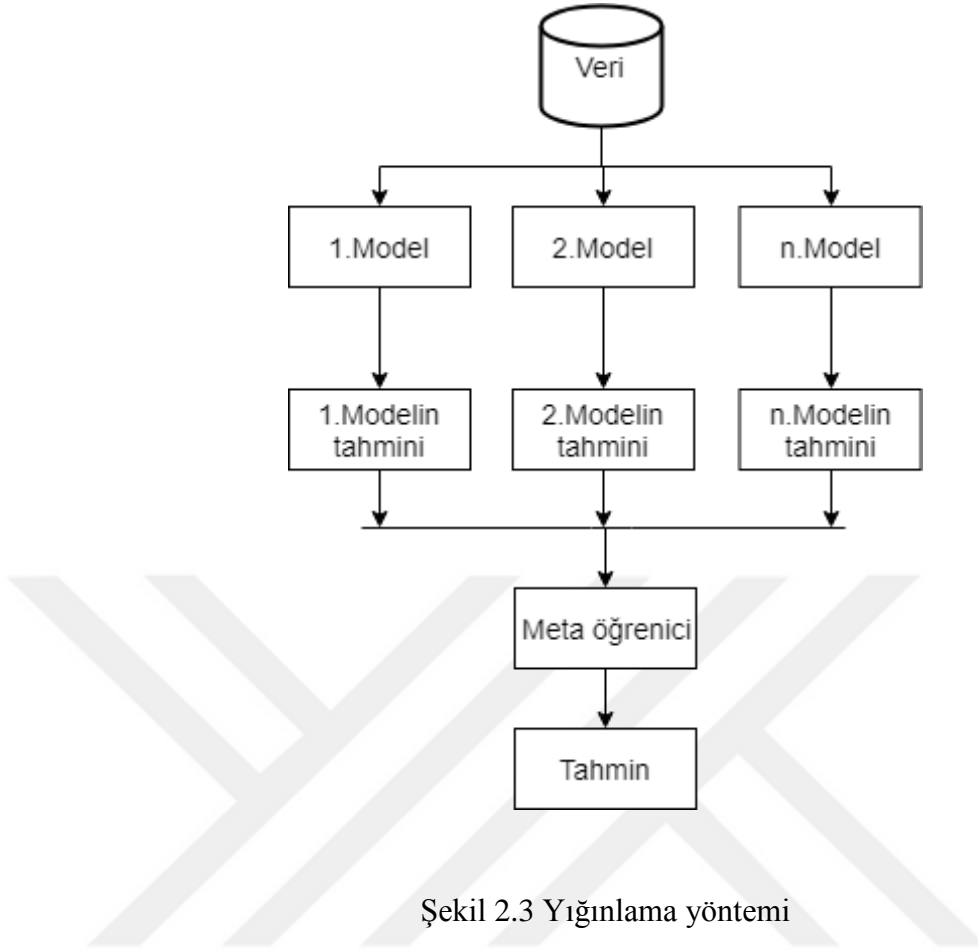
Kayıp fonksiyonu türevi alınabilir bir fonksiyon olmalıdır. Kullanılan bazı standart kayıp fonksiyonları mevcuttur. Sınıflandırma için logaritmik kayıp fonksiyonu; regresyon için ise kare kayıp fonksiyonu kullanılmaktadır. Arttırmada, her adımda elde edilen kayıp fonksiyonu değeri optimize edilir.

Gradyan arttırma makinesi yönteminde zayıf öğreniciler karar ağaçlarıdır. Algoritmanın her iterasyonunda yeni bir karar ağacı eklenir ve önceki karar ağaçlarında değişiklik yapılmaz. Yeni karar ağaçları eklendikçe kaybın azaltılması amaçlanır.

❖ Yığınlama (Stacking)

Önyükleyici birleştirme ve arttırma yöntemleri ile yapılan sınıflandırıcı topluluklarında aynı tipte sınıflandırıcıların birleştirilmesi uygulanmaktadır. Örneğin; rastgele orman yönteminde birçok karar ağacı birleştirilmektedir. Yığınlama yönteminde ise farklı türdeki sınıflandırıcılar bir araya getirilerek bir sınıflandırıcı topluluğu oluşturulmaktadır. Bir araya getirilen her bir sınıflandırıcı temel öğrenici olarak adlandırılmaktadır. Örneğin; derin öğrenme ve rastgele orman sınıflandırıcısı bir arada kullanılarak sınıflandırıcı yığını oluşturulabilir. Bu durumda derin öğrenme ve rastgele orman sınıflandırıcıları aynı ayrı eğitilir. İki sınıflandırıcının sonuçlarının birleştirilmesi oylama veya ortalama alma yöntemi ile değil, bir meta öğrenici ile yapılır. Meta öğrenici de bir sınıflandırma modelidir.

Sınıflandırıcı yığınındaki her bir temel öğrenici ayrı ayrı eğitim verisi ile eğitilmektedir. Temel öğreniciler eğitildikten sonra meta öğrenci temel öğrenicilerin çıktılarıyla eğitilmektedir. Yığının sınıflandırma başarısı, ayrı ayrı her bir modelden daha yüksek olması beklenmektedir.



Şekil 2.3 Yığınlama yöntemi

2.1.2 Gizli Markov Modeli

Markov Modeli, bir durum dizisindeki durumun olma olasılığının sadece bir önceki duruma bağlı olduğunu belirtir. x_n 'inci gözlemin olma olasılığı sadece x_{n-1} 'inci gözleme bağlı ise buna 1.derece Markov zinciri denilmektedir.

$$p(x_n | x_1, x_2, \dots, x_{n-1}) = p(x_n | x_{n-1}) \quad (2.4)$$

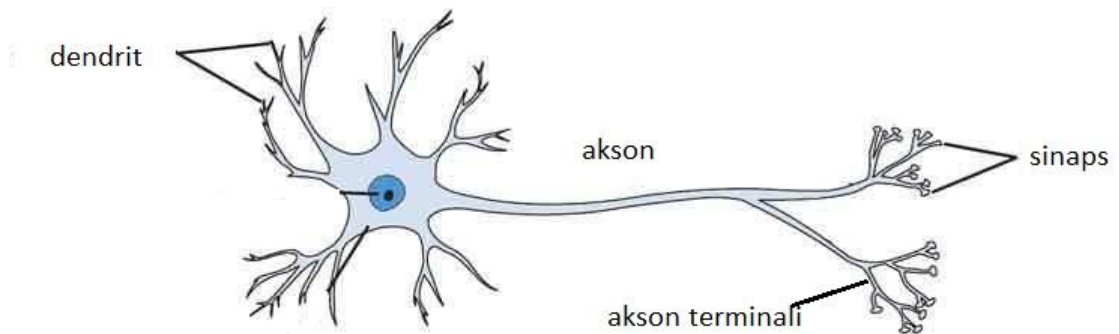
Eğer Markov zincirindeki n 'inci gözlemin olma olasılığı gizli bir değişken tarafından etkileniyorsa buna Gizli Markov Modeli tanımlaması yapılmaktadır. Gizli Markov Modelinde sınırlı sayıda durum ve bu durumlar arası geçiş olasılıkları vardır.

Gizli Markov Modeli ses tanıma, anomali tespiti ve biyoenformatik gibi alanlarda kullanılmaktadır.

2.1.3 Derin öğrenme

Derin öğrenme veya derin sinir ağları, makine öğrenmesinin bir alt dalıdır ve insan beyninin yapısı ve çalışmasından etkilenecek şekilde ortaya atılmıştır. Görüntü işleme, ses tanıma ve doğal dil işlemede yaygın olarak kullanılmaktadır ve başarı oranları yüksektir. Derin öğrenme mimarisinde çok seviyeli gizli katmanlar ve bağlı katmanlar mevcuttur. Derin öğrenme mimarisinin klasik yapay sinir ağlarından en temel farklı çok sayıda katman olmasıdır ve derin sinir ağları olarak adlandırılmaktadır.

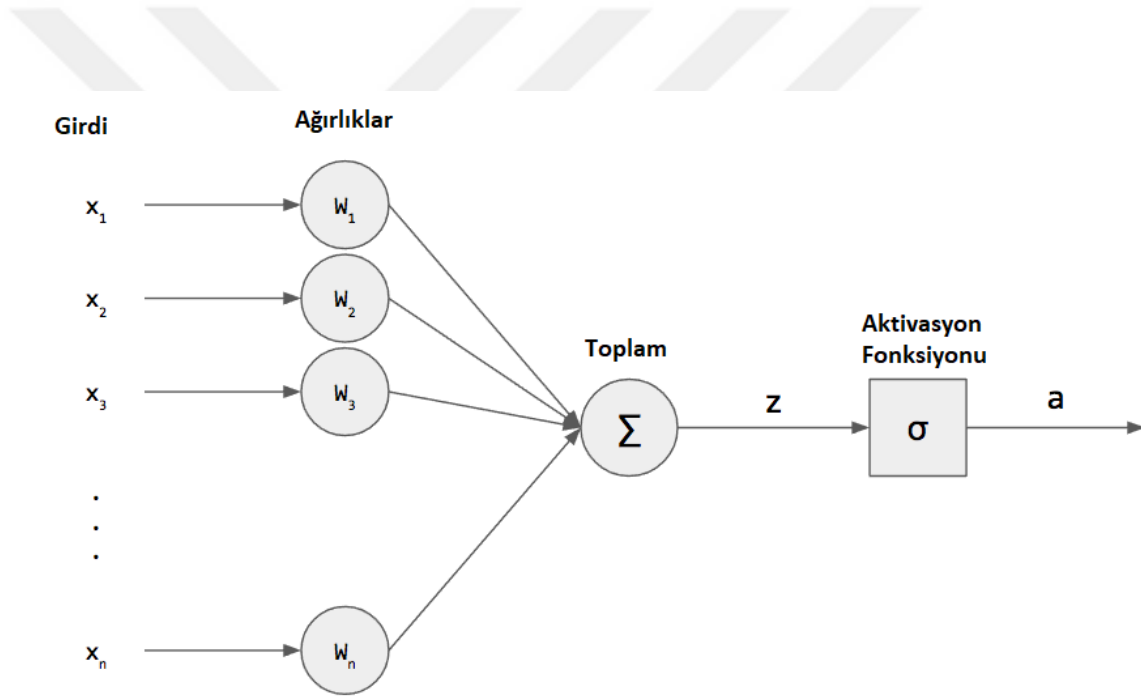
Sinir ağlarında en temel yapı taşı nörondur. Nöronların çalışma mantığında biyolojik sinir sistemlerindeki nöronlardan esinlenilmiştir. Biyolojik nöronlar girdi olarak bir elektrik sinyali alırlar ve yine elektrik sinyali olarak bir çıktı üretirler. Nöronlara girdiler dendritler üzerinden olmaktadır ve bu girdiler belirli bir eşik değeri üzerinde olduğu zaman nöron çıktı üretmektedirler. Üretilen çıktı, akson üzerinden akson terminallerine iletilir. Bir nöron başka bir nörona akson terminalleri ile bağlıdır ve akson terminali ile diğer nöronun dendriti arasında sinaps denilen boşluklar bulunmaktadır.



Şekil 2.4 Biyolojik nöronun yapısı

Yapay sinir ağılarındaki nöronların çalışma mantığı biyolojik nöronlara benzerdir. Bir nörona girdiler dendrit üzerinden gelmektedir. Şekil 2.4'teki yapay nöronun girdileri, biyolojik nöronlardaki akson terminallerinden gelen elektrik sinyallerine karşılık gelmektedir. Ağırlıklar ise dendrit ile akson terminali arasındaki sinaps boşluklarını temsil etmektedir. Toplam, lineer fonksiyondur ve ağırlıklar ile girdilerin çarpımının toplamına eşittir. Toplam değerine, sapma (bias) eklenebilir. Sapma, toplam lineer fonksiyonunun oluşturduğu doğrunun kaydırılmasında kullanılır.

$$z = \sum_{i=1}^n x_i w_i + b \quad (2.5)$$



Şekil 2.5 Yapay nöronun yapısı

❖ **Aktivasyon fonksiyonu:** Girdilerin ağırlıklarla çarpılmasının toplamından elde edilen değer, doğrusal olmayan aktivasyon fonksiyonundan geçirilir ve çıktı üretilir. Üretilen çıktı değeri genellikle 0-1 veya -1 ile 1 aralığı gibi belirli bir aralıkta olmaktadır. Yapay sinir ağlarında yaygın olarak kullanılan bazı aktivasyon fonksiyonları vardır. Sigmoid, tanjant hiperbolik ve ReLu aktivasyon fonksiyonları yaygın olarak kullanılmaktadır. Sigmoid fonksiyonu 0 ile 1 arasında; tanjant hiperbolik -1 ile 1 arasında değerler üretmektedir.

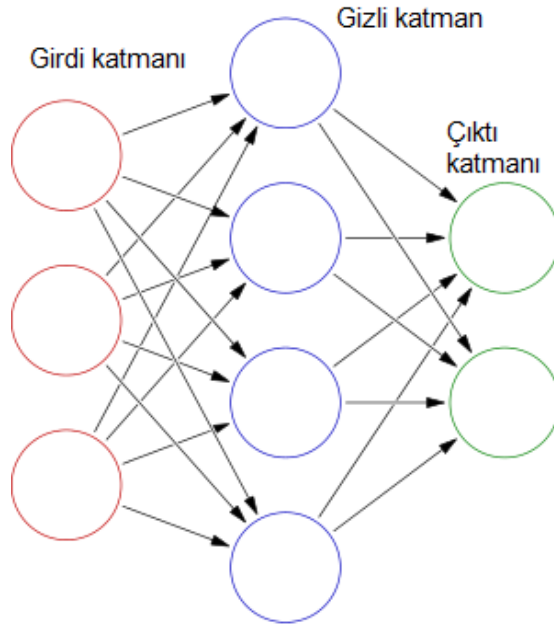
$$\text{sigmoid}(z) = \frac{1}{1 + e^{-z}} \quad (2.6)$$

$$\text{tanh}(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (2.7)$$

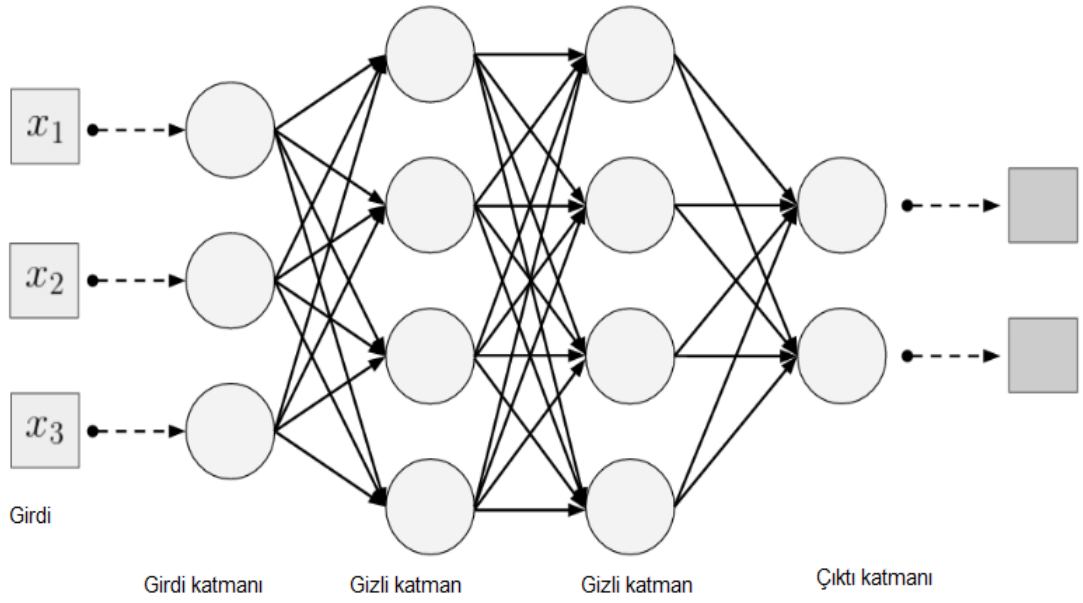
$$(\text{ReLU}) f(z) = \max(0, z) \quad (2.8)$$

Klasik yapay sinir ağlarında girdi katmanı, tek bir gizli katman ve çıktı katmanı vardır (Şekil 2.6). Derin sinir ağlarında çok sayıda gizli katman olabilmektedir. Sınıflandırma yapılacak verilerde çok sayıda özellik mevcut ise katman sayısı veya katmanlardaki nöron sayılarının artırılması gerekmektedir. Çok katmanlı sinir ağları ile yapılan öğrenme işlemine derin öğrenme denilmektedir. Derin sinir ağı mimarisi şekil 2.7’de verilmiştir.

Derin öğrenme ile sınıflandırma yapılırken hem öznitelikler çıkarılır hem de sınıflandırma yapılır. Derin sinir ağı mimarisinin öznitelikleri çıkarma yeteneği mevcuttur. Yüksek başarı elde etmek için çok sayıda veriye ihtiyaç duymaktadır. Bu yüzden çalışma zamanı fazla, sistem gereksinimleri yüksek olmaktadır. Bazı derin öğrenme işlemleri günler, haftalar hatta aylarca sürebilir. Bunun yanında yüksek işlemci gücü ve bellek ihtiyacı olabilmektedir. Hatta işlemcilerin yanında ekran kartları da yüksek işlem gücüne sahip olduğu için derin öğrenmede kullanılabilir.



Şekil 2.6 Klasik yapay sinir ağı mimarisi



Şekil 2.7 Derin sinir ağı mimarisi

Kredi kartı sahte işlem tespitinde derin öğrenme yaygın olarak kullanılan bir yöntem değildir ancak literatürde derin öğrenme ile yapılan çalışmalar mevcuttur. Gupta, sahte işlem tespiti için derin öğrenme mimarisini kullanmış ve bu yöntemi geleneksel makine öğrenmesi yöntemleri ile kıyaslamıştır. Örneklem yöntemleri (SMOTE, ROSE, Undersampling, Oversampling, Hybrid) ve Ensemble yöntemler (Random Forest, Gradient Boosted Model) derin öğrenme ile kıyaslanmıştır. Veri kümesi olarak UCSD-FICO 2009 Data Mining Contest Dataset kullanmıştır. Bu çalışma sonucunda derin öğrenmenin geri çağırma (recall) değeri en yüksek çıkmıştır. Yani derin öğrenme ile yapılan sınıflandırma işleminde sahte işlemlerin tespit edilme oranı en yüksektir. Bu da finansal kaybın en az olduğu anlamına gelmektedir. Kesinlik (precision) değeri, derin öğrenmede diğer makine öğrenmesi algoritmalarına kıyasla daha düşük çıkmıştır. Derin öğrenme, daha fazla gerçek işlemi sahte işlem olarak sınıflandırmıştır. Bu durum müşteri memnuniyetsizliğine yol açan bir durumdur. Gupta'nın çalışmasında aşırı örneklem (oversampling) yöntemi en büyük eğri altında kalan alana (AUC – Area Under the Curve) sahiptir. GBM (Gradient Boosted Model) en iyi kesinlik değerine ulaşmıştır. Topluluk (Ensemble) yöntemleri iyi AUC ve kesinlik değerlerine sahip olmuştur ancak çalışma zamanları oldukça fazladır. Rastgele Orman (Random Forests) ortalama değerlere sahiptir. Müşteri memnuniyetsizliği yani gerçek işlemlerin sahte olarak sınıflandırılması açısından bakıldığında derin öğrenme en kötü durumdadır. Topluluk yöntemleri müşteri memnuniyetsizliği açısından en iyi sonucu vermiştir. Derin öğrenme çapraz doğrulama ile kullanıldığında gerçek işlemlerin sahte olarak sınıflandırılması durumu daha az olmuştur. Finansal kaybın minimum olması açısından bakıldığı zaman derin öğrenmenin en iyi sonucu verdiği gözlemlenmiştir fakat çapraz doğrulama kullanıldığında bu başarı düşmüştür.

2.1.4 Yapay bağışıklık sistemi

Wong ve ekibi, kredi kartı sahte işlemlerinin tespitinde biyolojik bir yöntem olan yapay bağışıklık sistemini (YBS) kullanmışlardır. Yapay bağışıklık sistemi kart sahiplerinin alışveriş davranışını öğrenir. Sisteme giren yeni işlem gerçek ise normal davranış gösterip kabul eder; eğer işlem sahte ise işlemi reddeder. YBS daha önce karşılaşılmayan işlemlerin sınıflandırılması için uygun bir yöntemdir.

2.2 Örneklem Yöntemleri

Bir veri kümesindeki bir kategoriye ait veri sayısı, diğer bir kategoriye göre oldukça fazla ise bu veri kümesi dengesiz bir kümedir. Sınıflandırma algoritmalarını bu tarz veri kümeleri üzerinde çalıştırmak için dengeleme işlemi yapılmaktadır. Dengeleme işlemi sayesinde her bir sınıfa ait veri sayısı birbirine daha yakın hale getirilebilmektedir. Çoğunluğa sahip olan sınıftan alt örneklem yapıp veri azaltarak veya azınlığa sahip olan sınıftaki verilerden yenilerini türeterek bu dengeleme işlemi yapılabilir.

2.2.1 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) bir aşırı örneklem yöntemi. 2002 yılında Nitesh V. Chawla ve ekibi tarafından “SMOTE: Synthetic Minority Over-sampling Technique” başlıklı makale ile yayımlanmıştır. SMOTE algoritması azınlık sınıfındaki verileri kullanarak sentetik veriler üretir. Bu işlemi yaparken k-En Yakın Komşu algoritmasını kullanır. Sentetik numuneler üretirken, bir verinin özellik vektörü ile o verinin en yakın komşusunun özellik vektörü arasındaki farkı bulur ve bu farkı 0 ile 1 arasındaki rastgele bir sayı ile çarpar. Elde edilen sayıyı orijinal verinin özellik vektörüne ekleyerek sentetik bir numune oluşturur.

2.3 Performans Ölçümü

Sınıflandırma problemlerinde genellikle bir veri kümesi üzerinde bir sınıflandırma modeli eğitilir ve daha sonra bu model üzerinde test yapılır. Test işleminin başarı oranını ölçen çeşitli ölçüler mevcuttur.

İki sınıflı bir veri kümesi olan kredi kartı işlemleri kümesinde gerçek işlemler negatif (0), sahte işlemler de pozitif (1) sınıf olarak ele alınmıştır. Bu küme eğitim ve test kümesi olarak iki parçaya ayrılıp sınıflandırma yöntemleri ile sahte ve gerçek işlemler tespit edilmeye çalışılmıştır. Sınıflandırma yönteminin test kümesi üzerinde yaptığı tahminlerin doğruluğu bize yöntemin başarısını vermektedir.

❖ Hata Matrisi (Confusion matrix):

Hata matrisi, bir sınıflandırma algoritmasının performansını gösteren bir tablodur. Tablo, doğru tahmin edilen ve yanlış tahmin edilen veri sayılarını göstermektedir. Makine öğrenmesi modellerinin performanslarının ölçümünde kullanılan doğruluk, hassasiyet, kesinlik, f-ölçüsü, Matthews ilişki katsayısı gibi değerlerin hesaplanmasında hata matrisinden faydalanılmaktadır.

Çizelge 2.1 Hata matrisi

	Negatif (Tahmin)	Pozitif (Tahmin)
Negatif (Gerçek)	DN (TN)	YP (FP)
Pozitif (Gerçek)	YN (FN)	DP (TP)

- **DN (Doğru Negatif – True Negative) :** Veri kümesinde negatif olan bir örneğin sınıflandırma modeli tarafından negatif olarak sınıflandırılmasıdır.
- **DP (Doğru Pozitif – True Positive):** Veri kümesinde pozitif olan bir örneğin sınıflandırma modeli tarafından pozitif olarak sınıflandırılmasıdır.
- **YN (Yanlış Negatif – False Negative):** Veri kümesinde pozitif olan bir örneğin sınıflandırma modeli tarafından negatif olarak sınıflandırılmasıdır.
- **YP (Yanlış Pozitif – False Positive):** Veri kümesinde negatif olan bir örneğin sınıflandırma modeli tarafından negatif olarak sınıflandırılmasıdır.

Kredi kartı sahte işlem tespitinde yanlış pozitif, gerçek bir işlemin sahte olarak sınıflandırılması anlamına gelmektedir. Yanlış negatif ise sahte işlemin gerçek olarak tahmin edilmesidir. Kart sahibinin kendisinin yaptığı gerçek bir işlemin sahte olarak sınıflandırılması durumunda banka kart sahibi ile iletişime geçip teyit etmekte, işlemi iptal etmekte veya kartı kullanıma kapatmaktadır. Bu durumun sürekli yaşanması

müşteri memnuniyetinde düşüşe sebep olabilmektedir. Dolandırıcılar tarafından denenen ve başarılı olan sahte bir işlem ise finansal kayıp demektir.

❖ **Doğruluk (Accuracy)** : Doğru sınıflandırılan örneklerin tüm veri kümesine oranıdır.

$$\text{Doğruluk} = \frac{DP + DN}{DP + DN + YP + YN} \quad (2.9)$$

❖ **Hassasiyet (Sensitivity - Recall)**: Doğru sınıflandırılan pozitif örneklerin, tüm pozitiflere oranıdır.

$$\text{Hassasiyet} = \frac{DP}{DP + YN} \quad (2.10)$$

❖ **Kesinlik (Precision)**: Doğru sınıflandırılan pozitif örneklerin, pozitif olarak sınıflandırılan örneklere oranıdır.

$$\text{Kesinlik} = \frac{DP}{DP + YP} \quad (2.11)$$

❖ **Belirginlik (Specificity)**: Doğru sınıflandırılan negatif örneklerin, tüm negatiflere oranıdır.

$$\text{Belirginlik} = \frac{DN}{DN + YP} \quad (2.12)$$

❖ **Hata oranı (Error rate):** Hatalı sınıflandırma oranıdır.

$$\text{Hata oranı} = 1 - \text{Doğruluk} \quad (2.13)$$

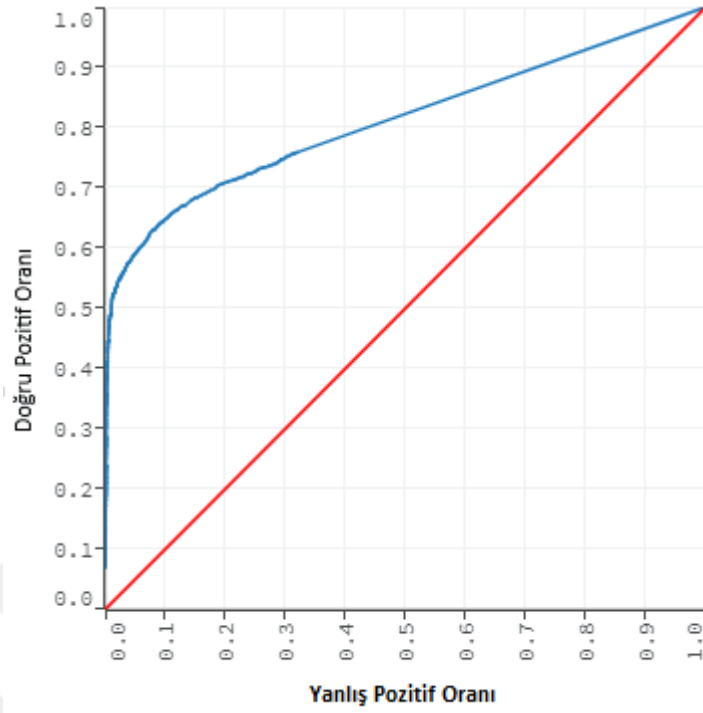
❖ **F-Ölçüsü (F-measure):** F-ölçüsü kesinlik ve hassasiyet değerinin harmonik ortalamasıdır. F-ölçüsünün yüksek olması sınıflandırma yönteminin pozitif sınıf üzerinde iyi sonuç verdiğini göstermektedir.

$$\text{F-ölçüsü} = 2 \times \frac{\text{Kesinlik} \times \text{Hassasiyet}}{\text{Kesinlik} + \text{Hassasiyet}} \quad (2.14)$$

❖ **Matthews İlişki Katsayısı (Matthews Correlation Coefficient):** İkili (binary) sınıflandırmaların kalitesini ölçer. Dengesiz veri kümeleri üzerinde yapılan sınıflandırmalarda iyi bir ölçüdür. Matthews ilişki katsayısı -1 ile 1 arasında değer almaktadır. -1 değeri sınıflandırıcının çok kötü, 1 değeri sınıflandırıcının çok iyi olduğunu göstermektedir. 0 değeri ise rastgele sınıflandırma anlamına gelmektedir.

$$\text{Matthews İlişki Katsayısı} = \frac{DP \times DN - YP \times YN}{\sqrt{(DP + YP)(DP + YN)(DN + YP)(DN + YN)}} \quad (2.15)$$

❖ **Alıcı İşletim Karakteristiği Eğrisi (ROC Curve):** Doğru pozitif oranının, yanlış pozitiflere göre çizilmiş grafiğidir.



Şekil 2.8 Örnek bir alıcı işletim karakteristiği eğrisi

❖ **Alıcı İşletim Karakteristiği Eğrisi altında kalan alan (AUC) :** ROC eğrisi altında kalan alan, bu eğriyi yorumlamak için bir ölçüdür. Dengesiz veri kümeleri üzerinde yapılan sınıflandırmalarda AUC değeri önemli bir performans ölçөгüdür. Mohamed Bekkar vd. tarafından yapılan çalışmada AUC değerinin aralıkları ve yorumlanması belirtilmiştir (Çizelge 2.2).

Çizelge 2.2 AUC değeri aralıkları

AUC Değeri	Performans
0.5-0.6	Kötü
0.6-0.7	Orta
0.7-0.8	İyi
0.8-0.9	Çok iyi
0.9-1.0	Mükemmel

2.4 Özellik Seçme

Bir veri kümesi üzerinde sınıflandırma yapmadan önce bazı ön işlemler yapılabilir. Ön işlemler sınıflandırmanın daha etkili veya daha hızlı çalışması için ya da verinin sınıflandırma modeline uygun hale getirilmesi için yardımcı olmaktadır. Özellik seçme işlemi veri kümesi özelliklerine uygulanan ön işlemlerden bir tanesidir. Sınıflandırma işleminde bütün özellikler sonuca ulaşmada aynı derecede katkı sağlamazlar, hatta bazı özellikler etkisizdir. Bu yüzden bu özelliklerin veri kümesinden çıkarılması çalışma zamanını kısıltacaktır.

Veri kümeleri resim veya ses dosyalarından oluşacağı gibi metin dosyalarından da oluşabilir. Metin şeklinde sunulan veri kümesindeki her bir satır, kayıt ya da örnek; her bir sütun özellik ya da nitelik olarak isimlendirilmektedir. Makine öğrenmesi modelleri nitelikleri kullanarak çıkarımlar yapmaktadır. Veri kümesindeki tüm nitelikler bu çıkarımları yapmak için gerekli değildir ya da veri kümesinden o niteliklerin silinmesi sonucu değiştirmez. Nitelik sayısı ne kadar fazla olursa, modelin çalışma zamanı o kadar uzun olacaktır. Niteliklerin elenmesi sınıflandırıcının başarısına da olumlu katkıda bulunabilmektedir. Kısacası özellik seçme, veriyi iyi bir şekilde temsil eden özellikler alt kümesi bulma işlemidir.

Veri kümelerindeki sınıf değişkeni ile herhangi bir ilişki içerisinde olmayan yani korelasyon göstermeyen özellikler, sınıflandırma işleminde sapmaya sebep olabilmektedir. Bu sapma sınıflandırma işleminin başarısının düşmesine sebep olmaktadır.

Özellik seçme için uygulanan üç yöntem bulunmaktadır. Bu yöntemler; filtreleme, sarmalayıcı yöntemler ve gömülü yöntemlerdir.

❖ **Filtreleme:** Filtreleme yöntemi her özelliğe bir puan verilerek, puanı belirli bir eşik değerinin altında kalan özelliklerin elenmesidir. Puan hesaplamasında bir özelliğin diğer

özelliklerle ve sınıf değişkeni ile ilişkisi (korelasyon) veya karşılıklı bilgi (mutual information) kullanılabilir.

Bir özelliğin korelasyon puanının hesaplanmasında Pearson korelasyon katsayısı kullanılabilir.

$$R(i) = \frac{kov(n_i + S)}{\sqrt{var(n_i) * var(S)}} \quad (2.16)$$

Çizelge 2.3 Pearson korelasyon katsayısı açıklaması

Değişken	Açıklama
n_i	veri kümesindeki i. özellik
S	veri kümesindeki sınıf değişkeni
kov	kovaryans
var	varyans

❖ **Sarmalayıcı yöntemler:** Bu yöntemde özellikler kümesinin farklı alt kümeleri ile sınıflandırma modelinin performansı test edilir ve en iyi sonucu veren alt küme bulunur. n adet özelliğin 2^n alt kümesi bulunmaktadır. Özellik sayısı arttıkça alt küme sayısı asimptotik olarak arttığı için, bütün alt kümeleri denemek NP-zor problem haline gelmektedir. Bu yüzden, yaklaşık en iyi alt kümeyi bulmak için buluşsal (heuristic) yöntemler kullanılmaktadır.

❖ **Gömülü yöntemler:** Bu yöntemde özellik seçme işlemi, eğitim işleminin bir parçası olarak yer almaktadır. Sarmalayıcı yöntemlere göre daha basittir. Özellik seçme eğitim işleminin bir parçası olduğu için, elde edilen özellik alt kümesi eğitilen modele özel olmaktadır. Bir model için uygun olan alt küme, başka bir model için uygun olmayabilmektedir.

2.5 Sahte İşlem Tespiti Üzerine Yapılan Çalışmalar

Quah ve Sriganesh, (2008) kredi kartı sahte işlemlerin tespitinde harcama alışkanlıklarını dikkate alan yeni bir yöntem geliştirmişlerdir. Kendini organize eden harita (KOH - self-organizing map) ile müşteri davranışını analiz etmişlerdir. Kendini organize eden haritalar gözetimsiz öğrenme yapan sinir ağlarıdır ve bu ağın nöronları veriye göre kendini organize etmektedir. Çalışmada öncelikle veriler üzerinde bazı ön işlemler uygulanmıştır. Sayısal olmayan özellikler sayısal hale getirilmiştir. KOH modeli eğitildiğinde, kredi kartı işlemleri sahte ve gerçek olarak sınıflandırılmış olmaktadır. KOH modelinin katmanlarının birçok amacı vardır. Bu amaçlar verinin sınıflandırılması ve kümelenmesi; verideki gizli örüntüleri tespit etmek ve diğer katmanlar için filtreleme mekanizması olarak görev yapmaktır. KOH yönteminin sınıflandırma ve kümeleme yeteneği, onu bu amaçlara ulaşmak için uygun bir yapay sinir ağı yapmaktadır. KOH'un çıktısı input vektörlerinden elde edilen örüntü ve kümelerdir. Geçmiş kredi kartı işlemleri ile eğitilen model, yeni işlemleri gerçek zamanlı olarak işlemektedir. Çalışmada 5 müşteri ve her birine ait 105 işlem ile KOH modeli test edilmiştir. Düğüm sayısı 70x70, iterasyon sayısı 200 olarak verilmiştir. Benzerlik metriği olarak Öklid uzaklığı kullanılmıştır.

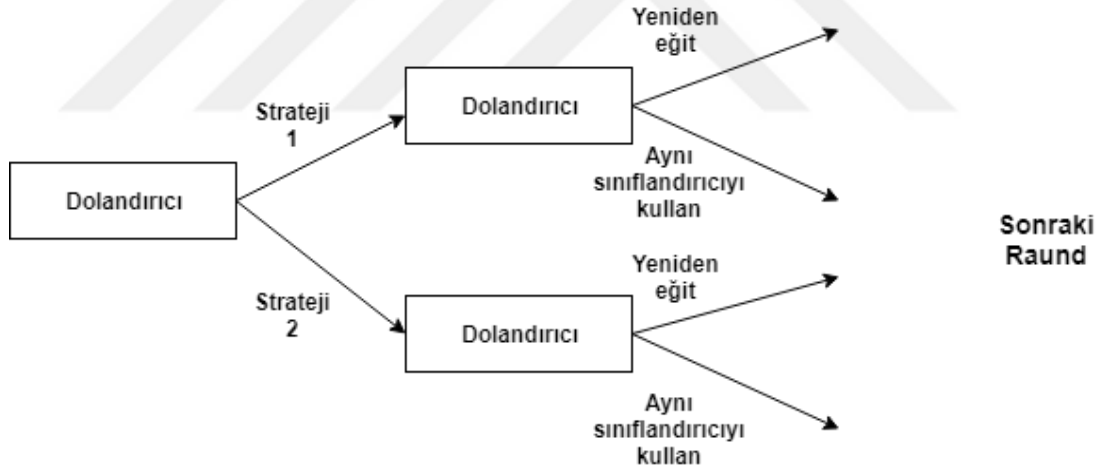
Pozzolo vd. (2015) tarafından yapılan çalışmada Belçika'da bir ödeme servisi sağlayıcısı tarafından sağlanan veri kümesi üzerinde çalışılmıştır. Bu çalışmada bir kaç makine öğrenmesi algoritması ve teknik kullanılmış; bu yöntemlerin sonuçları karşılaştırılmıştır. Ayrıca çalışmada veri kümesi dengeleme yöntemleri de incelenmiştir. Çalışmada tek bir işlem bilgisinin yeterli olmayacağı, bazı kümülatif değerlerin (günlük toplam harcama, haftalık işlem sayısı vb) özellik olarak eklenmesi gerektiği vurgulanmıştır. Rastgele orman, yapay sinir ağları, Destek Vektör Makinesi sınıflandırma yöntemleri ve alt örnekleme, SMOTE, EasyEnsemble örnekleme yöntemleri kullanılmıştır. Eğitilen modelin güncellenmesinin etkisi de incelenmiştir. Bir kez, her gün, 15 günde bir ve haftalık olarak model güncellemesi yapılarak sonuca olan etkisi gözlemlenmiştir. Her gün yapılan güncelleme işleminin en iyi katkıyı sağladığı görülmüştür. Bu güncellemede önceki 90 günün verileri ile her gün model yeniden

eğitilmektedir. Çalışmada en iyi başarı EasyEnsemble yöntemi ve geçmiş gerçek işlemlerin elenmesi yöntemi ile elde edilmiştir.

Zareapoor ve Shamsolmoali (2015) yaptıkları çalışmada kredi kartı sahte işlem tespitinde önyükleyici birleştirme topluluk sınıflandırıcısını kullanmışlardır. Bu çalışmada, literatürdeki sahte işlem tespiti çalışmalarının yetersizliğinden ve araştırmacılar için kredi kartı veri kümelerinin zor bulunmasından bahsetmişlerdir. Ayrıca bulunan veri kümelerinin oldukça dengesiz olması ve boyutlarının çok büyük olması; bu yüzden de yüksek işlemci gücü gerektirdiği vurgulanmıştır. Çalışmada sahte işlem tespiti tekniklerinin performanslarının ölçümünde genellikle yanlış pozitif ve yanlış negatif metriklerinin kullanıldığı belirtilmiş; veri kümeleri oldukça dengesiz olduğu için doğruluk değerinin sahte işlem tespitinde uygun bir metrik olmadığı vurgulanmıştır. Zareapoor ve Shamsolmoali, yanlış sınıflandırılan sahte işlemlerin maliyetinin yanlış sınıflandırılan gerçek işlemlerin maliyetinden daha fazla olduğunu; bu yüzden hem kesinlik hem de hassasiyet metriklerinin dikkate alınması gerektiğini belirtmişlerdir. Bu çalışmada Destek Vektör Makinesi (SVM), en yakın k-komşu (KNN), Naive Bayes ve önyükleyici birleştirme topluluk sınıflandırıcısı yöntemleri kullanılmıştır. Veri kümesi olarak UCSD-FICO Data Mining Contest 2009 veri kümesi kullanılmıştır. Veri kümesi, % 20, % 15, % 10 ve % 3 sahte işlem içerecek şekilde 4 farklı şekilde bölünerek kullanılmıştır. Ön yükleyici birleştirme yöntemi hem sahte işlemleri tespit etme oranında hem de gerçek işlemleri sahte olarak sınıflandırma oranında tüm veri kümesi bölümlerinde en iyi sonucu vermiştir.

Pandey'in 2017 yılında yaptığı çalışmasında kredi kartı sahte işlem tespiti için derin öğrenme yöntemini uygulamıştır. Pandey bu çalışmasını UCSD-FICO Data Mining Contest 2009 veri kümesi üzerinde gerçekleştirmiştir. Veri kümesini % 60'ı eğitim; % 20'si doğrulama ve % 20'si test için kullanılmak üzere üç parçaya bölmüştür. Pandey, oluşturduğu derin öğrenme modeli 200 nöronlu iki gizli katmandan oluşmakta ve aktivasyon fonksiyonu olarak Rectifier kullanmaktadır. Epok sayısı olarak 1 verilmiştir. Pandey'in bu çalışmasında doğrulama kümesi üzerindeki sonuçlar verilmiştir. Test kümesi ile yapılan tahminlerin sonucu bulunmamaktadır. Doğrulama esnasında MSE değeri 0.01661334; RMSE 0.1288928 olarak belirtilmiştir.

Zeager vd. (2017) tarafından adaptif sahte işlem tespiti çalışması yapılmıştır. Çalışmanın amacı dolandırıcıların davranışlarını sisteme dahil etmektir. Sahte işlem tespiti için sınıflandırma yöntemi olarak lojistik regresyon kullanılmıştır. Dolandırıcıların davranışlarını modellemek için oyun teorisi ile düşmanca öğrenme kullanılmıştır. Çalışmada bu yöntemin daha önce sahte işlem tespiti ve istenmeyen e-posta verileri için kullanıldığı belirtilmiştir ancak gerçek işlem verileriyle kullanılıp geliştirilmesi bu çalışma kapsamında yapılmıştır. Veri kümesi dengeleme için SMOTE yöntemi kullanılmıştır. Düşmanca öğrenme, makine öğrenmesinde kullanılan bir kavram olup, düşman ve rakibi arasındaki ilişkiler üzerine yoğunlaşır. Çalışma kapsamında bir finans kurumu tarafından sağlanan; bir yıla ait 86 milyon veri içeren kredi kartı işlemleri veri kümesi kullanılmıştır. Zeager vd. düşmanların sürekli olarak rakiplerinin davranışlarını öğrendiğini ve kendilerini adapte ettiklerini öne sürdükleri hipotezlerini test etmek için bir algoritma geliştirmişlerdir. Bu algoritma tekrar eden bir oyun şeklindedir, (Şekil 2.9).



Şekil 2.9 Düşmanca öğrenme oyunu

Sahte işlem algoritmalarında dolandırıcıların hangi stratejileri kullandığı bilinmemektedir. Bu yüzden modelin tekrar eğitilip eğitilmeyeceğinin kararı verilirken bu bilgi göz ardı edilir.

Zeager vd. yaptıkları çalışmada 10 raunttan oluşan düşmanca öğrenme oyunu ile test yapmışlardır. Dengesiz veri kümesi SMOTE ile daha dengeli hale getirilmiştir. SMOTE

ile veri kümesinin % 15'inin sahte işlem den oluşacağı şekilde üst örnekleme yapılmıştır. Oyunun ilk raundunda AUC değeri 0.78 iken son rauntta 0.84 olmuştur. Bu sonuç ekibin hipotezini destekler yöndedir. Aynı çalışmada 10 rauntluk bir oyun ile yeniden eğitim yapmadan test yapılmıştır. Bu kez, ilk rauntta 0.78 olan AUC değeri son rauntta 0.76'ya gerilemiştir.

Srivastava ve ekibi, Gizli Markov Model yöntemini kullanarak kart sahiplerinin normal alışveriş alışkanlıklarını modellemişlerdir. Tasarlanan bu model tarafından reddedilen işlemlerin büyük olasılıkla sahte olduğu kabul edilmiştir. Aynı zamanda bu modelin gerçek işlemleri reddetmemesi amaçlanmıştır.

Carneiro N. vd. (2017) kredi kartı sahte işlem tespiti çalışmalarında lojistik regresyon, rastgele orman ve destek vektör makineleri yöntemlerini kullanmışlardır. Çalışma ortak çalışılan bir şirketin verileri üzerinde gerçekleştirilmiştir. Veri kümesindeki kategorik değerler numerik değerlere dönüştürülmüş; kayıp değerler yerine 0 verilmiştir. Farklı birimlerde olan özellikler 0-1 aralığına ölçeklenerek standart bir hale getirilmesi sağlanmıştır. Çapraz doğrulama sonucunda rastgele orman sınıflandırıcısı ile 0.935, destek vektör makinesi ile 0.906 ve lojistik regresyon ile 0.907 AUC değeri elde edilmiştir. Rastgele orman modelinde 1500 ağaç kullanılmıştır. Test kümesi üzerinde gerçekleştirilen tahmin işleminde ise rastgele orman modeli 0.88 AUC değerine sahip olmuştur.

3. MATERYAL VE YÖNTEM

3.1 Ön İşlemler

3.1.1 Özellik seçme

Veri kümesinde bulunan alanların tamamı sınıflandırma için gerekli değildir. Bazı alanların elenmesi çalışma zamanını kısaltacaktır. Ayrıca ezberlemeyi önleme, bir yöne sapmayı engelleme ve tahmin performansını arttırma gibi etkilere sahiptir. Bu çalışmada WEKA veri madenciliği ve makine öğrenmesi programı ile kredi kartı işlemleri veri kümesi üzerinde özellik seçme işlemi uygulanmıştır. WEKA’da bulunan CfsSubsetEval özellik değerlendiricisi ve BestFirst arama yöntemi kullanılarak özellikler seçilmiştir. CfsSubsetEval, korelasyon tabanlı özellik seçme yöntemidir. Sınıf değişkeni ile korelasyonu yüksek fakat birbirleri ile korelasyonu düşük alt küme seçilir.

Çizelge 3.1’de WEKA ile uygulanan özellik seçme işlemi sonucu verilmiştir. Özellik seçme çalıştırılırken 10-katlamalı çapraz doğrulama kullanılmıştır. Bu sonuca göre Time, V22, V23 ve Amount özellikleri veri kümesinden silinmiştir.

Çizelge 3.1 Özellik seçme işlemi sonucu

Özellik	Katlama sayısı
Time	0(0 %)
V1	10(100 %)
V2	10(100 %)
V3	10(100 %)
V4	10(100 %)
V5	10(100 %)
V6	10(100 %)
V7	10(100 %)
V8	10(100 %)
V9	10(100 %)
V10	10(100 %)
V11	10(100 %)
V12	10(100 %)
V13	10(100 %)

Çizelge 3.2 Özellik seçme işlemi sonucu (devamı)

V14	10(100 %)
V15	10(100 %)
V16	10(100 %)
V17	10(100 %)
V18	10(100 %)
V19	10(100 %)
V20	9(90 %)
V21	10(100 %)
V22	2(20 %)
V23	0(0 %)
V24	10(100 %)
V25	6(60 %)
V26	10(100 %)
V27	9(90 %)
V28	5(50 %)
Amount	0(0 %)

3.1.2 Kategorik değerler

Kategorik özellikler, sınırlı sayıda elemana sahip olan, bir kategoriye ait olan alanlardır. Makine öğrenmesi yöntemlerinin bir kısmı kategorik değerler ile çalışmayı desteklememektedir. Bu yüzden bu alanların sayısal değerlere dönüştürülmesine ihtiyaç bulunmaktadır. Rastgele orman ve derin öğrenme modellerinin eğitilmesi sırasında kategorik değerlerin etiket kodlama (Label Encoding) yöntemi ile kodlanması sağlanmıştır. Kategorik değer kodlama yöntemi bir parametre olarak verilmektedir.

3.1.3 Numerik işlemler

Veri kümesindeki sayısal alanlar üzerinde bazı ön işlemler yapılabilmektedir. Sayısal alanların belirli bir aralığa ölçeklenmesi, yuvarlama, standartlaştırma gibi ön işlemler makine öğrenmesi problemlerinde kullanılan yöntemlerdir.

Sayısal değerler bir alt sınır ve üst sınır arasına ölçeklenebilir. Birçok makine öğrenmesi algoritması sayısal özelliklerin 0 ile 1 veya -1 ile 1 arasına ölçeklenmesini

beklemektedir. Ölçekleme, o özelliğin minimum ve maksimum değerlerini kullanarak belirlenen aralığa getirilmeyi sağlar. Formül 3.2’de x_i özellik; x'_i ise ölçeklenmiş özelliktir.

$$x'_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (3.2)$$

Nümerik alanlar, ölçeklemeye alternatif olarak ortalaması 0, standart sapması 1 olacak şekilde dönüştürülebilir.

$$x_i = \frac{x_i - \bar{x}}{\sigma} \quad (3.3)$$

3.2 Derin Öğrenme ile Sınıflandırma

Kredi kartı işlemlerinin sınıflandırılmasında H2O platformu üzerinde derin öğrenme ile sınıflandırma modeli oluşturulmuştur. Veri kümesi, % 70’i eğitim % 30’u test için kullanılmak üzere iki parçaya ayrılmıştır. Oluşturulan derin öğrenme modeli eğitim kümesiyle eğitilmiş; eğilirken çapraz doğrulama kullanılmıştır. Eğitimi tamamlanan modelin başarısı test kümesi ile sınanmış; AUC değeri ile sahte işlem tespit etme oranı hesaplanmıştır.

3.2.1 Geliştirilen model

Kullanılan sinir ağı mimarisi ileri beslemeli (feedforward) sinir ağıdır. Çizelge 3.1’de derin öğrenme modelinin parametreleri verilmiştir. Model, her biri 500 nörona sahip olan 2 tane gizli katmandan oluşmaktadır. Aktivasyon fonksiyonu olarak Rectifier kullanılmıştır. Veri kümesi %70’i eğitim; %30’u test için kullanılmak üzere iki parçaya ayrılmıştır. Eğitim aşamasında 10-katlamalı çapraz doğrulama kullanılmıştır.

Çizelge 3.3 Derin öğrenme modeli parametreleri

Hiper Parametre	Değer
Epok sayısı	100
Saklı katmanlar	[500,500]
Aktivasyon fonksiyonu	Rectifier
Katlama sayısı	10
Örnekleme çarpanı	[1,500]
Durdurma metriği	Logloss
Durdurma raund sayısı	10
Durdurma toleransı	0,001
Kategorik kodlama	LabelEncoder
L1	1e-6
L2	1e-6
Ağırlık Dağıtımı	Bernoulli
Yığın sayısı	10

3.3 Rastgele Orman ile Sınıflandırma

Rastgele orman yöntemi kullanılarak kredi kartı işlemlerinin sınıflandırılması sağlanmıştır. Veri kümesi % 70'i eğitim % 30'u test için kullanılmak üzere iki parçaya ayrılmıştır. Tasarlanan rastgele orman modeli eğitim kümesiyle eğitilmiştir. Eğitim aşamasında çapraz doğrulama kullanılmıştır. Eğitimi tamamlanan modelin başarısı test kümesi ile sınanmış; AUC değeri ile sahte işlem tespit etme oranı hesaplanmıştır.

3.3.1 Geliştirilen model

Model, 300 ağaçtan oluşmaktadır. Modelin eğitilmesi sırasındaki doğrulama için 10-katlamalı çapraz doğrulama yöntemi kullanılmıştır.

Çizelge 3.4 Rastgele orman modeli parametreleri

Hiper parametre	Değer
Ağaç sayısı	300
Katlama sayısı	10
Derinlik	20
Örnekleme oranı	0.9
Durdurma metriği	Logloss
Durdurma raund sayısı	3
Durdurma toleransı	0,001

3.4 Sınıflandırıcı Yığını

Sınıflandırıcı topluluğu yöntemlerinden birisi olan yığınlama (stacking) yöntemi ile yeni bir model oluşturulmuş ve sınıflandırma yapılmıştır. Ayrı ayrı kullanılan derin öğrenme ve rastgele orman modelleri sınıflandırıcı yığnında birleştirilerek tek bir model gibi kullanılmıştır. Sınıflandırıcı yığnında meta öğrenici rastgele orman modeli kullanılmıştır.

3.4.1 Geliştirilen Model

Çizelge 3.5 Sınıflandırıcı yığını parametreleri

Hiper parametre	Değer
Taban modeller	Derin öğrenme, rastgele orman
Derin öğrenme parametreleri	Çizelge 3.2
Rastgele orman parametreleri	Çizelge 3.3
Meta öğrenici	Rastgele orman

4. ARAŞTIRMA BULGULARI

4.1 Veri Kümesi

Bu çalışmada kullanılan veri kümesi Eylül 2013'te Avrupa'da iki günde yapılan kredi kartı işlemlerinden oluşmaktadır. Veriler Özgür Brüksel Üniversitesi'nde makine öğrenmesi grubunun çalışmaları sırasında toplanmıştır. Veri kümesinde 284.807 işlem bulunmakta ve bu işlemlerin 492 adedi sahte işlem olarak yer almaktadır. Sahte işlemlerin sayısının tüm veri kümesine oranı 0.172'dir. Çok az sahte işlem bulunduğundan dolayı veri kümesi oldukça dengesiz bir haldedir.

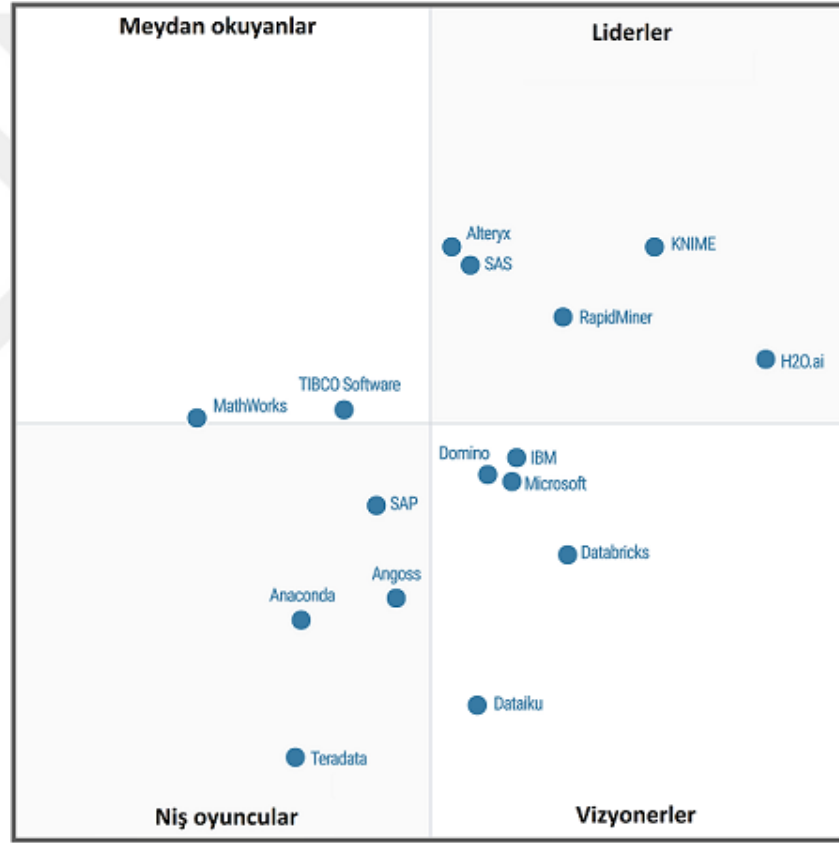
Veri kümesi kredi kartı işlemleri içerdiği için bilgilerin gizliliğinden dolayı PCA dönüşümü yapılmış halde sunulmaktadır. Bir kredi kartı işlemi, 31 alandan oluşmaktadır ve "Time" ile "Amount" dışındaki tüm alanlar PCA ile sayısal değerlere dönüştürülmüştür. "Class" alanında işlemin sahte mi gerçek mi olduğu bilgisi yer almaktadır. 0 gerçek işlemi, 1 ise sahte işlemi göstermektedir.

4.2 Uygulama Altyapısı

Bu çalışmada makine öğrenmesi algoritmalarının uygulanması için H2O Makine Öğrenmesi Platformu kullanılmıştır. Bu platform dünyadaki lider araştırma ve danışmanlık firmalarından olan Gartner'ın sunduğu veri madenciliği ve makine öğrenmesi platformu için Sihirli Çeyrekler (Magic Quadrants) raporunda lider konumda yer almaktadır.

H2O platformu Java programlama dili ile geliştirilmiş, açık kaynak kodlu bir platformdur. Python ve R dilleri için uygulama programı arabirimi (API) mevcuttur. Diğer programlama dilleri ile de REST ara yüzü üzerinden kullanılabilir. Bu çalışmada Windows 10 işletim sistemi üzerinde H2O 3.18.0.5 sürümü kullanılmıştır. H2O programının hem kendi web ara yüzü hem de Python arabirimi kullanılmıştır. H2O uygulaması tek bir bilgisayar üzerinde çalıştırılabileceği gibi bir çok bilgisayar üzerinde

paralel bir şekilde de çalıştırılabilir. Veri kümesinin çok büyük olduğu veya tek bir bilgisayarın belleğinin yeterli olmadığı durumlarda birkaç bilgisayar ile küme (cluster) yapılabilir. Ayrıca H2O Amazon AWS üzerinde de çalıştırılabilmektedir. Bu çalışmada kullanılan veri kümesi büyük boyutlarda olmadığı için, H2O tek bir bilgisayar üzerinde çalıştırılmıştır. Python uygulama kodları ile H2O farklı bilgisayarlar üzerinde bulunmaktadır. H2O'nun bulunduğu bilgisayar sunucu, Python kodlarının bulunduğu bilgisayar istemci olarak adlandırılmıştır. İstemci üzerinde bulunan Python kodları, sunucu bilgisayar üzerindeki H2O ile iletişime geçip sınıflandırma algoritmalarını çalıştırmaktadır.



Şekil 4.1 Gartner sihirli çeyrekler

Veri kümesi üzerinde yapılan ön işlemler için açık kaynaklı makine öğrenmesi ve veri madenciliği uygulaması olan WEKA 3.8.2 kullanılmıştır. WEKA ile özellik seçme, numerik kodlama ve veri kümesinin csv formatına dönüştürülmesi işlemleri yapılmıştır.

Çizelge 4.1 Bilgisayar konfigürasyonu

Konfigürasyon	Sunucu	İstemci
İşlemci	Intel Core i7 7700 3.6 GHZ	Intel Core i5 2430M 2.4 GHZ
Bellek	16 GB	4 GB
İşletim sistemi	Windows 10	Windows 10
Tip	Masaüstü	Dizüstü

4.3 Modellerin Karşılaştırılması

Bu çalışmada oluşturulan sınıflandırma modelleri veri kümesinin eğitim için ayrılan kısmı ile eğitilmiş ve test için ayrılan kısmı ile başarısı ölçülmüştür. Bütün modellerde çapraz doğrulama kullanılmıştır.

4.3.1 Derin öğrenme modeli sonuçları

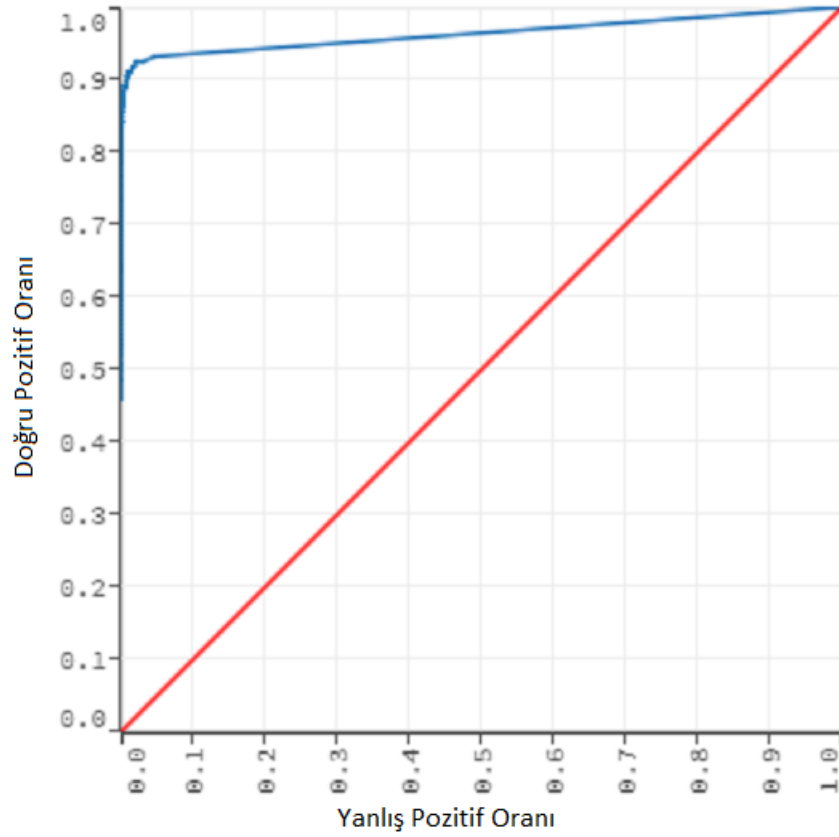
Derin öğrenme modeli çizelge 3.1’de verilen parametrelerle kredi kartı işlem veri kümesi üzerinde çalıştırılmıştır. Veri kümesinin yüzde 70’ini oluşturan eğitim kümesi ile model eğitilmiş; veri kümesinin %30’una karşılık gelen test kümesi ile model başarısı test edilmiştir. Çizelge 4.4’de verilen hata matrisi incelendiğinde, test kümesindeki 146 sahte işlem den 115 adedi doğru olarak sınıflandırılmıştır. 85331 adet gerçek işlemin 23 adedi yanlış sınıflandırılmıştır. Bu veriler ışığında, derin öğrenme modelinin sahte işlem tespit etme oranı % 78,7’dir. Gerçek işlemleri sahte olarak sınıflandırma oranı oldukça düşüktür; yani bankaların yanlış alarm sonrasında gerçek işlemlere sahte işlem muamelesi yapıp kart sahiplerinin memnuniyetsizliğine yol açması ihtimali düşük olmaktadır. Veri kümesi dengesiz olduğu için doğruluk oranının performans değerlendirmesinde bir önemi yoktur. Performans değerlendirmesinde MCC ve AUC değerleri ile sahte işlem tespit etme oranı dikkate alınmıştır. AUC değeri, şekil 4.2’de verilen ROC eğrisi altında kalan alandır.

Çizelge 4.2 Derin öğrenme modeli metrikleri

Metrik	Değer
Doğruluk	0,999
Kesinlik	0,833
Hassasiyet	0,787
MCC	0,809
F1	0,81
AUC	0,963

Çizelge 4.3 Derin öğrenme modeli hata matrisi

	Gerçek	Sahte	Hata Oranı
Gerçek	85402	20	0.00023
Sahte	31	115	0.212



Şekil 4.2 Derin öğrenme modeli ROC eğrisi

4.3.2 Rastgele orman modeli sonuçları

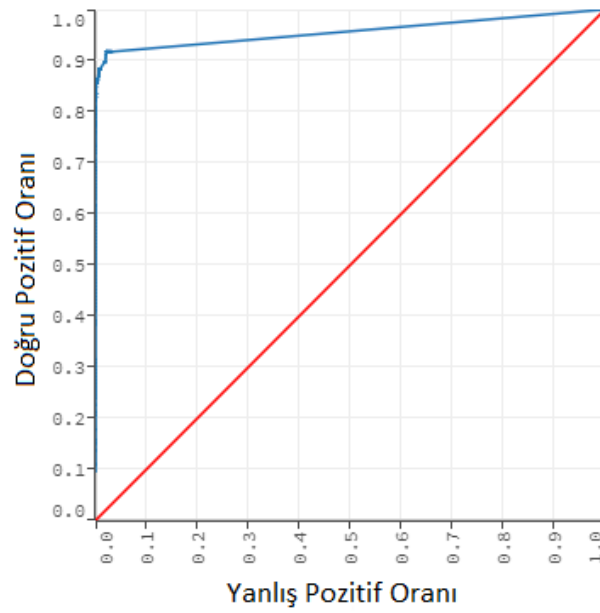
Rastgele orman modeli veri kümesi üzerinde çalıştırılmıştır. İlgili sonuçlar çizelge 4.6'da verilmiştir.

Çizelge 4.4 Rastgele orman modeli metrikleri

Metrik	Değer
Doğruluk	0,9996
Kesinlik	0,941
Hassasiyet	0,773
MCC	0,853
F1	0,849
AUC	0,956

Çizelge 4.5 Rastgele orman modeli hata matrisi

	Gerçek	Sahte	Hata Oranı
Gerçek	85324	7	0,00008
Sahte	33	113	0,22603



Şekil 4.3 Rastgele orman ROC grafiği

4.3.3 Sınıflandırıcı yığını modeli sonuçları

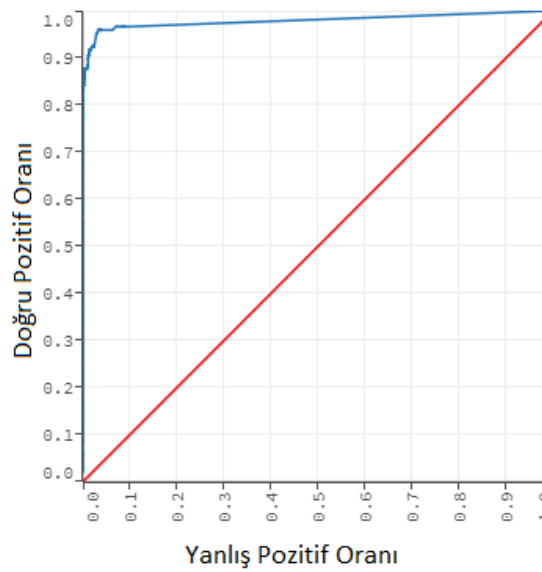
Sınıflandırıcı yığnında temel öğrenici olarak derin öğrenme ve rastgele orman modelleri kullanılmış, meta öğrenici olarak da yine rastgele orman algoritması kullanılmıştır. Meta öğrenici olarak kullanılan rastgele orman modelinde çapraz doğrulama için 2 katlama ve 200 ağaç kullanılmıştır.

Çizelge 4.6 Sınıflandırıcı yığını modeli metrikleri

Metrik	Değer
Doğruluk (Accuracy)	0.999
Kesinlik (Precision)	0.812
Hassasiyet (Recall)	0,823
MCC	0.87
F1	0.817
AUC	0.979

Çizelge 4.7 Sınıflandırıcı yığını hata matrisi

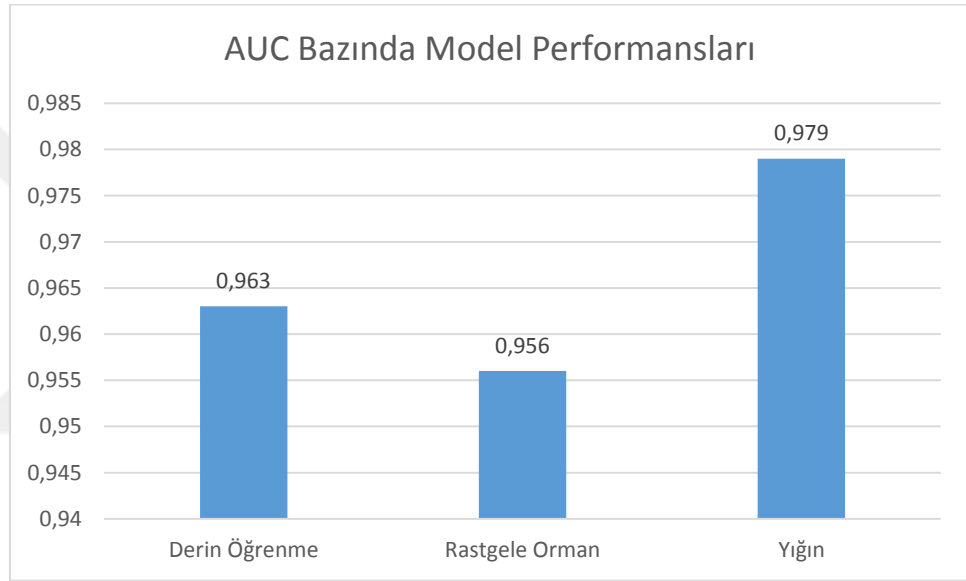
	Gerçek	Sahte	Hata Oranı
Gerçek	85303	28	0,00033
Sahte	26	121	0,18705



Şekil 4.4 Sınıflandırıcı yığını ROC eğrisi

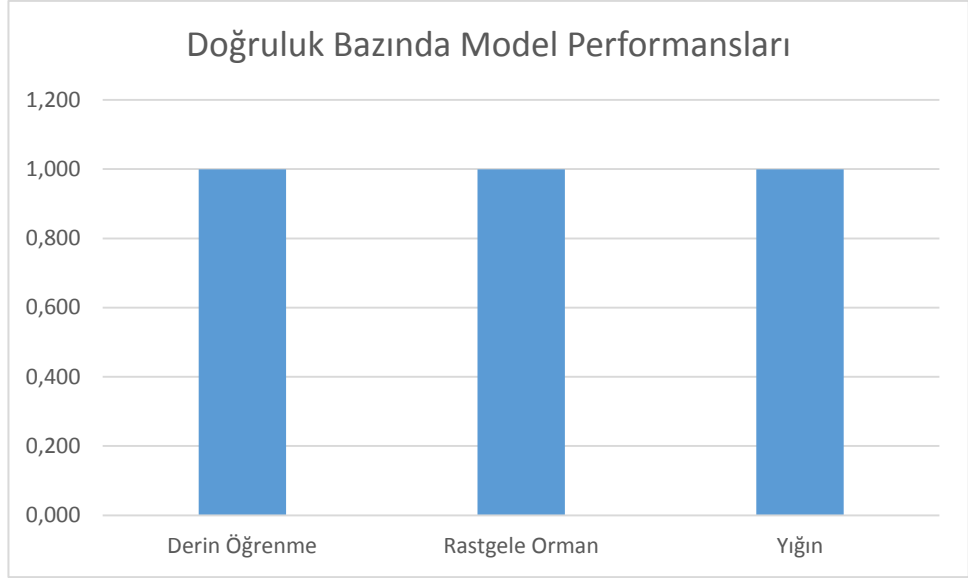
4.3.4 Sonuçların Karşılaştırılması

Geliştirilen üç modelin farklı metriklerdeki sonuçları görsel olarak karşılaştırılmıştır. AUC, doğruluk, kesinlik ve MCC değerleri bazında karşılaştırmalar gerçekleştirilmiştir. Şekil 4.5’de üç modelin AUC değerleri görsel olarak kıyaslanmıştır. Grafikten çıkarılan sonuca göre, en iyi AUC değerine sahip model sınıflandırıcı yığını olmuştur. En düşük AUC değeri ise rastgele orman ile elde edilmiştir. Sınıflandırıcı yığını ise derin öğrenme ile rastgele orman arasında bir AUC değerine sahip olmuştur.



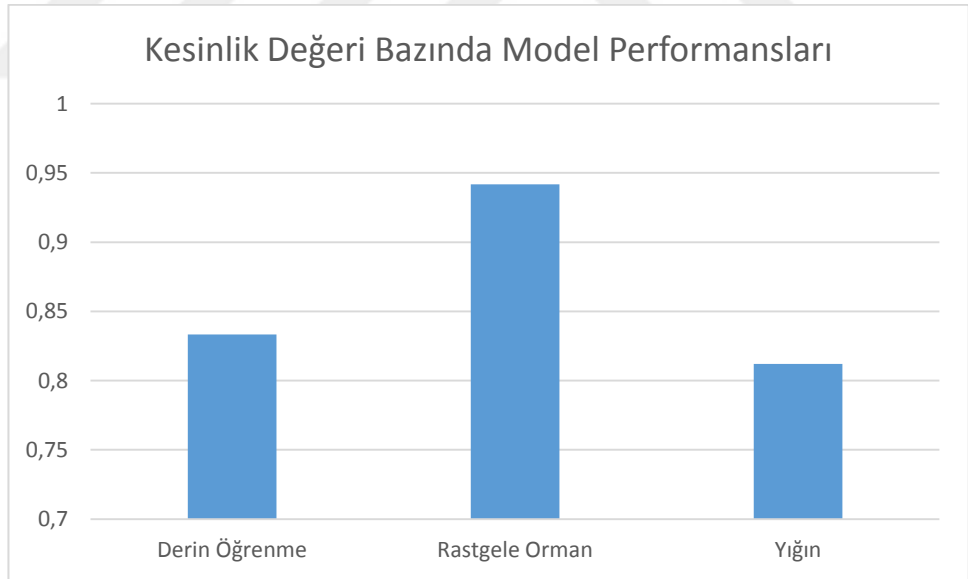
Şekil 4.5 Modellerin AUC bazında karşılaştırması

Her bir model ile yapılan sınıflandırma sonucunda elde edilen doğruluk değerleri Şekil 4.6’da karşılaştırılmıştır. Kullanılan veri kümesi oldukça dengesiz bir küme olduğu için tüm modeller çok yüksek doğruluk değerlerine sahiptir. Şekil 4.6’dan da anlaşıldığı gibi doğruluk değerinin bu modelleri kıyaslamak için doğru bir performans metriği olmadığı görülmektedir.



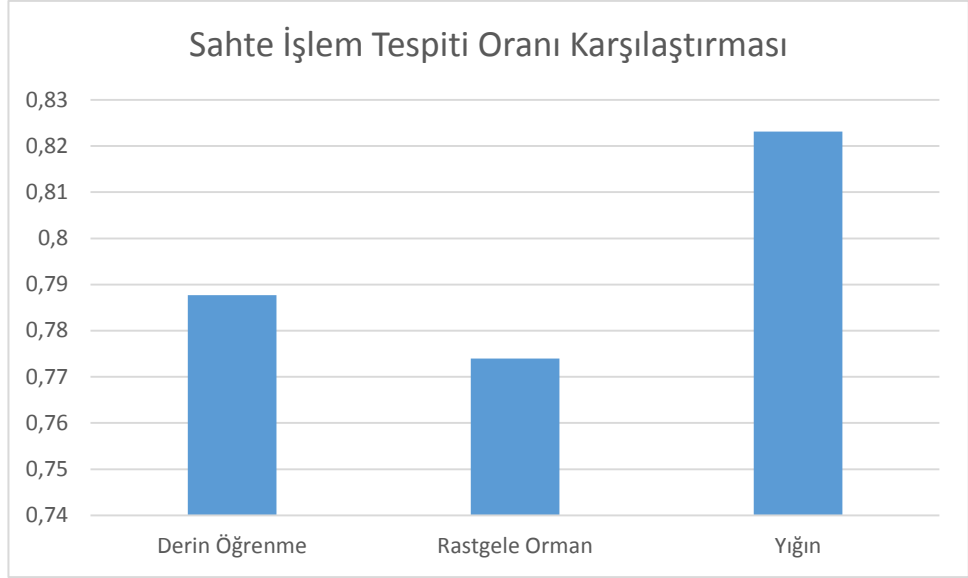
Şekil 4.6 Modellerin doğruluk karşılaştırılması

Kesinlik değerleri incelendiğinde (Şekil 4.7) en yüksek değere rastgele orman modeli ulaşmıştır.



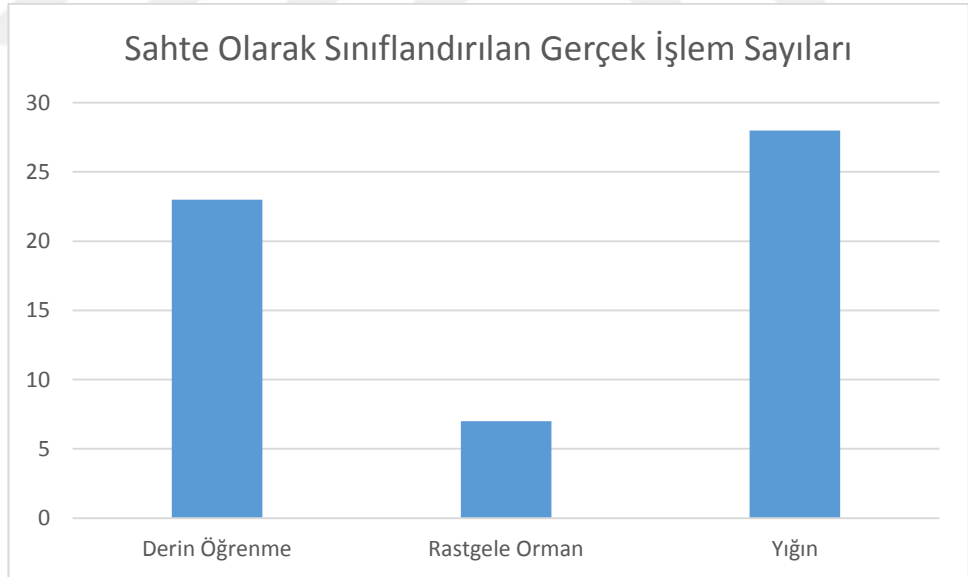
Şekil 4.7 Modellerin kesinlik karşılaştırması

Modeller sahte işlem tespiti bazında kıyaslandığı zaman en iyi orana sahip modelin sınıflandırıcı yığını olduğu görülmektedir.



Şekil 4.8 Sahte işlem tespiti oranı bazında modellerin kıyaslanması

Sahte işlem olarak sınıflandırılan gerçek işlem sayılarında en iyi model rastgele orman olmuştur.



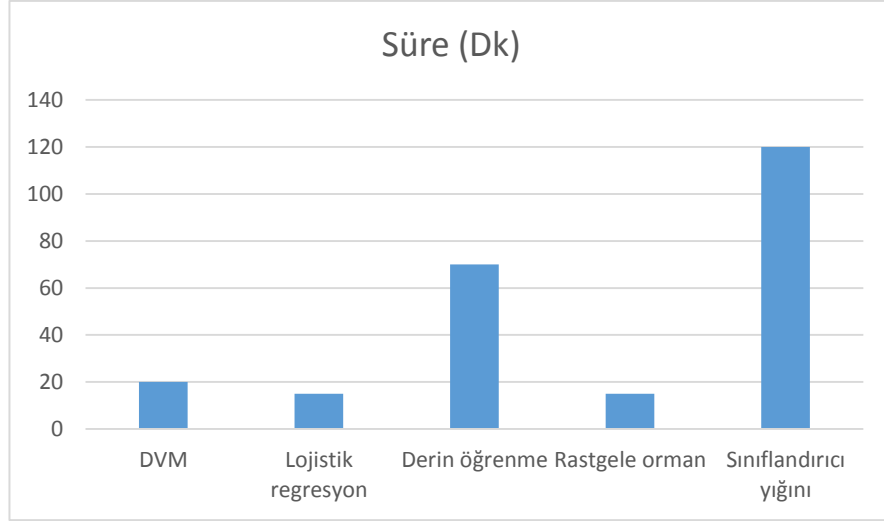
Şekil 4.9 Sahte olarak sınıflandırılan gerçek işlem sayıları

5. SONUÇ

Literatürde kredi kartı sahte işlem tespiti üzerine yapılan çalışmaların bu alandaki katkıları ile bu çalışmanın katkıları incelenmiş ve karşılaştırılmıştır. Kullanılan veri kümelerinin gizli olması, elde edilememesi ve farklı veri kümeleri üzerinde çalışılması sebebiyle literatürdeki çalışmalar ile tam anlamıyla doğru bir karşılaştırma yapılması güçtür. Bu yüzden, literatürde kullanılan destek vektör makinesi, lojistik regresyon gibi yöntemler de test edilmiş ve bu çalışmada kullanılan rastgele orman, derin öğrenme ve yığın yöntemi ile sonuçları karşılaştırılmıştır.

Derin öğrenme sınıflandırma modelinde katman sayısı arttırıldıkça başarının artıp artmadığı incelenmiştir. 500 nöron içeren iki katman kullanılan derin öğrenme modelinde sahte işlem tespit etme oranı 0,787 olmuştur. Katman sayısı 3'e çıkarıldığında hassasiyet, yani sahte işlem tespit etme oranı 0,781 elde edilmiştir. Katman arttırmanın başarıya olumlu bir etkisi olmamıştır fakat çalışma zamanı artmıştır. Katman sayısı ve nöron sayıları arttırılarak test edilmiş ancak erken durdurma (early stopping) mekanizmasından dolayı, 5 döngü boyunca bir iyileşme olmadığı için durdurulmuştur.

Çalışmada kullanılan bilgisayar (Çizelge 4.1) ile iki katmanlı derin öğrenme modelinin çalışması, bilgisayarın işlemcisi üzerindeki yüke bağlı olarak 60 dakika ile 75 dakika arasında zaman almıştır. Aynı nöron sayısına sahip yeni bir katman eklendiğinde çalışma zamanı yüke bağlı olarak 90 dakika ile 120 dakika arasında sürmüştür. Rastgele orman sınıflandırıcısının çalışma zamanı 10 dakika ile 20 dakika arasında değişmektedir. Şekil 5.1'de yöntemlerin yaklaşık çalışma zamanları verilmiştir.

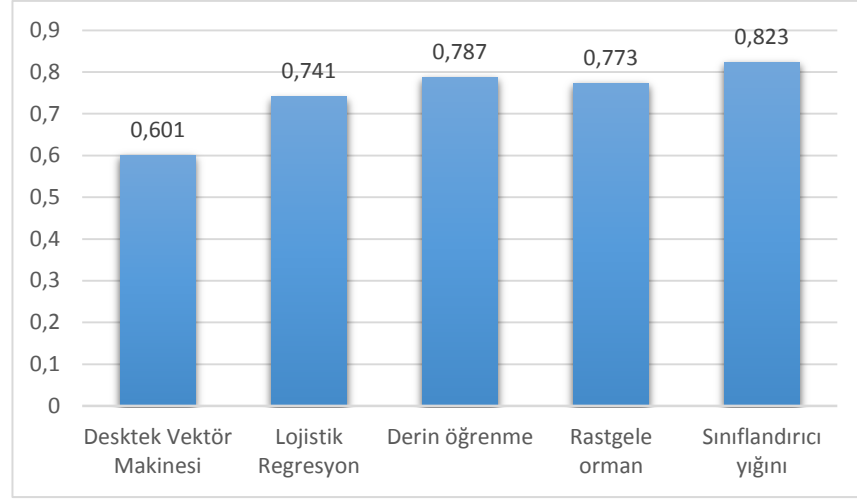


Şekil 5.1 Çeşitli yöntemlerin çalışma zamanları

Pozzolo vd. ve Carneiro N. vd.'nin çalışmalarında destek vektör makinesi yöntemi kullanılmıştır. Bu tez çalışmasında da destek vektör makinesi (SVM) ile kredi kartı işlemleri veri kümesi üzerinde sınıflandırma yapılması test edilmiştir. Sahte işlem tespit etme oranı 0,601 elde edilmiştir. Destek vektör makinesinde, çekirdek fonksiyonu olarak RBF (Radial Basis Function), gama değeri olarak 0.0464 ve c parametresi değeri olarak 10 kullanılmıştır.

Zeager vd. ve Carneiro N. vd.'nin çalışmalarında lojistik regresyon yöntemi kullanılan yöntemlerden bir tanesidir. Bu çalışmada lojistik regresyon yönteminin sonucunu görmek amacıyla, lojistik regresyon modeli eğitilmiş ve test edilmiştir. Sahte işlem tespit etme oranı 0,741 olarak elde edilmiştir. Lojistik regresyon modelinde çözücü (solver) parametresi “l_bfgs”, aile (family) parametresi “binomial” kullanılmıştır.

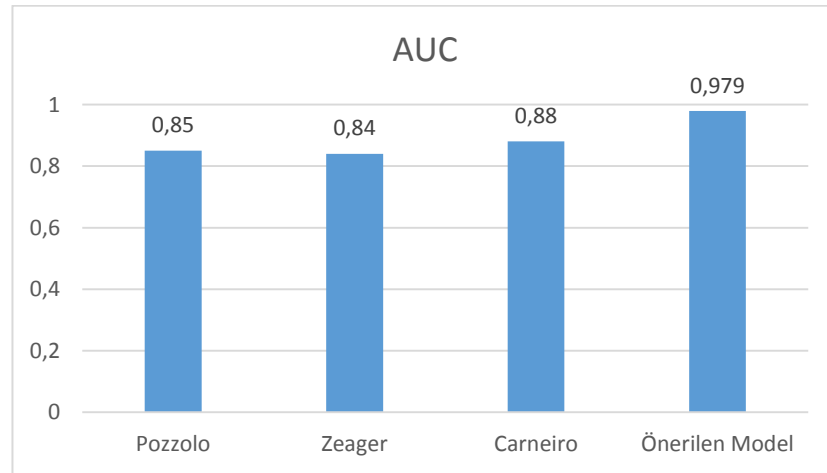
Zeager vd. tarafından yapılan düşmanca öğrenme üzerine tasarlanmış ve lojistik regresyon tabanlı model 0.84 AUC değerine sahip olmuştur. Mahmud M.S'nin çalışmasında test kümesi üzerinde sahte işlem yakalama başarısı en yüksek %45.94 olmuştur. Doğruluk değerinde de %98 elde etmiştir. Pozzolo, kredi kartı sahte işlem tespiti çalışmasında 0,85 AUC değerini yakalamıştır.



Şekil 5.2 Aynı veri kümesi üzerinde sahte işlem tespiti oranı bazında karşılaştırma

Şekil 5.2’de görüleceği gibi, hassasiyet (recall) yani sahte işlem tespit etme oranında en başarılı yöntem sınıflandırıcı yığını olmuştur. Bu kıyaslamada, bu tez çalışmasında kullanılan veri kümesi kullanılmıştır. En düşük sonucu veren yöntem ise destek vektör makinesi olarak görünmektedir.

Bu tez çalışması kapsamında AUC değeri en iyi sınıflandırıcı yığını yöntemiyle 0,979 elde edilmiştir.



Şekil 5.3 Farklı çalışmaların farklı veri kümeleriyle sonuçlarının kıyaslanması

Sahte işlem tespiti, veri kümelerinin çok dengesiz bir halde olmasından dolayı ve dolandırıcıların sürekli yöntem değiştirmesinden dolayı zor bir işlemdir. Kart sahipleri zamanla harcama alışkanlıklarını değiştirdiği gibi dolandırıcılar da müşterilerin alışkanlıklarını takip etmekte ve benzer davranışlar göstermeye çalışmaktadır. Bu sebeple, sahte işlem tespitinde çok yüksek başarılar elde etmek zordur.



KAYNAKLAR

- Albon, C. 2018. Machine Learning for Python Cookbook. O'Reilly Media, 366, USA.
- Anonim. 2018. Web sitesi: <http://bkm.com.tr/pos-atm-kart-sayilari/> Erişim Tarihi: 10.05.2018
- Bahnsen, A.C., Aouada, D., Stojanovic, A. and Ottersten, B. 2016. Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, 51, 134-142.
- Bekkar, M., Djemaa, H.K., Alitouche, T.A. 2013. Evaluation Measures for Models Assesment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*, 3(10)
- Carneiro, N, Figueria, G, Costa, M. 2017. A data mining based system for credit card fraud detection in e-tail. *Decision Support Systems*, 95, 91-101.
- Chandrashekar, G. and Sahin, F . 2014. A survey on feature selection methods. *Computers and Electrical Engineering*, 40(1), 16-28.
- Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, Volume 16, 321-357
- Dixit, A. 2017. Ensemble Machine Learning. Packt Publishing, 438, Birleşik Krallık.
- Duman, E., Ozcelik, M. H. 2011. Detecting credit card fraud by genetic algorithm and scatter search. *Expert Systems with Applications*, 38, 13057-13063.
- Halvaiee, N.S., Akbari, M.K. 2014. A novel model for credit card fraud detection using Artificial Immune Systems. *Applied Soft Computing*, 24, 40-49.
- Kotsiantis, S., Kanellopoulos, D. and Pintelas, P. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30, 25-36.
- Pandey, Y. 2017. Credit Card Fraud Detection Using Deep Learning, *International Journal of Advanced Research in Computer Science*, 8(5).
- Pawar, A., Patil, V., Martin, S. And Chaudhari, M.S. 2017. Credit card fraud detection using Hidden Markov Model. *Imperial Hournal of Interdisciplinary Research*, 3(4), 37-48.
- Pozzolo, A.D., Caelen, O., Johnson, R.A. and Bontempi, G. (2015) Calibrating Probability with Undersampling for Unbalanced Classification. *Symposium on Computational Intelligence and Data Mining (CIDM)*, IEEE.
- Pozzolo, A.D, Caelen, O, Borgne, Y.L, Waterschoot, S, Bontempi, G. 2014. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10), 4915-4928
- Quah, J.T.S., Sriganesh, M. 2008. Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications*, 35, 1721-1732.

- Zareapoor, M. and Shamsolmoali, P. 2015. Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. International Conference on Intelligent Computing, Communication and Convergence, India.
- Zeager, M.F., Sridhar, A., Fogal, N., Adams, S., Brown, D.E. and Beling, P.A. 2017. Adversarial Learning in Credit Card Fraud Detection, Systems and Information Engineering Design Symposium, Charlottesville, VA, USA.
- Zheng, A. and Casari, A. Feature Engineering for Machine Learning. O'Reilly Media, 218, USA.



ÖZGEÇMİŞ

Adı Soyadı : Kazım SOYLU

Doğum Yeri : Ankara

Doğum Tarihi : 10.05.1987

Medeni Hali : Evli

Yabancı Dili : İngilizce

Eğitim Durumu

Lise : İbni Sina Lisesi (2006)

Lisans : Ankara Üniversitesi Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü (2012)

Yüksek Lisans : Ankara Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı (Eylül 2012–Haziran 2018)

Çalıştığı Kurumlar

Bilişim Uzmanı, Türkiye Cumhuriyet Merkez Bankası, 2016-Devam ediyor

Uygulama Geliştirme Uzmanı, İnnova Bilişim Çözümleri, 2012-2016

Yazılım Uzmanı, Başarsoft, 2012-2012

Programcı, Ankira Elektronik, 2011 – 2012