

**T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
İSTATİSTİK ANA BİLİM DALI**



**MAKİNE ÖĞRENİMİNDE SINIFLANDIRMA YÖNTEMLERİ
KULLANILARAK ULAŞIM KARTI SİSTİMALİNİN TESPİT
EDİLMESİ**

Yüksek Lisans Tezi

Serhat DEMİRTÜRK

Danışman

Doç. Dr. Emre DÜNDER

SAMSUN
2023

TEZ KABUL VE ONAYI

Serhat DEMİRTÜRK tarafından, Prof. Dr. Emre DÜNDER danışmanlığında hazırlanan “MAKİNE ÖĞRENİMİNDE SINIFLANDIRMA YÖNTEMLERİ KULLANILARAK ULAŞIM KARTI SUİSTİMALİNİN TESPİT EDİLMESİ” başlıklı bu çalışma, jürimiz tarafından 13.2.2023 tarihinde yapılan sınav sonucunda oy birliği ile başarılı bulunarak Yüksek Lisans Tezi olarak kabul edilmiştir.

	Unvanı Adı Soyadı Üniversitesi Ana Bilim/Ana Sanat Dalı	Sonuç
Başkan	Prof. Dr. Mehmet Ali CENGİZ Ondokuz Mayıs Üniversitesi İstatistik Ana Bilim Dalı	☒ Kabul ☐ Ret
Üye	Doç.Dr. Üyesi Emre DÜNDER Ondokuz Mayıs Üniversitesi İstatistik Ana Bilim Dalı	☒ Kabul ☐ Ret
Üye	Dr.Öğr.Üyesi Tuba KOÇ Ondokuz Mayıs Üniversitesi İstatistik Ana Bilim Dalı	☒ Kabul ☐ Ret

Bu tez, Enstitü Yönetim Kurulunca belirlenen ve yukarıda adları yazılı jüri üyeleri tarafından uygun görülmüştür.

Prof. Dr. Ahmet TABAK
Enstitü Müdürü

BİLİMSEL ETİĞE UYGUNLUK BEYANI

Hazırladığım Yüksek Lisans tezinin bütün aşamalarında bilimsel etiğe ve akademik kurallara riayet ettiğimi, çalışmada doğrudan veya dolaylı olarak kullandığım her alıntıya kaynak gösterdiğimi ve yararlandığım eserlerin Kaynaklar'da gösterilenlerden oluştuğunu, her unsurun enstitü yazım kılavuzuna uygun yazıldığını ve TÜBİTAK Araştırma ve Yayın Etiği Kurulu Yönetmeliği'nin 3. bölüm 9. maddesinde belirtilen durumlara aykırı davranılmadığını taahhüt ve beyan ederim.

Etik Kurul Gerekli mi?

Evet ☐ (Gerekli ise ekler kısmına ekleyiniz)

Hayır ☒

... / ... / 2023
Serhat DEMİRTÜRK

TEZ ÇALIŞMASI ÖZGÜNLÜK RAPORU BEYANI

Tez Başlığı: MAKİNE ÖĞRENİMİNDE SINIFLANDIRMA YÖNTEMLERİ
KULLANILARAK ULAŞIM KARTI SUİSTİMALİNİN TESPİT EDİLMESİ

Yukarıda başlığı belirtilen tez çalışması için şahsım tarafından 03.01.2023 tarihinde intihal tespit programından alınmış olan özgünlük raporu sonucunda;

Benzerlik oranı : % 19

Tek kaynak oranı : % 3 çıkmıştır.

... / ... / 2023
Doç. Dr. Emre DÜNDER

ÖZET

MAKİNE ÖĞRENİMİNDE SINIFLANDIRMA YÖNTEMLERİ KULLANILARAK ULAŞIM KARTI SUİSTİMALİNİN TESPİT EDİLMESİ

Serhat DEMİRTÜRK

Ondokuz Mayıs Üniversitesi

Lisansüstü Eğitim Enstitüsü

İstatistik Ana Bilim Dalı

Yüksek Lisans, Ocak/2023

Danışman: Doç. Dr. Emre DÜNDER

Makine öğrenimi günümüzde hem bilim dünyasında hem de iş dünyasında birçok problemin çözülmesine somut ve hızlı çözümler üretmekte olup, yeni neden sonuç ilişkilerinin ortaya çıkmasına katkı sağlamaktadır. Bu nedenle son yıllarda makine öğreniminin kullanımı giderek yaygınlaşmış ve popüler bir bilim haline gelmiştir. Günümüzde ise ödeme sistemlerinde yaşanan birtakım gelişmeler yeni problemlerin ortaya çıkmasına neden olmuştur ve olmaya devam etmektedir. Bahsi geçen problemlerin başında akıllı kartlarda meydana gelen suistimaller yer almaktadır.

Bu çalışmada da akıllı kart olarak nitelendirilen ulaşım kartlarında yapılan suistimaller incelenerek literatürde en sık kullanılan sınıflandırma algoritmaları kullanılmış ve bu algoritmaların sınıflandırma performansları karşılaştırılarak en iyi algoritmaya göre değişkenlerin önem seviyeleri değerlendirilmiştir. Kullanılan algoritmalar ise, Karar Ağaçları, Rastgele Ormanlar, Destek Vektör Makineleri, Lojistik Regresyon, Naif Bayes, Adaboost, XGBoost, K-En Yakın Komşu ve Yapay Sinir Ağlarından Derin Öğrenme algoritmalarıdır. Bu algoritmaların sınıflandırma performanslarının ölçümünde doğruluk, MCC, F1 ölçütü ve AUC ölçütleri kullanılmıştır. Yine, bu çalışmada öncelikli olarak tüm değişkenlerin veri seti üzerindeki dağılımları ve nümerik değişkenlerin normal dağılım durumları incelenmiş, normal dağılım göstermeyen nümerik değişkenlere logaritmik dönüşüm uygulanmıştır. Çalışmada değişken seçme yöntemlerinden Boruta yöntemi tercih edilmiş ve bu yöntemle göre bütün değişkenler anlamlı bulunmuştur. Bütün değişkenler modele dahil edilmeden önce bu değişkenler üzerinde 10 katlı çapraz geçerlilik uygulanmış ve veri setinde yer alan tüm değişkenler çalışmaya konu olan tüm makine öğrenimi algoritmalarına dahil edilmiştir.

Tüm bu hususlar neticesinde XGBoost algoritmasının diğer algoritmalara göre daha yüksek doğruluk derecesine sahip olduğu gözlenmiştir. XGBoost algoritmasının doğruluk derecesi 0,881, MCC değeri 0,750, AUC değeri 0,953 ve F1 ölçütü de 0,875 olarak ölçülmüştür. Bu sonuca göre XGBoost algoritmasının doğruluk derecesinin en yüksek olması ve orta düzeyde model başarısına sahip olması nedeniyle ulaşım kartı suistimalinin tespit edilmesinde en başarılı sınıflandırıcı olduğu gözlenmiştir.

Anahtar Sözcükler: Makine öğrenimi, sınıflandırma algoritmaları, ulaşım kartı, akıllı kart, kart suistimali, ödeme sistemleri

ABSTRACT

DETECTION OF TRANSPORTATION CARD FRAUD BY CLASSIFICATION METHOD IN MACHINE LEARNING

Serhat DEMİRTÜRK
Ondokuz Mayıs University
Institute of Graduate Studies
Department of Statistics
Master, January/2023

Supervisor: Assoc. Prof. Dr. Emre DÜNDER

Today, machine learning produces concrete and fast solutions to solving many problems both in the scientific world and in the business world, and contributes to the emergence of new cause-effect relationships. Therefore, in recent years, the use of machine learning has become increasingly widespread and has become a popular science. Today, some developments in payment systems have led to the emergence of new problems and continue to do so. At the beginning of the aforementioned problems is the abuse of smart cards.

In this study, the most frequently used classification algorithms in the literature were used by examining the abuses made in the transportation cards, which are described as smart cards, and the classification performances of these algorithms were compared and the importance levels of the variables were evaluated according to the best algorithm. The algorithms used are Decision Trees, Random Forests, Support Vector Machines, Logistic Regression, Naive Bayes, Adaboost, XGBoost, K-Nearest Neighbor and Deep Learning algorithms from Artificial Neural Networks. Accuracy, MCC, F1 criteria and AUC criteria were used to measure the classification performance of these algorithms. Again, in this study, first of all, the distribution of all variables on the data set and the normal distribution of the numerical variables were examined, and the logarithmic transformation was applied to the numerical variables that did not show normal distribution. In the study, Boruta method was preferred among the variable selection methods and all variables were found significant according to this method. Before all variables were included in the model, 10-fold cross-validation was applied on these variables and all variables in the data set were included in all machine learning algorithms that were the subject of the study.

As a result of all these issues, it has been observed that the XGBoost algorithm has a higher degree of accuracy than other algorithms. The accuracy of the XGBoost algorithm was 0,881, the MCC value was 0,750, the AUC value was 0,953, and the F1 criterion was 0,875. According to this result, it has been observed that XGBoost algorithm is the most successful classifier in detecting transportation card abuse, since it has the highest accuracy and moderate model success.

Keywords: Machine learning, classification algorithms, transportation card, smart card, card fraud, payment systems

ÖN SÖZ VE TEŞEKKÜR

Yüksek lisans tezimin konusunu oluşturan “Makine Öğreniminde Sınıflandırma Yöntemleri Kullanılarak Ulaşım Kartı Suistimalinin Tespit Edilmesi”, ulaşım kartlarında gerçekleşen suistimallerin belediyelere getirmiş olduğu finansal zararların azaltılması ve kamu güvenliğinin korunması bakımından oldukça önem arz etmektedir. Çalışmada değerlendirilen konuya literatürde fazla yer verilmemiş olması nedeniyle yüksek lisans tezimi bu alanda yazma ihtiyacı hissettim.

Bu süreçte, yoğun çalışma programından bana vakit ayırarak her sorumu yanıtlayan, lisans ve yüksek lisans sürecimde gerek manevi gerekse de bilgi aktarımı konusunda desteğini hiçbir zaman esirgemeyen, tez çalışmama yorum ve fikirleri ile katkı sağlayan tez danışmanım Sayın Doç. Dr. Emre DÜNDER’e,

Tez çalışmamın araştırma ve uygulama aşamasında her sorumu sabırla yanıtlayan ve desteklerini esirgemeyen Arş. Gör. Fatih SAĞLAM’a,

Tez çalışmalarım boyunca benden desteğini esirgemeyen ve kişisel motivasyonumu arttıran kıymetli eşim Bilge YILDIZ DEMİRTÜRK’e,

Son olarak bugünlere gelmemde en büyük katkıyı sağlayan, gerek manevi gerek maddi olarak beni hiçbir zaman yalnız bırakmayan sevgili annem Sunay DEMİRTÜRK’e, sevgili babam Mehmet DEMİRTÜRK’e ve kardeşlerim Serap DEMİRTÜRK ÖZSOY ile Seda DEMİRTÜRK’e teşekkür etmeyi bir borç bilirim.

Serhat DEMİRTÜRK

İÇİNDEKİLER

TEZ KABUL VE ONAYI	i
BİLİMSEL ETİĞE UYGUNLUK BEYANI	ii
TEZ ÇALIŞMASI ÖZGÜNLÜK RAPORU BEYANI	ii
ÖZET	iii
ABSTRACT	iv
ÖNSÖZ VE TEŞEKKÜR	v
İÇİNDEKİLER.....	vi
SİMGELER VE KISALTMALAR	vii
ŞEKİLLER DİZİNİ	viii
TABLolar DİZİNİ	ix
1. GİRİŞ.....	11
2. TOPLU TAŞIMADA AKILLI KARTLARIN KULLANIMI.....	14
2.1. Toplu Taşıma	14
2.2. Toplu Taşımada Kullanılan Akıllı Kartlar ve İstanbulkart Örneği.....	14
2.2.1. Akıllı Kart	14
2.2.2. Akıllı Kartların Toplu Taşımada Kullanımı	14
2.2.3. Bir Akıllı Ulaşım Kartı Örneği Olarak: İstanbulkart	15
2.2.4. Akıllı Ulaşım Kartlarının Kamu Güvenliğindeki Rolü	17
2.2.5. Akıllı Ulaşım Kartlarında Yapılan Suistimaller ve Hukuksal Boyutu	18
2.3. Kredi Kartı Dolandırıcılığı ve Bu Dolandırıcılık Türünün Ulaşımında Kullanılan Akıllı Kartlar ile Benzerliği	20
2.3.1. Kredi Kartı Dolandırıcılık Türleri.....	23
2.3.2. Kredi Kartı Dolandırıcılığının Sınıflandırılması ile İlgili Örnek Çalışmalar	24
3. MAKİNE ÖĞRENİMİ.....	27
3.1. Veri Madenciliği	29
3.2. Makine Öğrenimi ile Veri Madenciliği Arasındaki Farklar	31
3.3. Denetimli ve Denetimsiz Öğrenme	32
3.4. Aşırı Öğrenme (Overfitting) ve Eksik Öğrenme (Underfitting).....	33
3.4.1. Makine Öğreniminde Sınıflandırma Teknikleri	34
3.5. Çalışmada Kullanılan Algoritmalar.....	34
3.5.1. Karar Ağaçları Algoritması	35
3.5.2. Rastgele Orman Algoritması	36
3.5.3. Destek Vektör Makineleri	38
3.5.4. Naif Bayes Algoritması.....	42
3.5.5. XGBOOST Algoritması.....	43
3.5.6. Lojistik Regresyon Analizi.....	43
3.5.7. K-En Yakın Komşuluk Algoritması	46
3.5.8. Adaboost Algoritması	48
3.5.9. Yapay Sinir Ağları ve Derin Öğrenme.....	49
3.6. Değişken Seçimi.....	54
3.6.1. Boruta ile Değişken Seçim Yöntemi.....	55
3.7. Sınıflandırma Algoritmalarında Performans Ölçütleri	56
3.8. Çapraz Geçerlilik.....	58
4. UYGULAMALAR.....	60
4.1. Veri Setinin Elde Edilmesi ve Düzenlenmesi	60
4.2. Veri Setinde Yer Alan Değişkenlerin Dağılımları ve Açıklamaları	62
4.3. Veri Ön İşleme	67
4.4. Sınıflandırma Algoritmalarının Veri Seti Üzerindeki Uygulaması	69
4.5. En İyi Modele Göre Değişkenlerin ve Kategorilerin Önem Seviyeleri	72
5. SONUÇ VE DEĞERLENDİRME	74
KAYNAKLAR	76
ÖZ GEÇMİŞ	83

SİMGELER VE KISALTMALAR

A.Ş.	: Anonim Şirketi
ACC	: Doğruluk Oranı (Accuracy)
AdaBoost	: Adaptive Bossting (Uyarlamalı Artırma)
ANN	: Yapay Sinir Ağları (Artificial Neural Network)
ATM	: Otomatik Vezne Makinesi (Automatic Teller Machine)
AUC	: Eğri Altında Kalan Alan (Area Under The Curve)
BELBİM	: İstanbul Belediyeler Bilgi İşlem Enerji Sanayi ve Ticaret A.Ş.
CART	: Sınıflandırma ve Regresyon Ağaçları
CNP	: Kart Mevcut Değil (Card Not Present)
DN	: Doğru Negatif
DP	: Doğru Pozitif
DVM	: Destek Vektör Makineleri
GB	: Gradyan Artırma (Gradient Boosting)
HES	: Hayat Eve Sığar
IF	: İzolasyon Ormanı (Isolation Forest)
İBB	: İstanbul Büyükşehir Belediyesi
İETT	: İstanbul Elektrik Tramvay ve Tünel İşletmeleri
KA	: Karar Ağaçları
KNN	: K-En Yakın Komşu (K-Nearest Neighbors)
LOF	: Yerel Aykırı Değer Faktörü (Local Outlier Factor)
LR	: Lojistik Regresyon
MKK	: Matthew'in Koreslasyon Katsayısı
MLP	: Çok Katmanlı Algılayıcı (Multi Layer Perceptron)
NB	: Naif Bayes (Naive Bayes)
OOB	: Torba Dışı (Out Of Bag)
PDKS	: Personel Devam Kontrol Sistemi
PIN	: Kişisel Tanıtım Numarası (Personal Identifier Number)
POS	: Satış Noktası (Point Of Sale)
QR	: Hızlı Cevap (Quick Response)
RO	: Rastgele Orman
ROM	: Salt Okunur Bellek (Read Only Memory)
SC	: Yaralama Kümesi (Scarring Cluster)
TCK	: Türk Ceza Kanunu

TDK : Türk Dil Kurumu
TL : Türk Lirası
XGBOOST : Aşırı Gradyan Artırma (Extreme Gradient Boosting)
YN : Yanlış Negatif
YP : Yanlış Pozitif



ŞEKİLLER DİZİNİ

Şekil 2.1. Kart sağlayıcı şirketlerin 2020 ve 2021 yılı kart adetleri (milyar)	21
Şekil 2.2. 2013/2019 yıllarında dolandırıcılık sebebiyle kaybedilen ve 2021/2027 yılları arasında kaybedileceği tahmin edilen dolandırıcılık tutarları (milyar dolar)	22
Şekil 3.1. Makine öğreniminde kullanılan algoritmalar	28
Şekil 3.2. Veri madenciliği sürecine ait akış şeması	30
Şekil 3.3. Veri madenciliğinde denetimli ve denetimsiz öğrenme işlemleri	33
Şekil 3.4. Aşırı öğrenme ve eksik öğrenmeye ait gösterim	34
Şekil 3.5 Karar ağacı algoritmasının yapısı	36
Şekil 3.6. Rastgele orman algoritmasının yapısı	38
Şekil 3.7. Destek vektör makinelerine ait örnek grafik	39
Şekil 3.8. Doğrusal ayrılabilen destek vektör makinelerinin gösterimi	40
Şekil 3.9. Doğrusal ayrılmayan destek vektör makinelerinin gösterimi	41
Şekil 3.10. İkili bağımlı değişkenin s ve ters şeklindeki olasılık fonksiyon grafikleri	45
Şekil 3.11. KNN algoritması sınıflandırma işlemi	46
Şekil 3.12. Adaboost algoritmasının gösterimi	49
Şekil 3.13. Biyolojik yapay sinir ağının gösterimi	49
Şekil 3.14. Yapay nörona ait gösterim	50
Şekil 3.15. Basit bir yapay sinir ağına ait katmanlar	50
Şekil 3.16. İleri beslemeli yapay sinir ağına ait örnek gösterim	51
Şekil 3.17. Geri beslemeli yapay sinir ağına ait örnek gösterim	52
Şekil 3.18. Derin öğrenmeye ait örnek gösterim	53
Şekil 3.19. Performans ölçütlerinin şekil üzerinde gösterimi	56
Şekil 3.20. AUC grafiğinin gösterimi	58
Şekil 3.21. Çapraz geçerliliğe ait örnek gösterim	59
Şekil 4.1. Nümerik değişkenlere ait yoğunluk grafikleri	68
Şekil 4.2. Logaritmlar alındıktan sonra nümerik değişkenlere ait yoğunluk grafikleri	68
Şekil 4.3. Boruta değişken seçimi ile değişken önem seviyeleri	69
Şekil 4.4. XGBOOST modeline göre değişken önem seviyeleri	73

TABLolar DİZİNİ

Tablo 2.1. Kredi kartı dolandırıcılığı ile ilgili yapılmış literatür çalışmaları.....	24
Tablo 4.1. Kitleye ait demografik özellikler.....	63
Tablo 4.2. Kitlenin kart bilgileri ve hareketleri	63
Tablo 4.3. Kitlelerin kart bilgilerinin ve kart hareketlerinin nümerik dağılımları	67
Tablo 4.4. Sınıflandırma algoritmalarına ait performans değerleri.....	70
Tablo 4.5. Sınıflandırma algoritmalarının hiper parametreleri	70



1. GİRİŞ

Büyük şehirlerde nüfus yoğunluğunun artmasıyla toplumun ulaşım talepleri artış göstermiştir. Toplumun ulaşım taleplerinin en iyi şekilde karşılanması belediyeler için büyük bir sorumluluk haline gelmiştir. Toplumun rahat, güvenilir ve konforlu ulaşım imkanı sağlaması belediyelerin en temel görevleri arasında yer almaktadır. Belediyelerin topluma bu imkanı sağlaması şehir içinde yaşanan motorlu araç trafiğinin azalmasına katkı sağladığı gibi çevre kirliliğinin azalması, enerji tüketiminin azalması ve insanların zamanlarını daha etkin bir şekilde kullanmalarına katkı sağlamaktadır.

Belediyeler insanların ulaşım taleplerini karşılayabilmek için birçok stratejik planlamalarda bulunmaktadır. Nüfus yoğunluğu artış gösteren bir bölgeye yeni ulaşım hatlarının dahil edilmesi, çalışma saatleri göz önünde bulundurularak ulaşım sefer sıklığının artış göstermesi ve ulaşım da ücret toplama sisteminin geliştirilmesi gibi örnekler bu stratejik planlara birer örnek olarak gösterilebilir. Bu örneklerden ücret toplama sistemini incelediğimizde ise belediyeler bu sistemi geliştirmek için akıllı kartları kullanmaya başlamıştır. Ulaşım da kullanılan akıllı kartlar her ne kadar ücret toplama sisteminin bir parçası olarak görülse de ulaşım yoğunluğunun incelenmesi adına hangi bölgelerin daha yoğun olduğunu tespit etmek ve hangi zaman dilimlerinde ulaşım talebinin artış gösterdiğini incelemek gibi durumlar akıllı kartların ulaşım alt yapısının geliştirilmesine katkı sağlayan en büyük sistemlerden biridir.

Ulaşım da kullanılan akıllı kartlar sadece belediyelere avantaj sağlamayıp toplum için de bu kartların kullanımının yaygın olmasının birtakım avantajları bulunmaktadır. İnsanların sahip oldukları niteliklere göre ulaşımın ücretsiz ve indirimli olması, akıllı kartların anlaşmalı kurumlar ile ulaşım dışında da kullanılabilir olması (Örneğin; marketler ve belediyelere ait sosyal tesisler) insanların bu kartları kullanımını yaygınlaştırmış ve insanlara büyük avantajlar sağlamıştır.

Ulaşım da kullanılan akıllı kartların yukarıda belirtildiği gibi topluma ve belediyelere her ne kadar avantaj sağladığı görülse de kötü niyetli insanlar tarafından bu kartlarda yapılan birtakım suistimaller belediyelere, topluma ve toplum güvenliğine birtakım zararlar vermektedir.

Belediyeler yapılan bu suistimalleri tespit etmek için birtakım senaryolar geliştirmiş ve bu senaryolar ile suistimallerin en aza indirilmesi sağlanmıştır.

Belediyeler tarafından geliştirilen bu senaryolar belediyelerin finansal kayıplarının önüne geçmesine katkı sağladığı gibi toplum güvenliğinin sağlanmasına da büyük katkılar sağlamıştır. Yapılan bu çalışmada ise makine öğreniminde kullanılan sınıflandırma teknikleri uygulanarak yapılan suistimaller incelenmiştir.

Bilişim teknolojilerinin insan hayatına girmesi ile beraber veri miktarı da her geçen gün artış göstermiş ve bu verilerin de anlamlı bir bilgiye dönüştürülmesi kaçınılmaz bir hale gelmiştir. Verilerden elde edilmek istenen bu anlamlı bilgi avcılığı ve veriye dayalı kararlar alıp verebilme makine öğrenimi olarak adlandırılmıştır. Makine öğrenimi uygulamaları bilimsel ve endüstriyel çalışmalarda birçok problemin çözümüne katkılar sağlayarak yeni merakların, yeni neden sonuç ilişkilerinin ve yeni fütüristik yaklaşımların ortaya çıkmasına katkı sağlamıştır. Ulaşımda kullanılan akıllı kartlarda yaşanan suistimallerin makine öğrenimi teknikleri ile tespit edilmesi ise bu yaklaşımların bir sonucu olarak ortaya çıkmış ve bu çalışma da literatürde yer alan kredi kartı dolandırıcılığının makine öğrenimi teknikleri ile tespit edilmesi çalışmalarından ilham alınmıştır.

Ulaşım kartlarında yapılan bu suistimallerin makine öğrenimi ile tespit edilmesi belediyelerin bu anlamda yaşayacağı finansal kayıpların indirgenmesi ve toplum güvenliğine katkı sağlamak amacıyla tasarlanmıştır. Yapılan çalışmada makine öğreniminde kullanılan sınıflandırma algoritmaları ile bir kart kullanıcısının suistimal yapmasını açıklayan en iyi değişkenlerin hangileri olduğu irdelenmiştir. Buradan amaçla makine öğrenimi sınıflandırma algoritmalarında sıkça kullanılan Karar Ağaçları, Rastgele Ormanlar, Destek Vektör Makineleri, Naif Bayes, Lojistik Regresyon, K-En Yakın Komşuluk, XGBoost, Adaboost ve Yapay Sinir Ağları-Derin Öğrenme algoritmalarının performansları karşılaştırılmış ve en yüksek doğruluk derecesine sahip olan algoritmaya göre değişkenlerin önem sevipleri incelenmiştir.

Çalışma dört bölümden oluşmaktadır. Çalışmanın ilk bölümünde toplu taşımada kullanılan akıllı kartlar, akıllı ulaşım kartlarının kamu güvenliğindeki rolü, akıllı ulaşım kartlarında yapılan suistimallerin hukuksal boyutu, kredi kartlarında yapılan dolandırıcılık türlerinin akıllı ulaşım kartlarında yapılan suistimaller ile benzerliği ve kredi kartı dolandırıcılığında kullanılan makine öğrenme tekniklerine ait literatür çalışmaları incelenmiştir. İkinci bölümde makine öğrenimi ve makine öğreniminde kullanılan birtakım sınıflandırma algoritmalarından bahsedilmiştir. Buna ek olarak makine öğrenimi tekniklerinde kullanılan değişken seçim yöntemi, çapraz geçerlilik

ve sınıflandırma performanslarında kullanılan ölçütlere değinilmiştir. Üçüncü bölüm çalışmanın uygulama bölümünü oluşturmaktadır. Üçüncü bölümde ilk olarak veri setinin özelliklerinden bahsedilmiş, veri seti modellere dahil edilmeden önce üzerinde yapılan dönüşümler incelenmiş ve bu dönüşümlerden sonra veri seti sınıflandırma algoritmalarına dahil edilerek, modellerin sınıflandırma performansları ile hiper parametreleri incelenmiştir. Uygulama bölümünün son aşamasında en yüksek doğruluk derecesine sahip algoritmanın değişken önem seviyeleri incelenmiştir. Çalışmanın dördüncü ve son bölümü ise sonuç ve değerlendirmeden oluşmakta olup, yapılacak olan benzer çalışmalara ait birtakım tavsiyelere yer verilmiştir.



2. TOPLU TAŞIMADA AKILLI KARTLARIN KULLANIMI

2.1. Toplu Taşıma

Büyükşehirlerde nüfus yoğunluğunun artması, insanların ulaşım ihtiyaçlarının çoğalmasına neden olmaktadır. Bununla beraber ülkemizde bakanlıklar ve belediyeler insanların ulaşım ihtiyaçlarını gidermek adına birçok yatırımda bulunmaktadır. Büyükşehirlerde kullanılan ve toplu taşıma aracı olarak nitelendirilen metro, metrobüs, tramvay, otobüs ve vapur gibi ulaşım araçlarının kullanılması toplumun ulaşım ihtiyaçlarının karşılanmasına büyük katkı sağlarken, şehir içi motorlu araç trafiğinin de azalmasına yardımcı olmaktadır. Örneğin İstanbul'da metro hatlarının yer altından gitmesi, Marmaray ve tramvay hatlarının raylı sistem üzerinde ilerlemesi ve Metrobüs hattının da kendine ait yol ve güzergah yapısının olması ulaşımı hızlandırmaktadır. Bunun bir sonucu olarak da insanlar ulaşmak istedikleri bölgelere daha hızlı şekilde varmaktadır.

2.2. Toplu Taşımada Kullanılan Akıllı Kartlar ve İstanbulkart Örneği

2.2.1. Akıllı Kart

Akıllı kart içinde yarı iletken çipler bulunduran, 1 ile 64 kilobyte hafızaya sahip ve ROM üzerinde yazılmış işletim sistemine sahip, plastikten üretilmiş kart olarak nitelendirilmektedir. Çeşitli türde veri saklayabilen bu kartlar günümüzde kredi kartı ve ulaşım kartı olarak kullanılmaktadır. Akıllı kartlar günlük hayatta temaslı ve temassız olarak iki gruba ayrılmaktadır. Temassız kartlar okuma terminalinde herhangi bir yuvaya temas etmeden, cihaza yakından tutularak işlem yapılabilen kartlardır. Temaslı kartlar ise okuma terminaline takılarak kart üzerindeki çipin terminal ile teması sonucunda işlemi gerçekleştirebilen kartlardır (Karakulah, vd., 2011).

2.2.2. Akıllı Kartların Toplu Taşımada Kullanımı

Günümüzde belediyeler insanların toplu taşımada seyahat edebilmeleri için akıllı kartları ödeme aracı olarak kullanmaya başlamış ve bu kartların kullanımı da giderek yaygınlaşmıştır. Bu yaygınlaşmada yer alan önemli etkenlerden birisi de vatandaşlara, sahip oldukları nitelikler nedeniyle birtakım imtiyazlar tanınmasıdır. Örneğin, öğrencilere indirimli, kolluk kuvveti çalışanı, asker ve sağlık çalışanı gibi kişilere yine indirimli, belirli bir yaşın üzerindeki vatandaşlara ve engelli vatandaşlara

ise ücretsiz şekilde ulaşım imkanı sağlanmaktadır. Böylece yukarıda da ifade edildiği üzere, ulaşım kartlarının kullanımları yaygınlaşmaktadır.

Günlük hayatta kullanılan bu kartlarda tüm işlemler, kartın yaşam süresi boyunca kayıt altına alınmaktadır. Kayıt altına alınan bu işlemler, karta yapılan yüklemeler ve yapılan yüklemelerin harcama terminallerinde işlem görmesinden oluşmaktadır. Bu çalışmada İstanbul şehrinde toplu taşımada kullanılan İstanbulkart ürünü incelenmiş ve ürün ile ilgili yapılan yükleme ve harcama bilgilerine aşağıdaki gibi yer verilmiştir.

2.2.3. Bir Akıllı Ulaşım Kartı Örneği Olarak: İstanbulkart

2.2.3.1. İstanbulkart'ın Tarihsel Gelişimi

BELBİM Elektronik Para ve Ödeme Hizmetleri A.Ş. 1987 yılında İstanbul'un toplu taşıma ağının tüm ödeme altyapısının yürütüldüğü, İstanbul Büyükşehir Belediyesi ve iştirak şirketlerinin yanı sıra ilçe belediyelerine de hizmet vermek amacıyla kurulan Türkiye'nin ilk elektronik ücret toplama sisteminin kurucusudur. Şirket 1994 yılında akıllı bilet Akbil projesinin testlerini başlatmış ve 1995 yılından itibaren ilk defa Metro ve İETT otobüslerinde kullanılmaya başlanmıştır. 1997'de ise özel halk otobüsleri ve banliyö tren hatlarına entegre edilmiştir. 2006 yılına gelindiğinde özel halk otobüsleri ve İETT otobüslerinde aylık abonman kullanımına başlanmış ve Akbil'in kullanımı toplu taşımada giderek yaygınlaşmıştır. 2009 tarihinden itibaren temassız kart okutma sistemine geçiş sağlanmış ve temassız kart okuma sistemi ile Akbil birlikte kullanılmaya başlanmıştır. 2011 yılında otobüslerde nakit para kullanımına son verilmiş ve İstanbulkart kullanımı zorunlu hale getirilmiştir. 2014 yılında ise Akbil kullanımı tamamen kaldırılarak sadece İstanbulkart kullanılmaya başlanmıştır. BELBİM A.Ş. 2015 yılında 6493 sayılı Menkul Kıymet Mutabakat Sistemleri, Ödeme Hizmetleri ve Elektronik Para Kuruluşları Kanunu kapsamında elektronik para kuruluşuna dönüştürülmüştür. 2016 yılında ise İstanbulkart bir ulaşım kartı olmaktan da öte ulaşım dışında da kullanılabilen elektronik para kartına dönüştürülmüştür. 2018 yılında İstanbulkart mobil uygulama ile bakiye yükleme, bakiye görüntüleme ve hesap hareketlerinin izlenebilme özelliği getirilmiş ve 2021 yılından itibaren de mobil uygulama üzerinden QR kod ile geçiş özelliği sisteme entegre edilmiştir (BELBİM, 2022).

2.2.3.2. İstanbulkart'ın Özellikleri

İstanbulkart ulaşım ve ulaşım dışında kullanılabilir olması özelliği ile Türkiye’de herhangi bir alternatifi bulunmayan, aktif 18 milyon kullanıcıya sahip ve kart teknolojisi olarak temassız özelliği bulunan ön ödemeli bir karttır. İstanbulkart ürününe ait 9 çeşit kart tipi bulunmaktadır. Bu kart tipleri şunlardır:

a. İstanbulkart Tam: Herhangi bir temsilciden veya biletmatikten temin edilebilen bir karttır bu kart aynı zamanda “kişisel olmayan ulaşım kartı”dır.

b. İndirimli İstanbulkart: Öğrencilerin, öğretmenlerin ve 60 yaşını doldurmuş vatandaşların kullanabildiği, toplu taşımada indirimli tariflerden faydalanabilen kişisel kartlardır.

c. Ücretsiz İstanbulkart: Toplu taşımada ücretsiz seyahat edebilme imkanına sahip, içine TL yüklenerek ulaşım dışında harcamanın yapılabilirdiği kart tipidir. Bu kartlar 65 yaş üstü vatandaşlar, 0-4 yaş aralığındaki çocuk sahibi anneler, basın mensupları, zabıtarlar, emniyet personelleri, engelli vatandaşlar, er ve erbaşlar, gazi ve gazi aileleri, şehit aileleri, milli sporcular ve muhtarlar tarafından kullanılan kartlardır.

d. Mavi İstanbulkart: Abonman olarak kullanılan bu kart tipi herhangi bir indirim tabi olmayan kişisel bir karttır. Mavi kartlar; bireysel mavi kart, kurumsal mavi kart ve harcırah mavi kart olmak üzere üç gruba ayrılmaktadır. Bireysel mavi kart kişilerin bireysel başvuru yaparak sahip olduğı kartlardır. Kurumsal mavi kart, belirli kurumların personelleri için toplu olarak tanımlanan ve abonman özelliğı taşıyan kartlardır. Harcırah kartlar ise kurumların kendi personellerine vermek üzere BELBİM’den temin ettikleri, aylık 100 toplu ulaşım geçiş hakkının verildiğı kartlardır.

e. Sosyal Destek Kartı: İstanbul Büyükşehir Belediyesi tarafından ekonomik gelir düzeyi düşük vatandaşlar için üretilmiş kart tipidir.

f. Personel PDKS Kartı: Kurumların personelleri için kullandıkları devam kontrol sistemi kartıdır.

g. Dijital İstanbulkart: İstanbulkart’ın mobil uygulaması üzerinden oluşturulan, karekod okutularak işlem yapılabilen kartlardır.

h. Sınırlı Kullanılan İstanbulkart: BELBİM ile anlaşmalı herhangi bir temsilciden ya da biletmatiklerden temin edilebilen, sınırlı kullanım hakkına sahip

kartlardır. Bu kartlar bir geç, iki geç, üç geç, beş geç ve on geç olarak sınırlandırılmıştır (BELBİM, 2022).

2.2.3.3. İstanbulkart'a Ait 2021 Yılı İstatistikleri

İstanbulkart yukarıda da detaylı şekilde ifade edildiği üzere, oldukça geniş kullanım ağına sahiptir. Bu doğrultuda BELBİM'in 2021 yılı faaliyet raporlarına göre aktif İstanbulkart sayısı 27.538.454 adet, ulaşımda yapılan geçiş adedi 1.536.142.246, ulaşım dışında yapılan işlem adedi 22.560.571 olarak tespit edilmiştir. Yine aynı 2021 yılı faaliyet raporlarına göre İstanbulkart'ın 12.757 adet yükleme noktasının bulunduğu ve ilgili yılda yapılan toplam yükleme işlemi adedinin 220.460.038 olduğu belirlenmiştir (BELBİM, 2022).

2.2.4. Akıllı Ulaşım Kartlarının Kamu Güvenliğindeki Rolü

Şehir içi toplu taşımada kart kullanımının yaygınlaşması, yolcu güvenliğinin üst düzeyde tutulmasına katkı sağlamakta ve yine ulaşım kartlarında yer alan kişisel bilgiler birtakım suçların aydınlatılmasına yardımcı olmaktadır. Ulaşım kartı ile terminallerde yapılan her bir işlem kayıt altına alındığından, bu bilgiler gerekli görüldüğünde kart sağlayıcı kurum (İstanbulkart bakımından BELBİM) tarafından emniyet vb. kolluk kuvvetleri ve savcılık makamıyla paylaşarak suça konu olayın çözümlenmesi sağlanabilir. Örneğin, toplu taşıma aracında gerçekleşen hırsızlık, taciz ya da tehdit gibi bir olayın aydınlatılması noktasında müşterdeki şahsın hangi hat numaralı araca, hangi duraktan, saat kaçta bindiği gibi bilgiler şüpheli şahsın tespitine yardımcı olur.

Benzer şekilde, pandemi sürecinde kullanımı yaygınlaşan HES uygulamasının ulaşım kartlarında zorunlu hale gelmesi toplu taşımada kamu sağlığının korunmasına katkı sağlamıştır. Ulaşım kartına tanımlanan HES kodu, riskli görünen vatandaşların toplu taşıma kullanımını engellemiş ve böylece bu vatandaşların toplum sağlığını bozmalarının önüne geçilmiştir.

Ulaşım kartları kaybedildiğinde ya da çalındığında bu durum vatandaşlar tarafından fazla önemsenmese de içinde bulundurmuş olduğu kişisel bilgilerden dolayı bu kartlar aslında banka kartlarına benzemektedirler. Her ne kadar bu bilgiler kart sağlayıcı kurum tarafından korunmakta ise de ulaşım kartı kullanıcıları da kişisel bilgilerinin güvenliği açısından kendi kart güvenliklerini sağlanması konusunda gayret göstermelidirler. Kart çalındığında veya kaybolduğunda kart sağlayıcı kurum ile

iletişime geçerek kartı blokeye aldirmek, ulaşım kartlarını başka birisi ile paylaşmamak gibi önlemler buna örnektir.

2.2.5. Akıllı Ulaşım Kartlarında Yapılan Suistimler ve Hukuksal Boyutu

Suistimal kelimesi Arapça kökenli olup görev, yetki vb. nin kötüye kullanma anlamına gelmektedir. (TDK, 2022). Buradan hareketle ulaşım kartlarında kullanılan akıllı kartların suistimali bu kartların sağladığı imkanların dışında kötüye kullanılmasını ifade eder. Kart güvenliğinin korunması veya kişisel olmayan ulaşım kartının güvenliğinin sağlanması adına, kart sağlayıcı kurumlar birtakım senaryolar hazırlayarak, suistimali en aza indirmeye yönelik çalışmalarda bulunmaktadır. Bu kurumların hazırlamış olduğu senaryolarla kişisel veya kişisel olmayan ulaşım kartlarının terminallerdeki hareketleri izlenerek aykırı bir durum tespit edildiğinde, kart kara listeye alınır ve kullanıma kapatılır. Örneğin; bir ulaşım kartı ile bir günde hayatın olağan akışı içerisinde yapılabilecek harcama miktarından fazla ulaşım harcaması yapılması, aylık belirlenen ulaşım harcama tutarının çok üzerinde harcama yapılması, belirli bir yaşın üzerindeki (65 yaş üstü kartlar) kişilerin hayatın olağan akışı içinde yapabilecekleri harcama miktarından fazla ulaşım harcamaları yapmaları, bir kişinin aynı terminal üzerinde dolum, harcama ve iade gibi birden fazla işlem yapmış olması, yine bir kişiyi ulaşım kartıyla herhangi bir terminalde belirli bir süre içinde birden fazla kullanımda bulunmuş olması ya da imtiyazlı kartın belirli bir sürede harcama limitini aşacak şekilde sık kullanılması. Bunlar ve bunlara benzer senaryolarla kart hareketleri izlenerek ulaşım kartlarında yapılan suistimallerin en aza indirgenmesi sağlanmaktadır.

Ulaşım kartlarında gerçekleşen ve yukarıda izah edilen durumlar, belediyeler ve diğer kurumlar için ciddi bir finansal kayba neden olduğundan bahsi geçen işlem ve uygulamaların önüne geçilmesi gerekmektedir. Yukarıda yer verilen bazı ihtimaller aynı zamanda TCK'de düzenlenen suç tiplerinden biri veya birkaçı ile uyuşmakta, bu nedenle suça karışan kişiler ceza yargılaması ve kimi durumlarda da mahkumiyet kararıyla karşı karşıya kalmaktadırlar. Örneğin; Yargıtay 15.Ceza Dairesi'nin 19.09.2019 tarihli, 2017/11313 esas ve 2019/8602 karar sayılı kararına konu olayda, turnike görevlisi şahıs ile özel güvenlik görevlisinin geçiş yapmak için gelen yolculardan 3.343,50 TL ücret aldıkları ve 2229 adet geçiş sağladıkları, buna karşın görevli kent kartını kullanmayarak elde ettikleri bedeli de kurum kayıtlarına işlemedikleri tespit edilmiş ve mahkemece sanıklar hakkında güveni kötüye kullanma

suçu işledikleri gerekçesiyle mahkumiyet kararının onanmasına karar verilmiştir (Kazancı, 2019).

Yine benzer şekilde Yargıtay 5. Ceza Dairesi'nin 08.10.2019 tarihli, 2015/5933 esas ve 2019/9440 karar sayılı kararına konu olayda; sanık Metrobüs durağında yolculardan aldığı ücret karşılığında İstanbulkart kullanmak ve tahsil ettiği parayı ilgili kuruma vermekle görevlendirilmiş ancak sanık turnike kollarını dışardan yaptığı müdahale ile serbest bırakmış, yolculardan geçiş paralarını tahsil etmiş hatta bu paraları üçüncü kişi ile paylaşmış ancak kuruma vermemiştir. Sanıklar hakkında ilk derece mahkemesince beraat kararı verilmişse de eksik yargılama yapılmış olması nedeniyle Yargıtay 5. Ceza Dairesi tarafından beraat kararının bozulmasına karar verilmiştir (Kazancı, 2019).

Benzer şekilde, Yargıtay 13. Ceza Dairesi'nin 15.04.2014 tarihli, 2013/13961 esas ve 2014/14051 karar sayılı kararına konu olayda suça sürüklenen çocuğun arkadaşı adına düzenlenen ancak üzerinde kendisine ait resmin bulunduğu İstanbulkartı turnikelerden geçmekte olan vatandaşlara para karşılığında kullandığı ve bu yolla haksız menfaat elde ettiği yerel mahkemece tespit edilmiştir (Kazancı, 2014).

Yargıtay 15. Ceza Dairesi'nin 19.02.2014 tarihli, 2012/9345 esas ve 2014/3022 karar sayılı kararına konu olayda ise sanıklar hakkında nitelikli dolandırıcılık suçlamasıyla iddianame düzenlenmiş, suça konu olayda sanıkların özel halk otobüsü sahibi oldukları, sanıklardan birinin biletçilik, birinin şoförlük yaptığı, sahibi oldukları otobüste bir adet İstanbulkartın 03.01.2007-04.01.2007 tarihlerinde 26 kere kullanıldığının tespit edildiği, bu şekilde sanıkların otobüsüne yolcu binmediği halde İstanbulkartı kullanarak haksız kazanç sağladıkları ileri sürülmüştür. Sanıklar savunmalarında aracı kendileri dışında da başka şoför ve muavinlerin kullandığını bu kullanımlara ilişkin resmi kayıtların tutulmadığını atılı suçlamayı kabul etmediklerini beyan etmişlerdir. Mahkemece yapılan araştırma neticesinde de kartı kullananların kimlikleri kesin olarak tespit edilememiş, dolayısıyla sanıklar hakkında beraat kararı verilmiştir (Lexpera, 2014)

Yukarıda yer verilen Yargıtay kararları incelendiğinde görülmektedir ki, güvenlik görevlisi gibi, yetkili kurumca görevlendirilmiş kişilerin, turnike geçişlerinde üçüncü kişilerden para tahsil etmelerine karşın bu tutarı yetkili kuruma teslim etmemeleri güveni kötüye kullanma suçunu oluştururken, kimi durumlarda ulaşım

kartlarının haksız kazanç elde edilmesi amacıyla kullanılması dolandırıcılık suçu kapsamında sayılmaktadır. Ulaşım kartları aracılığıyla gerçekleştirilen bu işlem ve eylemlerin suç teşkil edebilmesi için TCK’de düzenlenen şartların gerçekleşmiş olması gerekmektedir. Dolandırıcılık kavramı TCK’nin 157. Maddesinde yer verilmiş olup ilgili madde hükmü; *“Hileli davranışlarla bir kimseyi aldatıp, onun veya başkasının zararına olarak, kendisine veya başkasına bir yarar sağlayan kişiye bir yıldan beş yıla kadar hapis ve beş bin güne kadar adli para cezası verilir.”* şeklindedir. Bu nedenle ulaşım kartları aracılığıyla gerçekleştirilen tüm suistimalleri TCK kapsamında değerlendirmek hukuken mümkün değildir. Bu yorumlamanın yapılması kuşkusuz mahkemelere aittir.

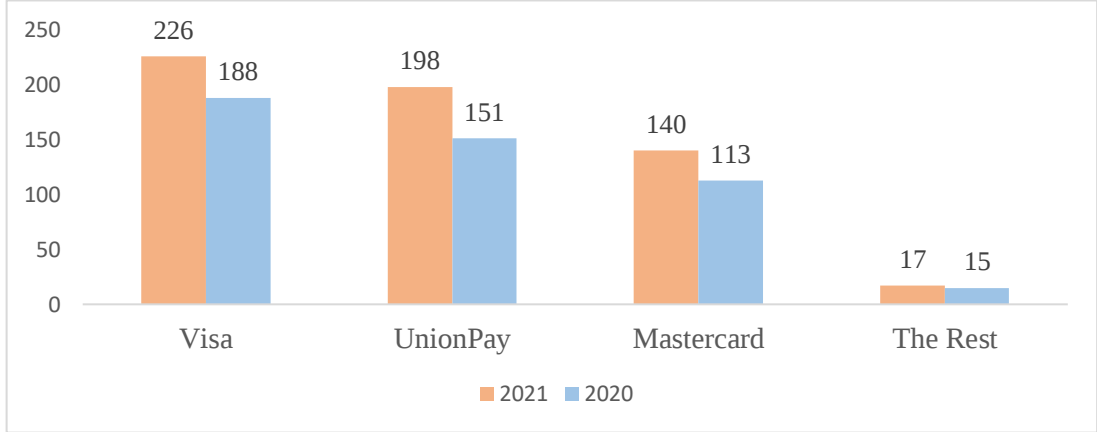
2.3. Kredi Kartı Dolandırıcılığı ve Bu Dolandırıcılık Türünün Ulaşımında Kullanılan Akıllı Kartlar ile Benzerliği

Bu çalışmanın konusunu her ne kadar ulaşım kartlarında yapılan suistimallerin tespiti oluşturmakta ise de ulaşım kartlarındaki suistimaller ile kredi kartlarında gerçekleşen suistimaller birbiriyle benzerlik göstermektedir. Bu benzerliklere çalışmanın ilerleyen kısımlarında değinilecek olup, şimdi kredi kartlarında gerçekleşen suistimaller üzerinde durulacaktır. Bununla birlikte kredi kartları üzerinde gerçekleşen suistimallerin çoğu dolandırıcılık kapsamına girdiğinden, bu başlık altında kredi kartı dolandırıcılığına değinilmiştir.

Bilindiği üzere son yıllarda kredi kartı kullanımı ciddi düzeyde artış göstermiştir ve göstermeye devam etmektedir. Ödeme sistemlerinde online ya da offline olarak birçok işlem kredi kart ile yapılmaya başlanmıştır. Ancak internet teknolojisinin hızla artmasıyla birlikte, banka veya kredi kartı dolandırıcılığı da endişe verici düzeyde artış göstermektedir. Bu nedenle banka ve kart kuruluş şirketleri kart dolandırıcılığını büyük bir tehdit olarak değerlendirmekte olup, bu dolandırıcılığı en az düzeye indirmek adına büyük bir güvenlik sistemi altyapısı için çalışmaktadırlar. Şüpheli işlemlerin hızlı tespit edilebilmesi hem kart kullanıcıları hem de bankalar için maliyet ve şirket itibarı için büyük öneme sahiptir (Itou, 2021).

Nilson Report (2022) verilerine göre 2021 yılında 581 milyar kart olduğu belirlenirken, 2020 yılında bu sayı 467 milyar olarak belirlenmiştir. Dünya çapında kart kuruluş şirketlerinin 2021 yılı dağılımlarını inceldiğimizde ise bu kartlarının %39’unun Visa, %34’ünün UnionPay, %24’ünün Mastercard ve %3’ünün The Rest

olduğu gözlenmiştir. Kart kuruluş şirketlerine ait Şekil 2.1’de 2020 ve 2021 yılı istatistikleri verilmiştir (Nilson Report, 2022).

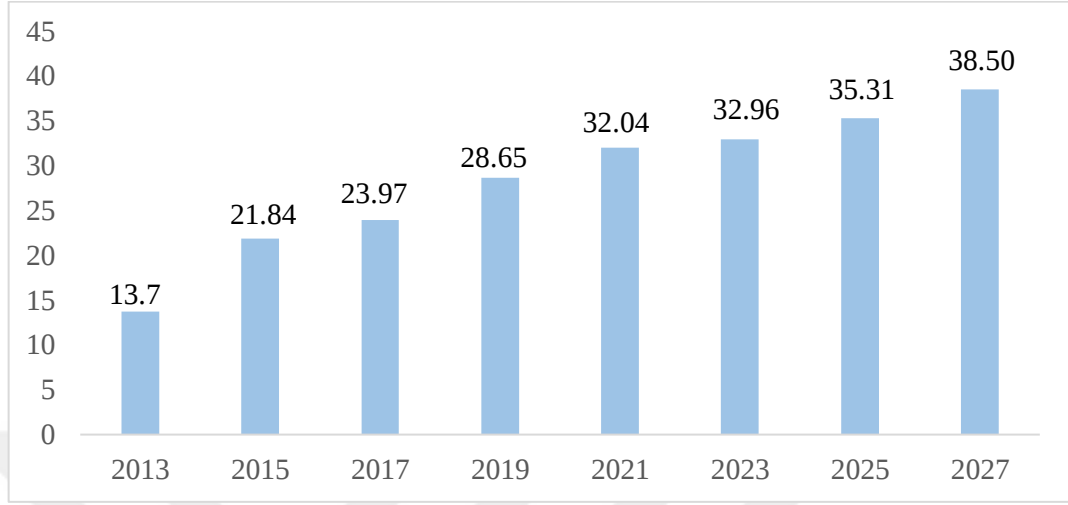


Şekil 2.1. Kart sağlayıcı şirketlerin 2020 ve 2021 yılı kart adetleri (milyar)

İnternette yapılan alışveriş miktarının artması, mobil bankacılığın artış göstermesi ve temassız ödeme gibi işlemlerin yaygınlaşması da yapılan işlem miktarının artmasına neden olmuştur. Kartlarda yapılan işlem miktarının artması kurumlar için veri miktarının artmasına ve bu verilerin güvenliğinin üst düzeyde sağlanmasına neden olmuştur. Kredi kartı dolandırıcılığı hem kart kullanıcıları hem de bankalar için büyük bir finansal kayba neden olduğundan, kredi kartı kullanım güvenliğinin en üst düzeyde tutulması adına bankalar ve kart kuruluş şirketleri verilerin güvenliğini sağlamak için teknolojik yatırımlarda bulunmaktadır. Ancak buna paralel olarak dolandırıcılar sistemlerdeki diğer boşlukları bularak yeni yöntemler geliştirmeye başlamış ve bu durum dolandırıcılık işlemlerinin tespit edilmesini zorlaştırmıştır. Ancak makine öğreniminin hayatımıza girmesiyle beraber gerçekleşen dolandırıcılık ve suistimal gibi işlemlerin tespit edilebilmesi kolaylaşmış; banka ve kart kuruluşları gibi şirketlerde sadece bu tür işlemlerin takibini sağlayan birimler oluşturulmaya başlanmıştır. Makine öğrenimi ile tahmine dayalı geliştirilen algoritmalar sistemlerde yaşanan bu tür işlemlerin tespit edilmesine büyük katkı sağlamaktadır.

Kredi kartı dolandırıcılığı dünyada milyarlarca dolar kayba neden olmaktadır. Nilson Report (2020) verilerine göre 2019 yılında dünya çapında kart tabanlı ödeme sistemlerinde brüt 28 Milyar dolar kayıp yaşanmıştır. Şekil 2.2’de gösterildiği üzere, 2013 yılından 2027 yılına kadar gerçekleşmiş kart tabanlı ödeme sistemi dolandırıcılığının finansal boyutu ve 2021,2023,2025 ve 2027 yıllarında yaşanması

öngörülen kart tabanlı ödeme sistemleri dolandırıcılığının finansal boyutu gösterilmiştir. Şekil 2.1 incelendiğinde bu tür vakaların endişe verici düzeyde artış gösterdiği görülmektedir (Nilson Report, 2020).



Şekil 2.2. 2013/2019 yıllarında dolandırıcılık sebebiyle kaybedilen ve 2021/2027 yılları arasında kaybedileceği tahmin edilen dolandırıcılık tutarları (milyar dolar)

Kredi kartı dolandırıcılığı kart sahibi ve bankaların bilgileri dışında herhangi bir yolla finansal kazanç elde etmek amacıyla yapılan tüm faaliyetleri ifade eder. Bu faaliyetler birden fazla yolla yapılabilir. Kartın çalınması, sahte kart üretimi, kullanıcı bilgilerinin çalınması vb. faaliyetler kredi kartı dolandırıcılığına örnektir (Jain, vd., 2019). Günümüzde bankalar ve kart kuruluş şirketleri tarafından bu tür vakaların önüne geçmek adına birtakım senaryolar oluşturularak kişilerin kart hareketlerini inceleme altına alınması söz konusudur. Ancak dolandırıcılığa maruz kalan her kartın fiziksel takibinin yapılması zordur ve eski yöntemlerin kullanılması yetersizdir. Bu nedenle kredi kartı dolandırıcılığı takibinde makine öğrenimi algoritmalarının kullanılması hem zaman hem de çalışmaların verimliliği açısından yarar sağlayacaktır. Ulaşımda kullanılan akıllı kartlara ilişkin açıklamalar kısmında da bu akıllı kartların sayıca fazla olduğuna yer verilmiştir. Bu nedenle aynı kredi kartı dolandırıcılığında olduğu gibi ulaşımda kullanılan akıllı kartlarda gerçekleşen suistimallerin takibi de önem arz etmektedir. Kredi kartı dolandırıcılığı takibinde kullanılan makine öğrenimi yöntemlerinin ulaşımda kullanılan akıllı kartların suistimallerinin takibinde de tercih edilmesi kuşkusuz yarar sağlayacaktır. Kaldı ki çalışmanın “Uygulamalar” başlığı doğrudan buna ilişkin olup ulaşımda kullanılan akıllı kartlarda gerçekleşen suistimleri makine öğrenimi algoritmaları kullanılarak tespiti yapılacaktır.

Son olarak kredi kartı dolandırıcılığıyla ilgili dolandırıcılık türlerine aşağıda yer verilmiştir.

2.3.1. Kredi Kartı Dolandırıcılık Türleri

a. Uygulama dolandırıcılığı: Dolandırıcılar kart sahibine ait kullanıcı adı ve şifre gibi bilgilere erişerek sistemin kontrolünü ele geçirir, dolandırıcı kart sahibinin kimliğine bürünür, böylece daha yüksek kredi limitine erişmeye çalışır ve kredi başvurusunda bulunur. Ayrıca dolandırıcı, sahte kredi başvurusunu kanıtlayabilecek belgeleri de ele geçirir.

b. Kredi kartının elektronik ya da manuel baskılanması: Dolandırıcı, kartın manyetik şeridinde yer alan bilgilere erişir ve sonrasında bu bilgileri kullanarak yeni bir sahte kart basar.

c. CNP card not present (kart mevcut değil): Dolandırıcı, kartın son kullanma tarihini ve hesap numarasını bildiğinde, kartın uzaktan herhangi bir ödeme terminaline teması olmadan yapılan işlemdir. Bu işlem genellikle telefon ve elektronik posta yolu ile gerçekleşir.

d. Sahte kart dolandırıcılığı: Kartın manyetiğinde bulunan tüm bilgiler kopyalanarak sahte bir kart çıkarılır, çip ya da PIN sisteminin desteklenmediği ülkelerde kart ile alışveriş yapılır ve bu kartlar ATM den nakit para çekme için de kullanılır.

e. Kayıp ve çalıntı kart dolandırıcılığı: Gerçek kart sahibinin kartını kaybetmesi ve dolandırıcının eline geçmesi durumunda kullanılan kartlardır. Ödeme işlemi sırasında kart şifresi gerekli olduğu için herhangi bir terminalde kullanılması zordur ancak internet alışverişlerinde dolandırıcılar için oldukça kolaydır.

f. Kart kimlik bilgilerinin çalınması: Bu dolandırıcılık türü bir kartı kullanmak ya da yeni bir hesap açmak için orijinal karta ait ayrıntılar elde edilerek oluşturulan kartlardır.

g. Basılan kartın kart sahibinin eline ulaşmaması: Kartı kullanacak kişinin kart başvurusunda bulunmasından sonra normal şartlarda kart, posta ile kart sahibine ulaştırılır. Bu dolandırıcılık türünde ise kart gerçek sahibine ulaşmadan başkasının eline geçmiş olur. Bu durum genellikle apartman, öğrenci yurtları, toplu konutlarda yaşayan kişilerde gerçekleşmektedir.

h. Hesabın ele geçirilmesi: Bu yöntem en çok kullanılan yöntemlerden biri olup; dolandırıcı gerçek kart kullanıcısının hesap bilgilerine ve ilgili belgelere erişerek hesabı kendi üstüne alabilir. Sonrasında kartın tedarik edildiği banka ile iletişime geçerek gerçek kart kullanıcısıymış gibi davranabilir ve hatta adres değişikliğinde bile bulunabilir. Dolandırıcının çıkardığı bu kart dolandırıcının adresine ulaşır ve bu kart kullanılmaya başlanır.

i. Sahte satıcı siteleri: Bu durum, müşterinin bilinen ve orijinal bir web sitesine çok benzeyen, dolandırıcılar tarafından oluşturulan sahte bir web sitesinde kişinin tüm kimlik bilgilerinin toplanarak kullanılan yöntemdir. Kişi bu tür web adreslerinde işlemlerin indirimli ya da ücretsiz yapılacağına aldanarak tüm kimlik bilgilerini bu siteye teslim eder ve bu bilgiler de dolandırıcıların eline ulaşmış olur.

j. Satıcı ile gizli anlaşma: Kart satıcısının (banka ya da kart üretici kurum) kart bilgilerini, kart sahibinin haberi olmadan, kasten başka biri ile paylaşması durumudur (Jain, vd., 2019).

Yukarıda yer verilen dolandırıcılık türlerinin bazıları ulaşımda kullanılan akıllı kartlarda da görülebilir.

2.3.2. Kredi Kartı Dolandırıcılığının Sınıflandırılması ile İlgili Örnek Çalışmalar

Kredi kartı dolandırıcılığı ile ilgili literatürde yapılmış pek çok çalışma bulunmaktadır. Bu çalışmalarda kullanılan makine öğrenimi algoritmaları ve bu algoritmaların değerlendirme metrikleri Tablo 2.1’de verilmiştir. Ulaşımda kullanılan akıllı kartlarda gerçekleşen suistimallerin tespitinde de bu çalışmalarda kullanılan algoritmalar ve değerlendirme metrikleri dikkate alınmıştır. Bu nedenle kredi kartı dolandırıcılığına ilişkin Tablo 2.1’de yer verilen örnek çalışmalar bu çalışma için oldukça önem arz etmektedir.

Tablo 2.1. Kredi kartı dolandırıcılığı ile ilgili yapılmış literatür çalışmaları

Yazarlar	Veri Seti	Algoritma	Değerlendirme Metrikleri			
			ACC	MCC	F1	AUC
Varmedja vd., 2019	European Cardholders	RO	0,999	-	-	-
		LR	0,974	-	-	-
		NB	0,992	-	-	-
		MLP	0,999	-	-	-
Dornadula ve Geetha, 2019	European Cardholders	LOF	0,899	0,017	-	-
		IF	0,901	0,104	-	-
		DVM	0,998	0,525	-	-

		LR	0,999	0,676	-	-
		KA	0,999	0,835	-	-
		RO	0,999	0,826	-	-
Sailusha vd., 2020	European Cardholders	RO	0,999	-	-	-
		AdaBoost	0,999	-	-	-
Lakshmi ve Kavilla, 2018	European Cardholders	LR	0,900	-	-	-
		KA	0,943	-	-	-
		RO	0,955	-	-	-
		RO	0,949	-	0,951	-
		NB	0,918	-	0,917	-
		LR	0,904	-	0,921	-
		DVM	0,939	-	0,934	-
		KNN	0,949	-	0,910	-
		KA	0,909	-	0,927	-
		GA	0,940	-	0,939	-
Jain vd., 2020	European Cardholders	RO	0,999	-	-	-
		KA	0,999	-	-	-
		XGBoost	0,999	-	-	-
		RO	0,962	-	-	-
		LR	0,947	-	-	-
		KNN	0,942	-	-	-
		DVM	0,938	-	-	-
		KA	0,908	-	-	-
		NB	0,937	-	-	-
Ito vd., 2020	European Cardholders	LR	0,959	-	0,909	0,918
		NB	0,915	-	0,789	0,829
		KNN	0,751	-	0,441	0,633
		KA	0,911	-	0,879	0,906
		RO	0,967	-	0,953	0,955
		LR	0,951	-	0,931	0,942
		NB	0,951	-	0,928	0,933
		DVM	0,934	-	-	-
Asha ve KR, 2021	European Cardholders	KNN	0,998	-	-	-
		YSA	0,999	-	-	-
		NB	0,905	-	-	-
		LR	0,998	-	-	-
		RO	0,999	-	-	-
		KNN	0,969	0,590	-	-
Selimoğlu ve Yılmaz, 2021	European Cardholders	ANN	0,999	-	-	-
		NB	0,982	-	-	-
		RO	0,978	-	-	-
		KNN	0,974	-	-	-
		LR	0,944	-	-	-
		NB	0,919	-	-	-
		SC	0,952	-	0,950	-
		RO	0,945	-	0,950	-
		XGBoost	0,945	-	0,950	-
		KNN	0,942	-	0,910	-
		LR	0,939	-	0,940	-
		GB	0,935	-	0,940	-
		MLP	0,932	-	0,930	-
		DVM	0,932	-	0,930	-
		KA	0,908	-	0,910	-
		NB	0,905	-	0,910	-

Tablo 2.1'e göre makine öğreniminde kullanılan sınıflandırma algoritmalarında genellikle rastgele orman, lojistik regresyon, Naif Bayes, k-en yakın komşuluk ve

karar ağaçları algoritmalarının tercih edildiği gözlenmektedir. Bununla birlikte kredi kartı dolandırıcılığında en çok tercih edilen veri setinin European Cardholders veri setinin olduğu görülmüştür.



3. MAKİNE ÖĞRENİMİ

Öğrenme, insanlar için kişiselleştirilmiş bir olgudur. Öğrenme sadece insanlarla sınırlı kalmayıp, en basit canlı türlerinin bile bu davranışı sergiledikleri bilinmektedir. Bu nedenle yaşamak ve öğrenmek, canlıların ömürleri boyunca bir arada götördükleri bir olgudur. İnsanlar ya da diğer canlılar gibi öğrenebilen makinelerin insan kılığında bir görüntüye sahip olması günümüz insanlığının en büyük övünç kaynağıdır (Mohammed, vd., 2016).

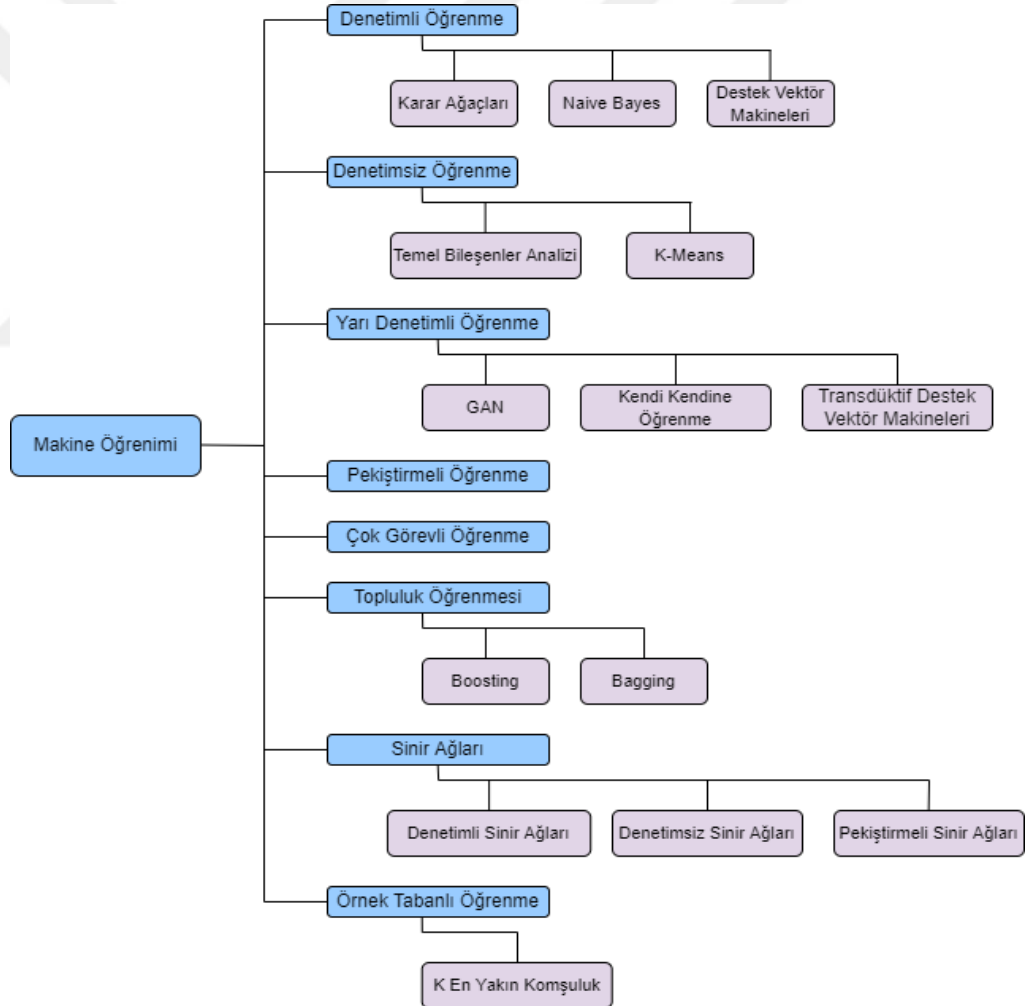
İnsanoğlu gündelik hayatta birtakım işleri daha basit yapabilmek için birçok araç türü kullanmıştır. İnsan beyninin yaratıcılığı farklı makinelerin icat edilmesine yol açmış, bu makineler endüstriyel ve bilgi işlem dahil olmak üzere pek çok yaşam ihtiyaçlarının karşılanmasını ve işlerinin kolaylaşmasını sağlamıştır ve makine öğrenimi de bunlardan biridir (Mahesh, 2020).

Makinelerin doğası gereği akıllı olmadığı bilindiği gibi bu makineler başlangıçta demiryolunda ilerleyen bir araç veya trafiğin akışını kontrol eden bir ışık olarak karşımıza çıkmış ya da hareket eden nesnelere ateş etmek gibi belirli görevleri yerine getirmek için tasarlanmıştır. Makinelerin görevlerini insanlara kıyasla daha hızlı ve daha yüksek bir hassasiyetle yapıyor olmaları insanların hayatını daha kolay hale getirmiştir. Ancak insan ile makine arasındaki temel fark zeka olup, makinelerde olmayan bu olgu bilgisayar biliminin gelişmesi ile makinelere aktarılmaya başlanmış ve günümüzde en zorlu matematik işlemlerini milisaniyelerde gerçekleştirebilen, kendi kendine strateji geliştirebilen robotlar üretilmiş ve bu robotlar makine öğrenimi ile öğrenerek günlük hayatımıza girmeye başlamıştır (Mohammed vd. 2016). Bununla birlikte çoğu kurum verileri depolayabilme sistemine sahip oldukları için eldeki mevcut verilerden anlamlı bilgiler çıkarmaya başlamış ve sektörde daha verimli çalışmaya başlamıştır. Veriler tek başına bir anlam ifade etmedikleri için veriye dayalı yöntemler ile yeni modeller geliştirilmiştir. Böylece geliştirilen bu modeller verilerin anlamlı hale getirilmesini sağlamışlardır (Ersöz ve Çınar, 2021).

Makine öğrenimi geniş kapsamlı uygulamalarıyla bilgisayar biliminin en hızlı büyüyen alanlarından birisidir. Makine öğrenimi araçları, programlara öğrenme ve uyum sağlama yeteneği kazandırmaktadır. Giderek artan veri miktarının kullanılabilir hale gelmesi, büyük veri setlerinden anlamlı bilgilerin elde edilmesi, makine öğrenimini bilgi teknolojilerinin temel dayanağı haline getirmiştir. Büyük veriden

anlamli bilgiler elde etme ve verilerde yer alan gizli kaliplari bulabilme merakini beraberinde getiren akilli ve nano teknolojilerin evriminin bir sonucu olarak veri madenciligi ve makine ogrenimi buyuk bir ilerleme kaydetmistir.

Istatistik, matematik ve bilgi teknolojileri makine ogreniminin ortaya cikmasini saglamistir. Her gezen gun veri boyutunun artmasi ve makine ogreniminin ogrenme surecinin karmaşik hale gelmiş olması, denetimli ve denetimsiz makine ogrenimi için çeşitli algoritmaların kullanılmasını gerek kılmıştır (Osisanwo, vd., 2017). Makine ogreniminde kullanılan algoritma türü çözmek istediğimiz problemin türüne, değişken sayısına ve ona en uygun modelin türüne bağı olarak değişim göstermektedir. Aşağıda Şekil 3.1’de makine ogreniminde yaygın olarak kullanılan bazı algoritmalar gösterilmiştir (Mahesh, 2020).



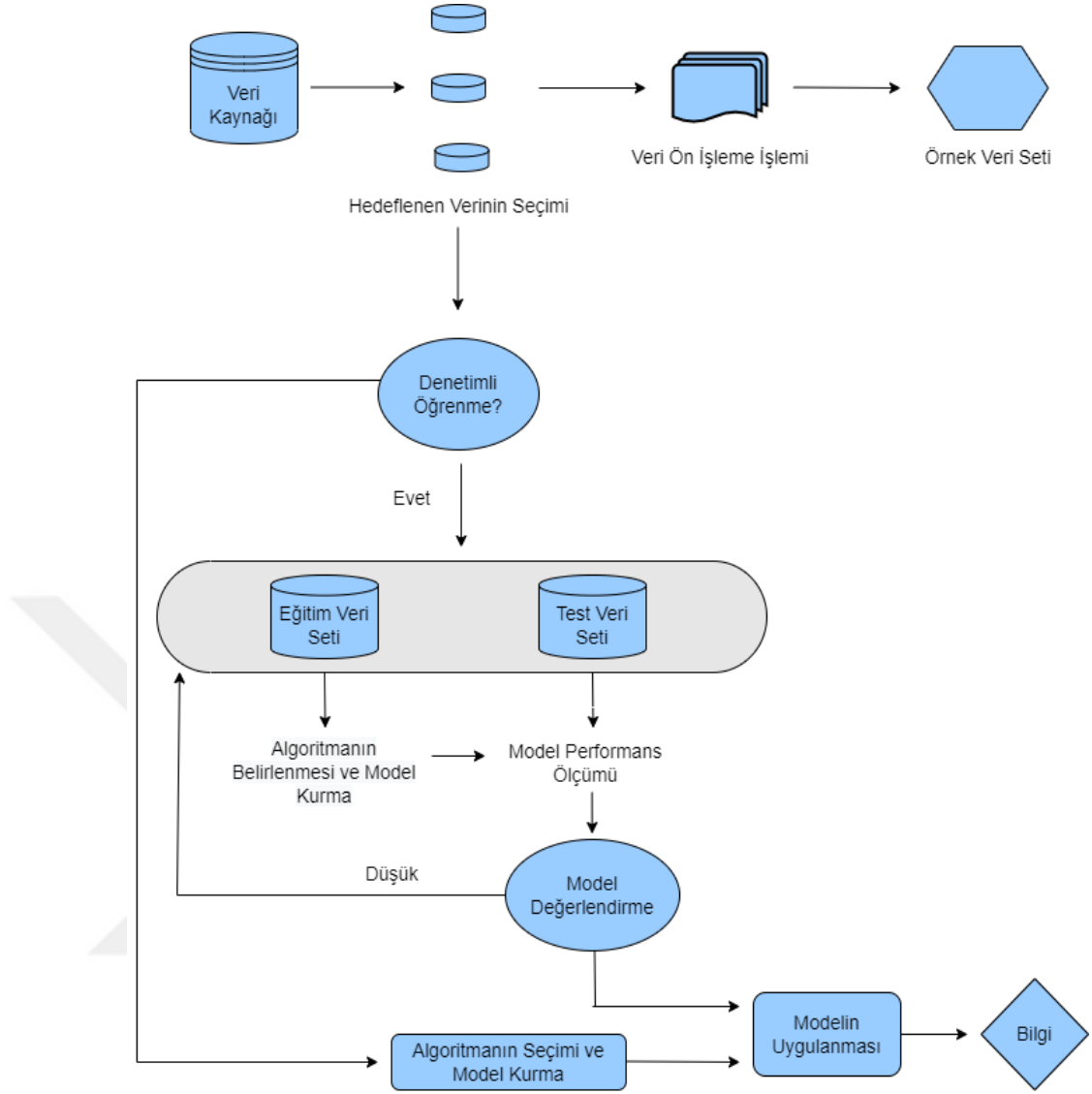
Şekil 3.1. Makine ogreniminde kullanılan algoritmalar (Mahesh, 2020)

Makine ogrenimi için birkaç uygulama bulunmaktadır ve bunlardan en önemlisi veri madenciligidir. İnsanlar analiz sırasında birden fazla özellik arasında ilişki

kurarken hata yapma eğiliminde olduklarından, bu tür problemlere çözüm bulmakta zorlanmaktadırlar. Makine öğrenimi bu tür problemleri başarıyla uygulayarak sistemlerin verimliliğini ve makinelerin tasarımını geliştirebilir. Makine öğreniminde kullanılan örnek veri setlerindeki her bir değişken aynı özellik kümesini kullanarak, bu özellik kümesini temsil etmektedir. Bu değişkenler sürekli, kategorik ya da ikili olabilir (Kotsiantis, vd., 2007). Makine öğreniminin büyük ve karmaşık verileri doğru tahmin edebilen bir algoritma olarak tanımlandığı çalışmalar bulunmaktadır. Ancak makine öğreniminin en temel görevi tüm sektörlerde sahip olan verilerin istatistik ve veri madenciliği tekniklerini kullanarak anlamlı bilgilerin ortaya çıkmasını sağlamaktır (Ersöz ve Çınar, 2021).

3.1. Veri Madenciliği

Veri madenciliği günümüzde finans sektöründen sağlık sektörüne kadar çoğu alanda problemlerin çözümü için kullanılmaktadır. Örneğin kampanya yapmak isteyen bir şirketin müşterilerini segmentlere ayırması, bir hastalığın teşhis edilmesi, müşteri sadakatinin artırılması, belirli bir ürünü satın alan tüketicilerin satın alma davranışlarının incelenmesi, elektrik üretim veya tüketiminin tahmin edilerek enerji talep tahminin yönetilmesi gibi alanlarda veri madenciliği teknikleri uygulanmaktadır. Veri madenciliğinin adımları Şekil 3.2’de gösterilmiştir:



Şekil 3.2. Veri madenciliği sürecine ait akış şeması (Balaban ve Kartal, 2015)

Yukarıdaki şemada anlatıldığı üzere, veri madenciliğinin ilk aşaması problemin tanımlanmasıdır. Çalışmanın hangi amaçla yapılacağı, sonuçlarının nasıl ölçüleceğine dair bilgiler iyi tanımlanmalıdır. Problem tanımlandıktan sonra mevcut veri kaynakları belirlenmelidir. Hedeflenen verinin seçimine başlamadan önce analist, gerekli alanların kontrolünü sağlamalıdır. Burada sadece verilerin depolanmış olması tek başına bir anlam ifade etmediği gibi anlamlı veri setinin oluşturulması için gerekli stratejilerin belirlenmesi de oldukça önemlidir. Doğru strateji belirlendikten sonra hedeflenen verilerin seçimi için gerekli işlemler gerçekleştirilir. Burada hedeflenen veri seti, veri ambarında tek bir tablodan oluşmuyorsa ilgili birleştirme işlemi ile birleştirme işlemleri gerçekleştirilir. Yeterli örneklem belirlendikten sonra veri seti analize hazır bir aşamaya getirilir. Bu aşamadan sonra analistin kuracağı modele göre

makine öğrenme algoritmalarından denetimli ya da denetimsiz öğrenme modelleri belirlenir. Eğer denetimli öğrenme seçilmeyecekse algoritma seçilip gerekli bilgiye ulaşmak için model kurulur. Eğer denetimli bir öğrenme algoritması kullanılacaksa veri seti, test ve eğitim seti olarak ayrılır.

Eğitim veri setinden seçilen algoritma ile model kurulur, test veri setinden ise modelin performans ölçütü belirlenir. Model performans ölçütü düşükse eğitim ve test veri seti tekrardan gözden geçirilir ve model yeniden kurulur. Model performans ölçütü yüksekse gerekli bilgiye ulaşmak için model çalıştırılır. Burada anlatılan süreç analistin gündelik hayatta kullanabileceği süreci içermektedir. Ancak bu süreçte analistin kayıp veya aykırı değerlerle karşılaşması muhtemeldir. Bu durumda analist sürece başlamadan önce bahsettiğimiz gibi anlamlı veri setine ulaşabilmesi için doğru stratejiyi belirlemesi gerekmektedir. Analistin doğru stratejisi sonucunda oluşturduğu veri setinde kayıp veya aykırı değerlerden oluşan veriler bulunuyorsa, analist tarafından bu değerlerin düzeltilerek modelin kurulması önerilir.

3.2. Makine Öğrenimi ile Veri Madenciliği Arasındaki Farklar

Veri madenciliği ve makine öğrenimi arasında kavramsal olarak ortak özellikler olsa da uygulama, teknik ve içerdiği özellikler bakımından birtakım farklılıklar bulunmaktadır. Bu farklılıklar şunlardır:

- Veri madenciliği istatistik, makine öğrenimi ve veri tabanı sistemlerine dayanarak, büyük veri kümesinde yer alan bilgileri ortaya çıkarmakta olup, makine öğrenimi algoritmalarını kullanarak sonuçları tahmin etmeye yönelik çalışmaktadır. Makine öğrenimi ise algoritmaların matematik ve programlama dillerine dayanarak ve mevcut verilerden kendi kendine öğrenerek, verileri tahmin eder ve verilerin nasıl öğretileceğine odaklanır.

- Veri madenciliği birçok makine öğrenimi algoritmalarından faydalandığı gibi makine öğrenimi dışında diğer teknikleri de kullanır. Makine öğrenimi algoritmaları ise denetimsiz öğrenme ya da veri ön işleme adımı gibi veri madenciliği tekniklerini kullanır ve makine öğreniminde kullanılan sınıflayıcı teknikler genellikle aynıdır.

- Veri madenciliğinde anlamlı bilgi elde etmek ve sonuçları tahmin etmek gibi makine öğrenimi araçlarından faydalanırken, makine öğreniminde kümeleme, sınıflandırma ve tahmin etme gibi veri madenciliği yöntemlerini kullanılır.

- Veri madenciliği makine öğrenimi için bir girdi kaynağı görevi görür ve veri setinden anlamlı bilgi çıkarmak için verileri derinlemesine inceler. Makine öğrenimi ise veri kümesini yinelemeli besleyerek, makineyi mükemmel hale getirmek için karmaşık algoritmalar geliştirir.

- Veri madenciliğinde anlamlı bilgi çıkarılmaya çalışıldığında insan müdahalesi gereklidir. Bu süreçte akıllı özellikler insan ile modele öğretilerek model akıllı hale getirilebilir. Ancak değişkenin değişmesi gibi durumlarda model de değişir. Makine öğreniminde ise algoritmalar sürekli olarak çalışır ve böylece model bir hatanın ne zaman ortaya çıkacağını da analiz edebilir. Burada değişkende değişiklik olduğunda, insan müdahalesine gerek kalmadan değişiklikleri yeniden programlayabilir.

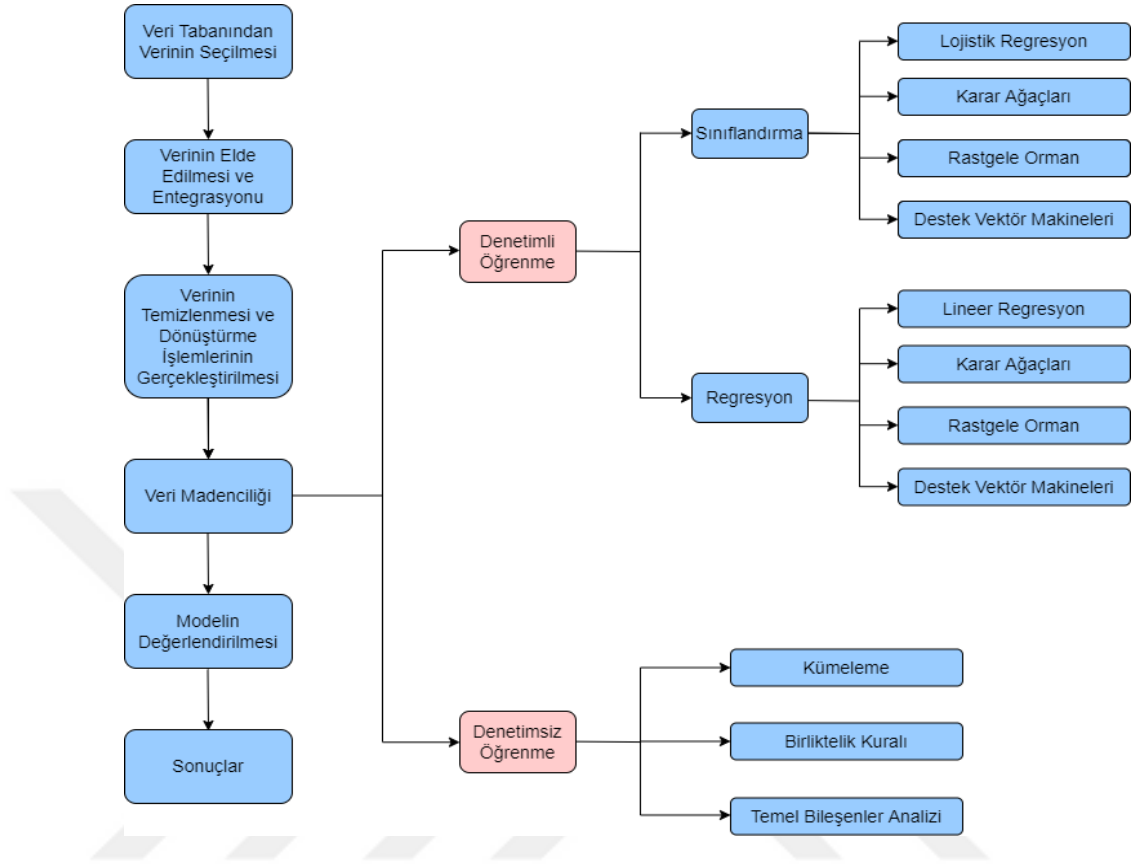
- Veri madenciliğinde kurulan bir model sonsuza kadar kullanılmaz, makine öğreniminde ise algoritma öğrendikten sonra insana ihtiyaç duymadan sonsuza kadar çalışabilir.

- Veri madenciliğinde aykırı değerler model performansını azalttığından verinin temizlenmesi oldukça önemlidir. Model kurulurken veriye uygun modelin kurulması durumunda sonuçlar güvenilir olur. Makine öğreniminde ise model performansı artırılırken model kendi kendine öğrenme algoritmalarını kullanır (Ersöz ve Çınar, 2021).

3.3. Denetimli ve Denetimsiz Öğrenme

Öğrenme problemleri, denetimli ve denetimsiz olmak üzere iki gruba ayrılmaktadır. Denetimli öğrenme, makine öğrenimi yöntemlerinde en çok kullanılan yöntemlerden biridir. Veri seti içinde yer alan değişkenlerin hangi girdi değerlerinin hangi çıktı değerlerini oluşturduğunun belli olduğu durumlarda denetimli öğrenme yöntemleri kullanılır. Denetimli öğrenmede tahminleme yapılırken regresyon ya da sınıflandırma yöntemleri kullanılmaktadır. Örnek olarak enerji üretim ya da tüketiminin tahmin edilmesi ya da kredi kartı kullanıcılarının kart hareketlerine göre sahtecilik yapıp yapmama durumlarının tespit edilmesi gibi durumlarda kullanılan öğrenme yöntemidir. Denetimsiz öğrenme ise etiketli veri olmadan veri seti üzerinde gerçekleşen öğrenme yöntemidir. Denetimsiz öğrenmede sadece girdi değerleri üzerinden çıkarım yapılmaya çalışılır ve bu yüzden bu tür veriler üzerinde değerlendirme yapmak oldukça zordur. Denetimsiz öğrenmede ise genellikle kümeleme yöntemleri kullanılmaktadır (Keskin ve Yıldız, 2018). Veri madenciliğinde

denetimli ve denetimsiz öğrenmeye ait algoritmalar ve bunlara ait süreçler aşağıdaki Şekil 3.3'te aktarılmıştır.

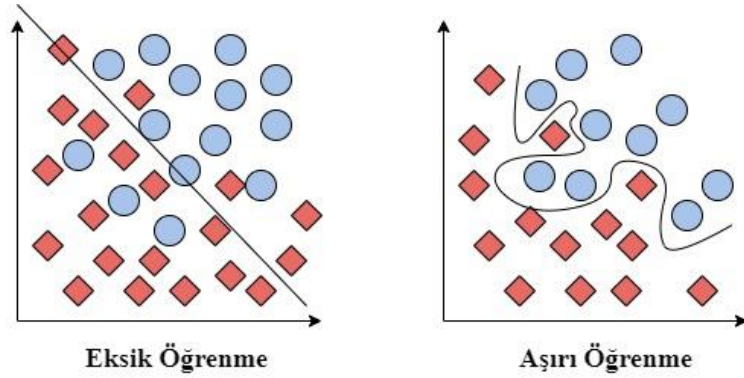


Şekil 3.3. Veri madenciliğinde denetimli ve denetimsiz öğrenme işlemleri (Wu, vd., 2021)

3.4. Aşırı Öğrenme (Overfitting) ve Eksik Öğrenme (Underfitting)

Aşırı öğrenme ve eksik öğrenme, modelin uyumunu irdeleyen ve modelin daha performanslı çalışması için gerekli aksiyonların alınmasını sağlayan kavramlardır. Eğitilen bir modelde çok fazla özellik varsa modelde aşırı öğrenme problemi ortaya çıkmaktadır. Yani modelin eğitim verilerine çok iyi uyum sağlayıp, test verilerinde genelleme yapamadığı durumlarda ortaya çıkmaktadır. Aşırı öğrenme durumlarında eğitim verilerinde yüksek tahmin doğruluğu olurken, aşırı öğrenme durumları yeni verilerde çok doğru tahminlerin yapılamamasına neden olmaktadır. Modelin daha az karmaşık hale getirilmesi ve veri miktarının artırılması aşırı öğrenme problemi için çözüm sağlayabilir. Model daha az özellik üzerinden eğitilmek istendiğinde ise makine bu model için ön yargılı olacaktır ve bu durum modelin eksik öğrenme problemini ortaya çıkaracaktır. Bu nedenle modelin daha fazla özellik üzerinden eğitilmesi bu

probleme çözüm sağlayabilir. Şekil 3.4'te aşırı öğrenme ve eksik öğrenmenin gösterimi yer almaktadır (Kolluri, 2020; Chaoji, 2016).



Şekil 3.4. Aşırı öğrenme ve eksik öğrenmeye ait gösterim (Kolluri, 2020)

3.4.1. Makine Öğreniminde Sınıflandırma Teknikleri

Sınıflandırma, verilerin grup üyeliğini tahmin etmek için kullanılan bir veri madenciliği (makine öğrenimi) tekniğidir. Makine öğreniminde sınıflandırma algoritmaları özellikle gelecek planlama ve bilgi keşfinin sağlanmasında en yaygın kullanılan tekniklerden biridir. Sınıflandırma, makine öğrenimi ve veri madenciliği araştırmacıları için en fazla çalışılan problemlerden biri olarak kabul edilmektedir (Soofi ve Awan, 2017). Makine öğreniminde kullanılan sınıflandırma algoritmaları geniş bir çalışma yelpazesine sahip olup yüksek boyutlu karmaşık veriler için daha yüksek doğruluk üretme eğilimindedirler (Maxwell, vd., 2018). Sınıflandırma algoritmalarının amacı girdi değişkenleri ile çıktı değişkenleri arasındaki ilişkiyi irdelemektir. Yapılan bu çalışmada ise sınıflandırma algoritmalarının performansları değerlendirilmiş ve ücretsiz bir istatistiksel yazılım programı olan R Studio programı ve bu yazılımın sahip olduğu caret kütüphanesi kullanılmıştır. Sınıflandırma algoritmalarında kullanılan birtakım algoritmalar bu çalışmada uygulanmış ve bu algoritmaların performansları karşılaştırılmıştır.

3.5. Çalışmada Kullanılan Algoritmalar

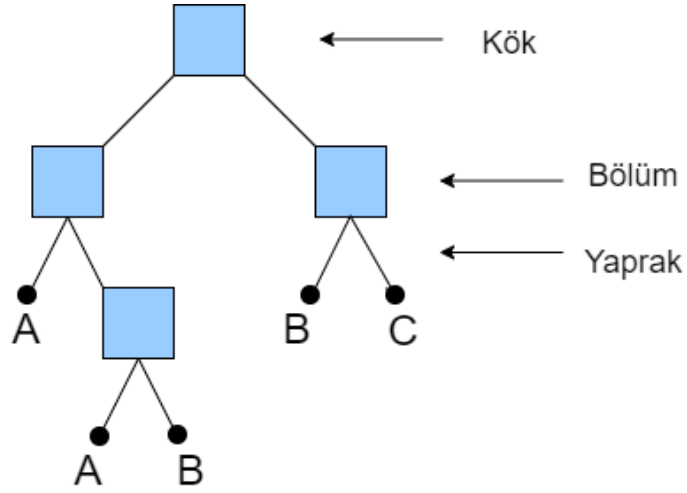
Çalışmanın bu bölümünde çalışmanın konusuna uygulanan algoritmalara yer verilecektir. Ancak bu algoritmalar makine öğreniminin sınıflandırma teknikleri arasında da yer alır. Dolayısıyla şimdi kısaca makine öğreniminde sınıflandırma tekniklerine ve ardından bu algoritmalara ilişkin açıklamalara değinilecektir.

3.5.1. Karar Ağaçları Algoritması

Karar ağacı yöntemi; sınıflandırma, tahmin, yorumlama ve veri manipülasyonu için kullanılan güçlü istatistiksel yöntemlerden biridir (Song ve Ying, 2015). Karar ağacı algoritması tüm veri öğelerini belirli bir sınıfa ait olana kadar derinlik ölçekli yaklaşım ya da genişlik ölçekli yaklaşım kullanarak, veri kümesini yinelemeli olarak bölümlere ayıran bir tekniktir. Bir karar ağacı sınıflandırma, özellik ve hedeflerine göre kök, karar düğümleri ve yapraklardan oluşur. Karar ağacında bulunan her bir karar düğümü karar kurallarını oluşturur (Jadhav ve Channe, 2016). Karar ağacı, ağaçtaki her bir karar düğümünde tanımlanmış bir dizi test veri setini yinelemeli olarak daha küçük alt bölümlere ayıran bir sınıflandırma sürecini barındırır. Bir veri seti algoritma tarafından tanımlanan karar çerçevesine göre sıralı olarak alt bölümlere ayrılarak sınıflandırılır ve her gözleme gözlemin düştüğü yaprak düğüme göre bir sınıf etiketi atanır. Karar ağaçları değişkenler ve sınıflar arasındaki doğrusal olmayan ilişkileri ele alarak eksik değerlere izin verirler ve hem nümerik hem de kategorik değişkenleri işleyebilme yeteneklerine sahiptirler. Karar ağaçları, değişkenler ve sınıflar arasındaki ilişkilerin rahatlıkla öğrenilebildiği yüksek kaliteli eğitim verilerine ihtiyaç duymaktadırlar. Bu yüzden karar ağaçlarının oluşması için sınıflandırılacak kitleyi temsil eden eğitim veri setine ihtiyaç vardır (Friedl ve Brodley, 1997)

Karar ağaçları sınıflandırma tekniği, ağaç oluşturma ve ağaç budama tekniği olarak iki aşamada gerçekleşmektedir. Ağaç oluşturma yukarıdan aşağıya doğru gerçekleşir ve bu aşamada ağaç tüm veri öğeleri aynı sınıf etiketlerine ait olana kadar yinelemeli olarak bölümlenir. Budama işlemi ise aşağıdan yukarıya doğru yapılır ve bu işlem modelin sınıflandırma doğruluğunu geliştirmek için kullanılır (Jadhav ve Channe, 2016).

Karar ağaçları kullanılabilirliği açısından kolay ve basit olmasından dolayı gerçek hayat problemlerinde sıklıkla kullanılan algoritmalarından biridir. Karar ağaçlarının kayıp verilerde ve çarpıklığın olduğu verilerde herhangi bir dönüşüme gerek duymadan kullanılabilir olmaları büyük avantaj sağlamaktadır. Ancak eğitim süresinin uzun olması ve küçük veri setleri kullanıldığında aşırı öğrenme gibi problemlerin oluşması modelin sağlamlığını etkileyen faktörler arasında yer almaktadır (Song ve Ying, 2015). Karar ağaçlarının yapısı Şekil 3.5'te gösterildiği gibidir.



Şekil 3.5 Karar ağacı algoritmasının yapısı (Friedl ve Brodley, 1997)

Karar ağaçları algoritmalarında kullanılan birçok yöntem bulunmaktadır. Bu çalışmanın uygulama bölümünde CART algoritması kullanılmıştır.

3.5.1.1. CART Algoritması

Karar ağaçlarında kullanılan CART algoritması sınıflandırma ve regresyon ağaçları olarak tanımlanmaktadır. CART algoritması gizli kalan veri ilişkilerini ortaya çıkaran karar ağaçlarında yaygın olarak kullanılan algoritmalarından biridir (Batra ve Agrawal, 2018). CART algoritması kategorik ve nümerik değişkenlerde kullanılabilen ve değişken seçimini otomatik yapabilen bir karar ağacı algoritmasıdır. CART algoritması dallanma işleminde Gini katsayısını kullanmaktadır (Gupta, vd., 2017).

3.5.2. Rastgele Orman Algoritması

Rastgele orman algoritması örnek veri setindeki verilerin torbalama (bagging) yöntemi ile eğitilerek, birden fazla karar ağacının birbirinden bağımsız olarak bir araya gelmesiyle oluşan denetimli makine öğrenme tekniklerinden biridir (Breiman, 2001). Her bir karar ağacı bootstrap yöntemi ile orijinal veri setinden rastgele örneklem seçerek her karar düğümüne tüm değişkenler için belirlenen sayıda rastgele değişken seçer ve rastgele ormanı oluşturur. Bu yöntemde oluşturulan karar ağaçları CART algoritması ile oluşur ve oluşan karar ağaçlarında budama işleminin yapılmasına gerek yoktur (Akman, vd., 2010). Her bir düğüm rastgele oluşturulurken rastgele seçilen alt küme arasından en iyi özelliği arar ve bu arama özelliği iyi bir modelleme oranını yakalayana kadar devam eder. Bunun sonucunda model en çok oyu alan sınıfı belirlemiş olur. Her bir düğüm için yapılan özellik seçimi rassal olduğu için ağaçlar

arasındaki korelasyon azdır ve bu durum modelin hata oranının azalmasını sağlar (Archer ve Kimes, 2007). Rastgele orman algoritması yüzlerce değişkeni silmeden özelliklerini sınıflandırabilir ve değişkenlerin sınıflandırmadaki önemini tahmin eder. Rastgele ormanın tek bir karar ağacından daha dayanıklı ve tahminleme performansı diğer denetimli makine öğrenimi tekniklerine göre daha güçlüdür (Jin, vd., 2018). Rastgele orman algoritması kategorik, sürekli ya da her ikisini bulunduran veri setleri ile çalışabildiği gibi hem büyük hem de küçük veri setleri için anlamlı sonuçlar verebilen bir algoritmadır. Rastgele orman algoritması sınıflandırma işlemi gerçekleştirilirken veri setinde yer alan kayıp verilere karşı dirençli bir yapıya sahiptir. Çok fazla değişkenin bulunduğu veri setleri için de kullanıma uygun bir algoritmadır. Rastgele orman algoritması birden fazla karar ağacından oluştuğu için görselleştirilmesi zor bir algoritma olmasıyla beraber her bir karar ağacını oluştururken çok fazla bellek kullandığı için düşük bellekli bilgisayarlarda uygulanması zor bir algoritmadır (Akman, vd., 2010).

Rastgele orman sınıflandırıcısı öznitelik seçim ölçütü olarak Gini katsayısını kullanmaktadır (Pal, 2005). Gini katsayısı her bir düğümde en düşük kirlilik ile bölünmeyi seçmek için kullanılır. Gini katsayısının yüksek olması düğümlerin ayrımı için oldukça önemlidir. Böylece belirli bir özelliğin kombinasyonu kullanılarak budama olmadan maksimum derinliğe ulaşmak için karar ağacı yapılır (Sonobe, vd., 2014).

Rastgele orman algoritmasına ait işlem süreci aşağıdaki gibidir.

a. Orijinal veri setinden n tane numune ile bootstrap yönteminin kullanılmasıyla örneklem seçimi yapılır.

b. Örnek kümesine ait rastgele ormanda CART algoritması ile ağaçlar oluşturulur. Ağaçların büyümesi sırasında aşağıdaki 1 ve 3'ncü adımları her bir düğüm için yinelemeli olarak tekrarlar. Böylece düğüm sayısı önceden belirlenmiş bir boyut sınırına ulaşır.

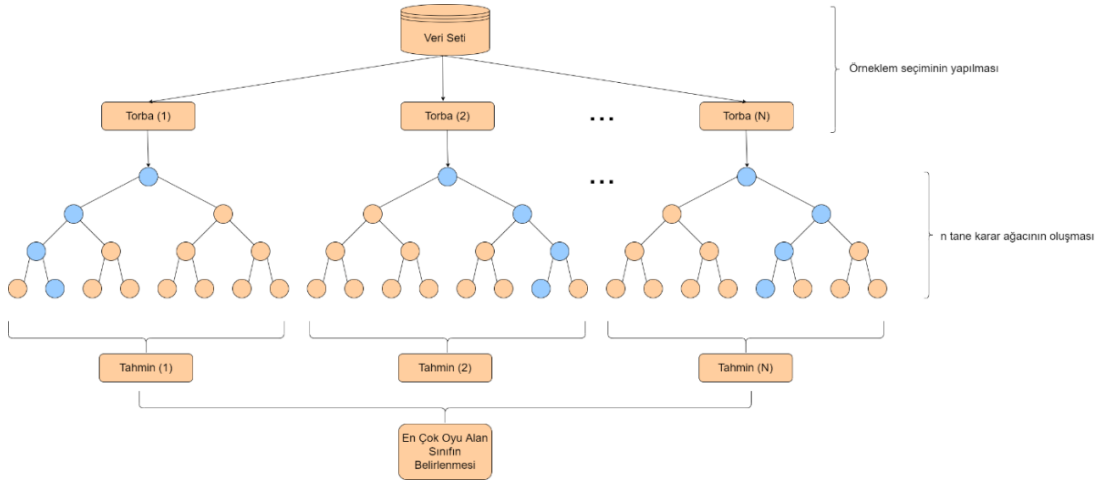
i. Tüm p değişkenlerinden rastgele r değişkeni seçilir ($r < p$)

ii. r değişkenlerinden en iyi bölme değişkeni ve bölme noktası seçilir,

iii. Mevcut düğüm seçilen bölme değişkenine ve bölme noktasına göre iki alt düğüme ayrılır.

c. Yaprak düğüm elde edilinceye kadar 1 ve 2'nci işlem devam eder.

d. Bu işlem sonucunda en çok oyu alan sınıf belirlenmiş olur. (Wang ve Wang 2020).



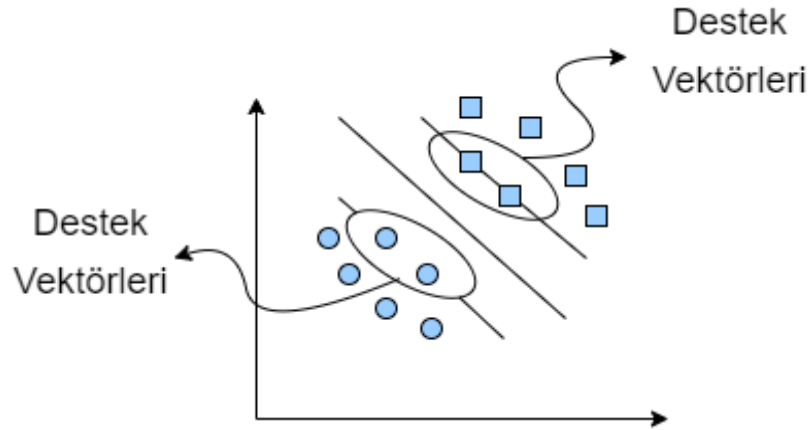
Şekil 3.6. Rastgele orman algoritmasının yapısı (Daneshvar ve Behnood, 2022)

Rastgele ormanda veri seti test ve eğitim veri seti olmak üzere iki bölüme ayrılır. Eğitim veri seti ön yükleme örnekleri ile oluşturulur. Out of bag (OOB) verisi olarak adlandırılan eğitim veri setinin kalan kısmını ise hataları test etmek için kullanılır. Bu adımlardan sonra CART algoritması ile geliştirilen ağaçlarda herhangi bir budama işlemi olmadan model devam eder ve her bir düğümde değişkenler tüm değişkenlerin arasından rastgele seçilir. Böylece bu değişkenler arasından en iyi dal belirlenmiş olur (Akar ve Güngör, 2012).

3.5.3. Destek Vektör Makineleri

Destek vektör makineleri Cortes ve Vapnik tarafından 1995 yılında ortaya atılan istatistiksel öğrenme teorisine ve yapısal risk minimizasyonuna dayanan, yüksek boyuttaki verileri hiper düzlemde iki ya da daha fazla sınıfa ayırabilen denetimli makine öğrenme algoritmalarından biridir. Destek vektör makinelerindeki amaç farklı sınıflara ait vektörler arasındaki mesafenin maksimize edilerek optimal ayırma hiper düzleminin belirlenmesidir. Destek vektör makinelerinin yapısal risk minimizasyonunu kullanması, verinin aşırı öğrenme problemlerine karşı güçlü olmasını sağlar (Güner ve Çomak, 2011). Destek vektör makineleri verinin boyutuna göre doğrusal, düzlemsel ve hiper düzlemde ayırabilme mekanizmaları ile veriyi iki ya da daha fazla sınıfa ayırabilme yeteneğine sahiptir (Güran, vd., 2014). Destek vektör makineleri sınıflandırma problemlerinde veriler sınıflandırılırken çekirdek fonksiyonları kullanılır. Veriler, doğrusal olarak ayrılabilen yüksek boyutlu uzaya taşınır. Hiper düzlemde ayrılan iki sınıfın birbirine en yakın olduğu noktalardan iki

doğru çizilir ve bu doğrular üzerinde yer alan nesnelere destek vektörleri adı verilir. Bunun gösterimi Şekil 3.7’de verilmiştir.



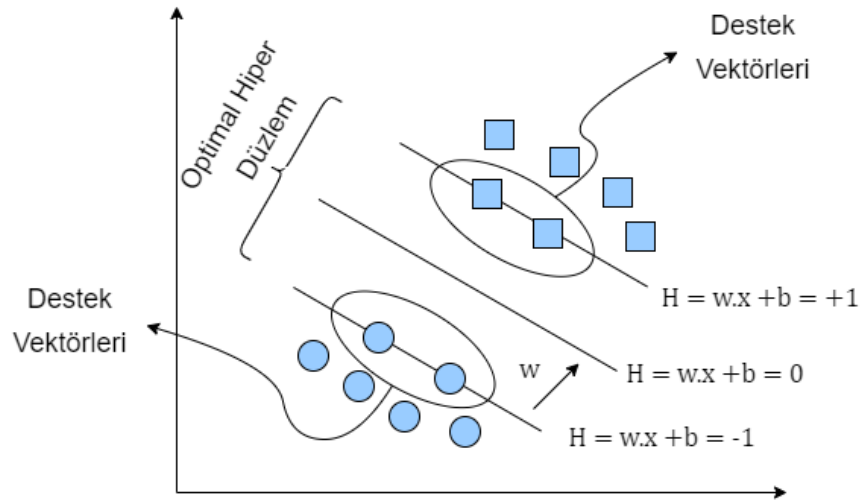
Şekil 3.7. Destek vektör makinelerine ait örnek grafik (Karagül, 2013)

Destek vektör makinelerinde çizilen düzlem ile destek vektörleri arasındaki marjin (uzaklık) maksimum olacak şekilde sınıflandırma işlemi gerçekleştirilir.

Destek vektör makineleri doğrusal ayrılma ve ayrılmama durumu olmak üzere iki alt başlıkta incelenmektedir.

3.5.3.1. Doğrusal Ayrılabilen Destek Vektör Makineleri

Doğrusal destek vektör makinelerinde destek vektörler hiper düzlemde problemsiz olarak dağılır. Her bir hiper düzlemin daha iyi ve daha doğru karar verdiğini görmek için maksimum marj sınıflandırıcısından faydalanılır. En doğru hiper düzlemin belirlenebilmesi için hiper düzlemin veri kümelerine eşit uzaklıkta olması gerekmektedir. Destek vektör makineleri sınıflandırma işleminde sınıflar, birbirinden ne kadar uzak olursa yani marjin genişliği ne kadar fazla olursa o kadar doğru sınıflandırma olur. Doğrusal ayrılabilen destek vektör makinelerinin gösterimi Şekil 3.8’de verilmiştir.



Şekil 3.8. Doğrusal ayrılabilen destek vektör makinelerinin gösterimi (Küçüksille ve Ateş, 2013)

Doğrusal olarak ayrılabilen iki sınıflı destek vektör makinelerinde, k sayıda örneklemden oluşan eğitim verisi $\{x_i, y_i\}$ ve $i = 1, \dots, k$ olarak kabul edilirse, optimal hiper düzleme ait eşitsizlik aşağıdaki gibi;

$$\begin{aligned} y = +1 \text{ için;} & \quad w \cdot x_i + b \geq +1 \\ y = -1 \text{ için;} & \quad w \cdot x_i + b \leq -1 \end{aligned}$$

Şeklinde olup, burada $x \in R^N$ N boyutlu uzay için $y \in \{-1, +1\}$ ise sınıf değerleri w ağırlık vektörünü ve b ise eğilim değerini göstermektedir. Tüm destek vektörlerindeki amaç düzlemi iki sınıfa bölüp örnek grubuna eşit uzaklıkta bulunmasını sağlamaktır. Grafik üzerinde yer alan $H = w \cdot x + b = -1$ ve $H = w \cdot x + b = 1$ destek vektörlerinin üzerinden geçen doğrular yumuşak ayrımı gösterirken $H = w \cdot x + b = 0$ doğrusu ise sert ayrımı göstermektedir. Bu işlemlerin sonucunda w değerinin minimum olması optimum hiper düzlemi sağlayacağı gibi marjin değerini maksimum yapacak ve en iyi ayrılmaya neden olacaktır (Küçüksille ve Ateş, 2013).

3.5.3.2. Doğrusal Olmayan Destek Vektör Makineleri

Gerçek hayat problemlerinde destek vektörler doğrusal düzlem üzerinde ayrılmadığında sınıflandırma işleminin gerçekleşmesi zorlaşır ve doğrusal sınıflandırıcılar yerine doğrusal olmayan sınıflandırıcılar kullanılır.

Person çekirdeği ise diğer çekirdek fonksiyonlarına göre daha karmaşık bir yapıya sahiptir. Person çekirdeğinin kullanımında en uygun parametre seçimi oldukça önemlidir. Person çekirdeğinin matematiksel gösterimi aşağıdaki gibidir;

$$K(x, y) = \frac{1}{\left[1 + \left(\frac{2\sqrt{|x-y|^2 \sqrt{2(1/w)} - 1}}{\sigma}\right)^2\right]^w} \quad (\text{Person genişliği parametreleri } (\sigma, w))$$

3.5.4. Naif Bayes Algoritması

Naif Bayes sınıflandırıcısı Bayes teoremine dayanan bir sınıflandırma algoritmasıdır. Naif Bayes algoritması, Bayes teoremine göre bir değişkenin birbirinden bağımsız olması gerektiği varsayımına dayanan ve koşullu sınıf olasılıklarına göre sonuç değişkenini tahmin eden istatistiksel bir sınıflandırma algoritmasıdır (Potur ve Erginel, 2021). Naif Bayes sınıflandırıcısı, Bayes teoreminin güçlü bağımsızlık varsayımlarına dayanan basit bir olasılıksal sınıflandırıcıdır. Değişkenlerin bağımsızlık varsayımları, değişken sırasının rassal olmasını sağlar ve sınıflandırma esnasında bir değişkenin mevcut olması diğer değişkenleri etkilemez ve bu varsayımlar bayes sınıflandırma algoritmasının hesaplamasını daha verimli hale getirir. Naif Bayes algoritmasında gerekli parametreleri tahmin edebilmek için az miktarda eğitim verisi çok verimli bir şekilde eğitilebilir. Naif Bayes sınıflandırıcıları çoğu zaman birçok karmaşık gerçek dünya problemlerinde beklenenden çok daha iyi çalıştığı ve bu sınıflandırıcının şaşırtıcı derecede iyi performans gösterdiği gözlenmiştir. Naif Bayes'in dezavantajı ise üstün performans gösteren sınıflandırma etkinliği ile destek vektör makineleri gibi diğer ayırt edici algoritmalara kıyasla daha düşük sınıflandırma performansı göstermesidir. Naif Bayes uzun yıllardır uygulanabilirliğinin kolay olması, hızlı çalışabilir olması ve iyi sonuçlar vermesi nedeniyle popüler makine öğrenme yöntemlerinden birisi olmuştur. (Khan, vd., 2010)

Naif Bayes algoritması aşağıdaki gibi formülize edilmiştir;

$$P(H|x) = \frac{P(H).P(x|H)}{P(x)}$$

Bu özelliğe göre Naif Bayes algoritması H sınıfının bilinmesi durumunda x değişkeninin var olma olasılığı ve x değişkeninin bilinmesi durumunda H sınıfının var olma olasılığı olarak adlandırılır. Bu varsayıma göre en iyi olasılık hangi sınıfa aitse o sınıf karar sınıfı olarak adlandırılır (Solmaz, vd., 2014).

3.5.5. XGBOOST Algoritması

XGBoost algoritması, temeli karar ağaçları algoritmasına dayanan, son zamanlarda regresyon ve sınıflandırma problemlerinde sıklıkla kullanılan, denetimli makine öğrenme algoritmalarından birisidir. XGBoost algoritması, sınıflandırma problemlerine iyi çözümler üreten ve modelin eğitilmesi sırasında aşırı öğrenme durumuna karşı model performansını artıran bir algoritmadır (Üstüne, vd., 2020). XGBoost algoritmasının birden fazla hiper parametresi bulunmaktadır. Algoritmanın parametreleri aşağıdaki gibidir:

- Gamma (Veri kaybını kısıtlar.)
- Min Child Weight (Bir yaprağın minimum ağırlık toplamını tanımlar.)
- Subsample (Her bir ağaçta değerlendirilecek özelliklerdir. Bu bölümde aşırı öğrenme engellenir ve modelin eğitim süresi kısalmır.)
- Colsample Bytree (Ağacın her seviyesinde bölme noktaları olarak kullanılacak değişkenlerin sayısını kontrol eder.)
- Max Depth (Maksimum dallanmayı temsil eder.)
- Eta (Modelin öğrenme oranı.)

Bu parametreler ne kadar büyük olursa algoritma o kadar güçlü olacaktır (Bardina, vd., 2020). XGBoost algoritmasının başarısındaki en önemli faktör diğer algoritma çözümlerinden 10 kat daha hızlı olması ve milyonlarca veriyi hafızasında ölçeklendirebilmesidir (Chen ve Guestrin, 2016).

3.5.6. Lojistik Regresyon Analizi

Lojistik regresyon analizi bağımsız değişkenler ile bağımlı değişken arasındaki ilişkinin analiz edilmesi için kullanılan regresyon yöntemlerinden biridir. Lojistik regresyonu doğrusal regresyondan ayıran en büyük özellik lojistik regresyonda bağımsız değişkenlerin kategorik veya ikili değişkenlerden oluşmasıdır. Lojistik regresyonda doğrusal regresyonun varsayımları olan normal dağılım, doğrusallık ve varyans-kovaryans matrisinin eşitliği gibi varsayımlar aranmaz. Lojistik regresyon en az değişkenle en iyi uyuma sahip bağımlı değişken ve bağımsız değişken arasındaki ilişkiyi açıklayabilen bir model kurmayı amaçlamaktadır (Çırak ve Çokluk, 2012).

Lojistik regresyonun parametreleri maksimum olabilirlik tekniği ile tahmin edilmektedir. Tahmin edilecek olan değer ile gözlem değerlerinin arasındaki olabilirliğin maksimum yapılması amaçlanmaktadır. Lojistik regresyonun gerektirdiği

varsayımlar; gözlem değerlerinin birbirlerinden bağımsız olması, bağımsız değişkenler ile bağımlı değişkenler arasında doğrusal bir ilişki olması ve bağımsız değişkenler arasında çoklu bağlantı probleminin olmamasıdır. Bu yöntem ile açıklanan kategorik bağımsız değişkenlerin 0 ya da 1 olması olasılığı hesaplanır ardından bağımlı değişkenin logaritmaları alınarak logit değişkene dönüştürülür. Böylece seçeneklerden herhangi birinin gerçekleşme oranının tahminleri yapılır (Özdemir, vd., 2011)

Lojistik Regresyon bağımlı değişkenin yapısına göre üç grupta incelenmektedir. Kategorik bağımlı değişkenin iki gruptan oluştuğu durumlarda ‘İkili Lojistik Regresyon Modeli’, kategorik bağımlı değişkenin üç ya da daha fazla olduğu durumlarda ‘Multinomial Lojistik Regresyon Modeli’, kategorik bağımlı değişkenin üçten fazla olduğu ve sıralı bir yapısı olduğunda ise ‘Sıralı (Ordinal) Lojistik Regresyon Modeli’ kullanılmaktadır (Şenel ve Alatlı, 2014).

Lojistik regresyon aşağıdaki gibi ifade edilmektedir;

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}$$

Bu eşitlikler aşağıdaki gibi ifade edilir;

- $\pi(x)$: İncelenen olayın gözlenme olasılığı,
- β_0 : Bağımsız değişken sıfır değerini aldığı anda, bağımlı değişken üzerindeki etkisi,
- $\beta_1, \dots, \beta_p$: Bağımsız değişkenlerin regresyon katsayıları,
- $X_1, \dots, X_p$: Bağımsız değişkenleri,
- p : Bağımsız değişken sayısını,
- e : 2,718 sayısını göstermektedir (Akyol, vd., 2012).

Bu çalışmada ikili lojistik regresyon kullanıldığı için matematiksel gösterimi aşağıdaki gibidir aktarılmıştır. Doğrusal regresyon modelinin genel olarak gösterimi,

$$E(y|x_{i1}, \dots, x_{ip}) = \sum_{k=0}^p \beta_k x_{ik}; i = 1, \dots, n \text{ için}$$

şeklinde olup, açıklayıcı değişkenler sabit varyanslıdır. Ayrıca y bağımlı değişkeninin sürekli olması koşulu vardır. Burada herhangi bir i 'nci gözlem değeri için aşağıdaki eşitlik sağlanır;

$$y_i = \sum_{k=0}^p \beta_k x_{ik} + u_i$$

Burada doğrusal regresyon modelinde bağımsız değişkenler üzerinde bir kısıt bulunmadığından y_i değeri $-\infty$ ile $+\infty$ aralığında değerler alabilir. Bağımlı değişken 0, 1 gibi değerler aldığı anda bu kural bozulmakta ve i 'nci gözlemin 1 değerini alma olasılığı $P(y_i = 1)$ olup beklenen değeri;

$$E(y_i) = P(y_i = 1) = \sum_{k=0}^p \beta_k x_{ik}$$

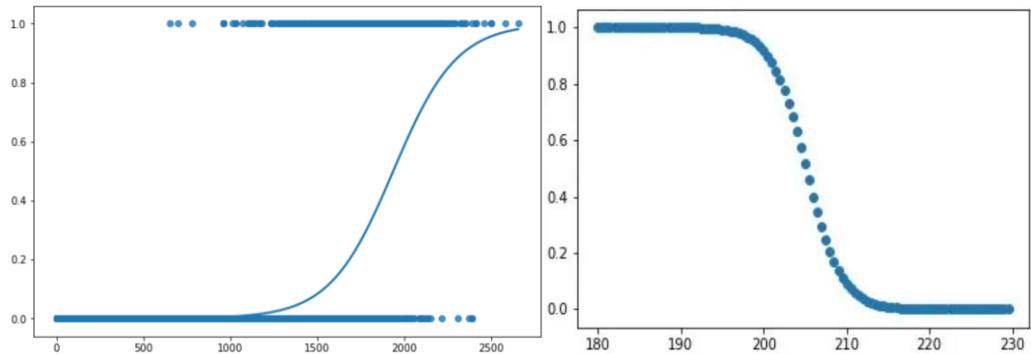
Olarak bulunur ve bu denkleme doğrusal olasılık modeli adı verilir. Açıklayıcı değişkenler $-\infty$ ile $+\infty$ arasında değerler aldığı için söz konusu eşitlik her zaman sağlanmaktadır. Bu nedenle burada logit ve probit dönüşümler uygulanmaktadır. Logit dönüşümde $P/1-P$ dönüşümü yapılmaktadır. Bu dönüşümden sonra elde edilen yeni fonksiyon;

$$E(y_i) = L_i = \log(P/(1 - P_i)) = \sum_{k=0}^p \beta_k x_{ik}$$

Olarak yazılır. Logit model olarak belirlenen bu modelin P_i olasılık değeri;

$$P_i = \exp\left(\sum_{k=0}^p \beta_k x_{ik}\right) / (1 + \exp\left(\sum_{k=0}^p \beta_k x_{ik}\right))$$

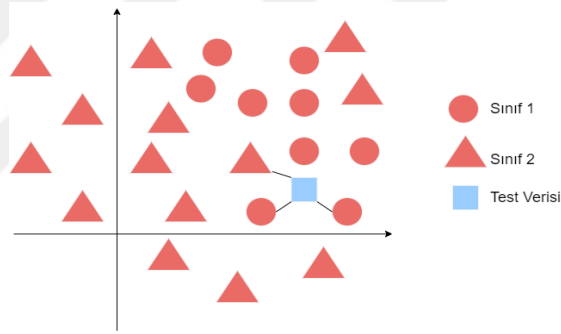
Olarak tanımlanır ve bu denkleme lojistik fonksiyon adı verilmektedir (Burmaoğlu, vd., 2009). Lojistik regresyon modeli grafiksel olarak, bağımlı değişken ikili sonuç verdiği anda cevap fonksiyon $P/(1 - P)$ şeklinin s ya da ters s şeklinde gösterilmesidir (Akyol vd. 2012).



Şekil 3.10. İkili bağımlı değişkenin s ve ters s şeklindeki olasılık fonksiyon grafikleri

3.5.7. K-En Yakın Komşuluk Algoritması

KNN algoritması sınıflandırma ve regresyon problemlerinde kullanılan basit bir denetimli makine öğrenme tekniğidir. KNN algoritması sınıfı belli olmayan bir verinin eğitim veri setindeki her bir örneğe olan uzaklığını hesaplar, hesaplanan uzaklığa göre veriye ait en optimal sınıf bulunur ve veri sınıflandırılır (Dilki ve Başar, 2020). Optimal sınıfa ait olacak verinin eğitim veri setindeki her bir sınıfa olan uzaklığı hesaplanır ve burada ifade edilen k sayısı sınıfın belirlenmesinde kullanılan en yakın komşu sayısını ifade etmektedir. Burada k değeri 1, 3, 5 gibi tek bir sayı olarak belirlenir. Burada optimal sınıfa ait olacak veri en yakın komşuların dahil olduğu sınıfa göre belirlenir. Şekil 3.11’de $k = 3$ için hesaplama yapılmış ve veriye ait en yakın 3 komşu belirlenmiştir. Burada veriye yakın 3 komşudan 2 tanesi birinci sınıfa yakın olduğundan verinin sınıf 1’e ait olduğu gözlenmektedir (Çavuşoğlu ve Kaçar, 2018).



Şekil 3.11. KNN algoritması sınıflandırma işlemi (Çavuşoğlu ve Kaçar, 2018).

KNN algoritmasının genel formülü aşağıdaki gibidir;

$$y(x_i) = \begin{cases} \infty, & \text{eğer } d(x_i, q) = 0 \text{ ise} \\ \frac{1}{d(x_i, q)}, & \text{aksi takdirde} \end{cases}$$

Bu denklemden x_i örnek uzayından seçilen rastgele örnekleme, d komşular arası uzaklığı ve q ise komşu ile sınıf arasındaki mesafenin tersini göstermektedir (Filiz, 2022.)

KNN algoritmasında optimal sınıf değerinin bulunabilmesi için mesafe ölçütleri kullanılır. Bu sayede sınıfı bilinmeyen verinin sınıfa olan yakınlığı hesaplanır. Bu uzaklık ölçütleri aşağıdaki gibidir;

a. Minkowski uzaklığı, $A = (a_1, a_2, \dots, a_n)$ ve $B = (b_1, b_2, \dots, b_n)$ sorgu vektörü olmak üzere Euclidean ve Chebyshev uzaklıklarının genelleştirilmiş halidir (Liu, vd., 2008).

$$D_{Min}(x, y) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p}$$

b. Euclidean uzaklığı, Minkowski uzaklığının özel bir ölçütü olup, $p=2$ olması durumudur.

$$D_{L2}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

c. Standardize edilmiş Euclidean uzaklığı, Euclidean uzaklığının normalizasyon işlemi uygulanmış veriler üzerinde hesaplanan mesafe ölçütüdür.

$$D_{SL2}(x, y) = \sqrt{\sum_{i=1}^n \frac{1}{V} (x_i - y_i)^2}$$

d. Mahalanobis uzaklığı, kovaryans matrisine dayalı normalleştirilmiş Euclidean mesafesine karşılık gelmektedir.

$$D_{Mah}(x, y) = \sqrt{\sum_{i=1}^n c^{-1} (x_i - y_i)^2}$$

e. Manhattan uzaklığı, Minkowski özelliğinin özel bir ölçütü olup, $p=1$ olması durumudur.

$$D_{Man}(x, y) = \sum_{i=1}^n |x_i - y_i|$$

f. Chebychev uzaklığı, Minkowski özelliğinin özel bir ölçütü olup, $p = \infty$ olması durumudur (Belattar ve Mostefai, 2015)

$$D_{Cheb}(x, y) = \max |x_i - y_i|$$

KNN algoritması kullanımının kolay olması, lineer olmayan eğitim süreçlerini içermemesi ve gürültülü verilere karşı oldukça başarılı sonuçlar vermesi nedeniyle büyük avantajlara sahiptir. Ancak veri seti boyutunun büyük olması yüksek düzeyde

güç kullanımına neden olduğundan algoritmanın çalışması zorlaşmaktadır (Pala, vd., 2019).

3.5.8. Adaboost Algoritması

Adaboost algoritması zayıf sınıflandırıcıların birtakım önemli özelliklerini yinelemeli olarak seçen ve güçlü bir sınıflandırıcı oluşturmak için doğrusal bir ağırlık kombinasyonu elde eden boosting yöntemidir. (Ma ve Du, 2010). Adaboost algoritması yineleme sürecinde bir önceki yinelemede doğru öğrenilmeyen örneklerle her turda daha fazla ağırlığın atandığı bir süreçtir (Javed, vd., 2020). Her zayıf sınıflandırıcı, girdi vektöründe yalnızca bir boyutun sınıflandırılmasına odaklanır. Algoritma eğitim süreci boyunca sınıflandırmanın genel doğruluk oranını iyileştirmek ve temel özelliklere odaklanmak için zayıf sınıflandırıcıların sayısını kendince uyarlayabilir. Zayıf sınıflandırıcı eklendikten sonra, algoritma bu zayıf sınıflandırıcının ağırlığını hesaplamak için minimum hatayı kullanır ve her eğitim örneğinin ağırlık değerini yeniden ayarlar, sonraki adımda da değeri bir sonraki zayıf sınıflandırıcıya iletir. Böylece yeni eklenen zayıf sınıflandırıcıya dayalı paralel sınıflandırıcının etkisi iyileştirilmiş olur (Wen ve Ding, 2013).

Adaboost algoritması kullanım kolaylığı açısından basit ve popüler bir algoritma olsa da bazı araştırmacılar gürültülü verilerde aşırı öğrenmeye eğilimli bir algoritma olarak nitelendirmiştir (Cao, vd., 2012). Ayrıca eğitim sürecinde her yinelemede yanlış sınıflandırılan örneklerin ağırlıklarını artırmaktadır. Örnekler sınıflandırma esnasında pek çok kez yanlış sınıflandırılabilir ve bu durum örneklerin ağırlıklarının sürekli artmasına neden olmaktadır. Bu durumda örneklerin ağırlık dağılımında ciddi bir dengesizlik problemi ortaya çıkar (Tang, vd., 2020).

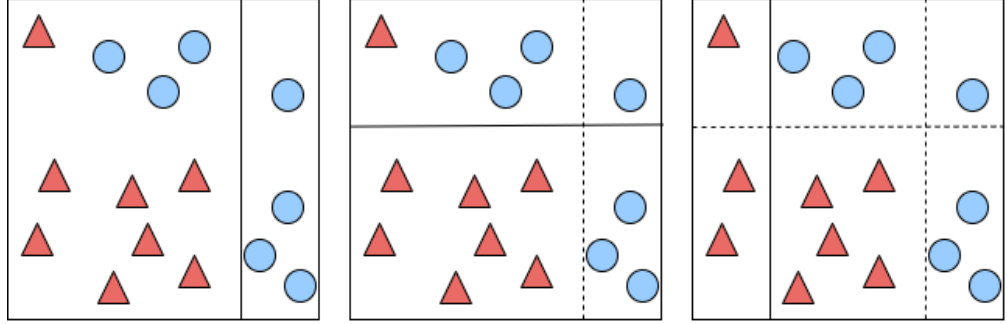
Adaboost algoritmasının işlem süreci aşağıdaki gibidir:

- a. Eğitim verisi üzerinden dağılımın başlaması $D_1(i) = 1/m$
- b. $t = 1, \dots, T$:
 - Zayıf sınıflandırıcının eğitilmesi D_t
 - Ağırlığın belirlenmesi $\alpha_1 \in R$
 - Eğitim setinde dağılımın güncellenmesi:

$$D_{t+1}(i) = \frac{D_t(i)e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$

- c. Nihai ağırlıklı oyun belirlenmesi:

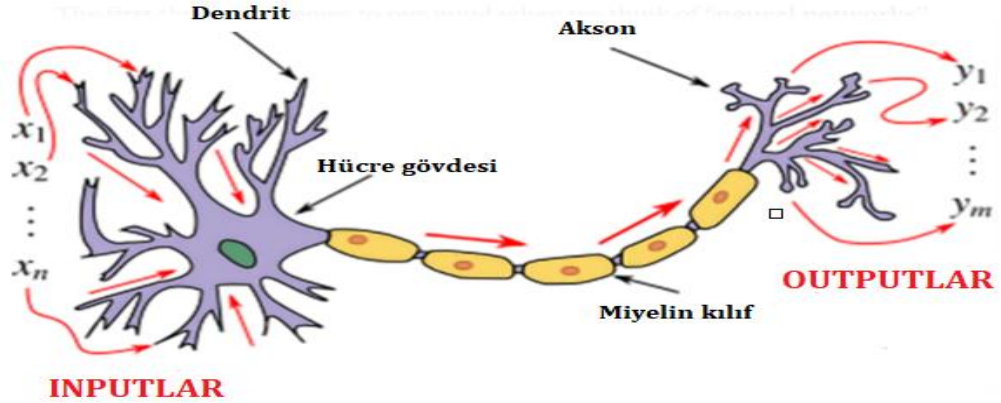
$$H(x) = \text{sign}(f(x)) = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right)$$



Şekil 3.12. Adaboost algoritmasının gösterimi

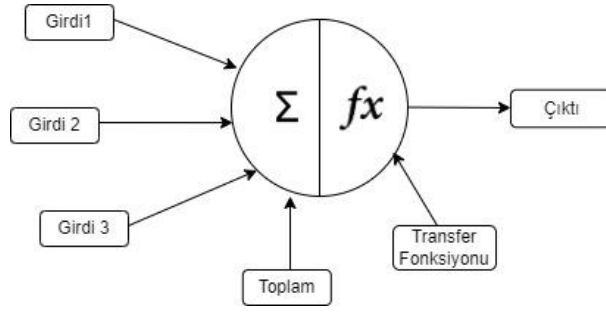
3.5.9. Yapay Sinir Ağları ve Derin Öğrenme

Yapay sinir ağları temeli biyolojik sinir sistemine dayanan basitleştirilmiş bir modeldir. İnsan beyinde bulunan nöronlar dendrit adı verilen yapıyla sinyalleri toplar, dendritler de nöronlardan gelen elektrokimyasal sinyalleri sinaptik eklemler yoluyla alır ve nöronlara iletir. Çıkış kanalları da aksonlar olarak adlandırılır ve bir akson dendritlere sinaptik eklemler aracılığıyla bağlanır (Bhagya ve Dash, 2022).



Şekil 3.13. Biyolojik yapay sinir ağının gösterimi (Wikipedia, 2022)

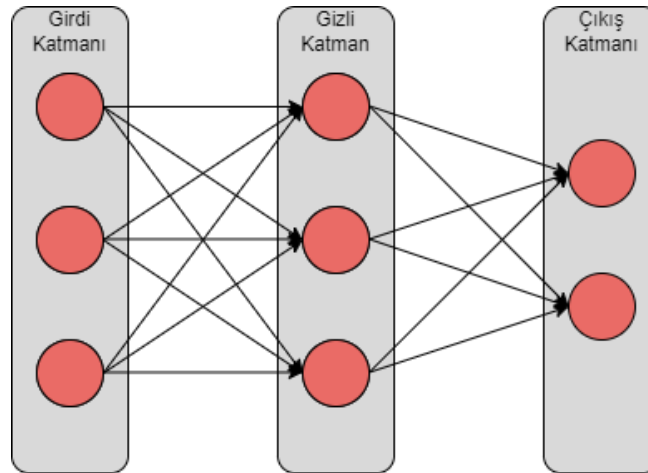
Yapay sinir ağlarının temel yapı taşı yapay nöronlar olup biyolojik nöronların basitleştirilmiş bir halidir. Yapay nöronlar girişlerinde bulunan bilgileri içeri alır, içeri alınan bu bilgilerin her biri belirlenen bir ağırlık değeri ile çarpılarak yeni değerler türetilir ve bu değerler de toplanarak transfer fonksiyonları ile işlenir, işlenen bu bilgiler yapay bir nöronla iletilir. Bunun basit bir gösterimi Şekil 3.14’de verilmiştir (Da Silva, vd., 2017; Suzuki, 2011).



Şekil 3.14. Yapay nörona ait gösterim (Suzuki, 2011)

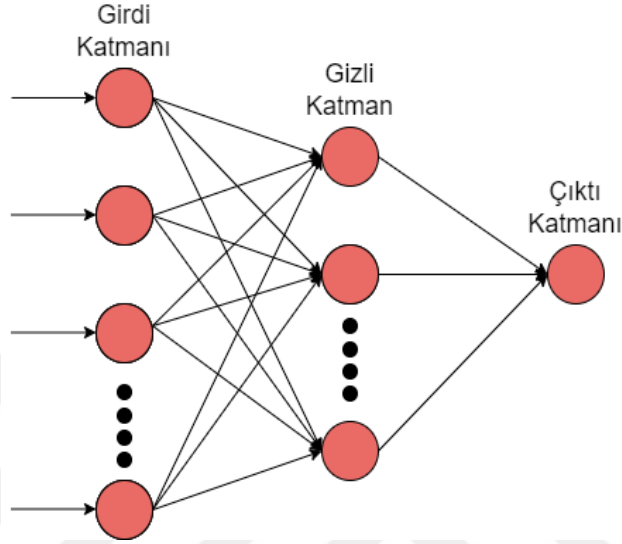
Yapay sinir ağları insan beyninin çalışma prensibine göre farklı problemleri çözmek için benzer öğrenme sürecini kullanmaktadır. Yapay sinir ağları veriye dayalı bir model olduğundan öğrenme sürecinden oluşan ve girdi değişkenleri ile çıktı değişkenleri arasındaki ilişkiyi güncelleyen bir modeldir. Yapay sinir ağları kısa sürede hesaplama yapma, yüksek doğruluk oranı verme, girdi ve çıktı değişkenleri arasındaki ilişkileri doğrusal olmayan bir şekilde haritalama yeteneğine sahiptir. Bu nedenle popüler bir algoritma haline gelmiştir. Yapay sinir ağları sınıflandırma, kümeleme, tahmin, örüntü tanıma ve görüntü işleme gibi alanlarda sıklıkla kullanılmaktadır (Jawad, vd., 2021).

Yapay siniri ağları giriş katmanı, gizli katman ve çıkış katmanı olarak üç ana katmandan oluşmaktadır. Giriş katmanı dışarıdan gelen bilgi, sinyal veya özellik gibi girdilerden oluşmaktadır. Gizli katman analiz edilme sürecini ve sisteme ilişkin kalıpları ortaya çıkarmaktadır. Bu katmandaki bütün nöronlar diğer nöronlar ile bağlantılıdır. Çıkış katmanı ise önceki katmanlarda bulunan nöronlardaki bilgiyi aktaran katmandır (Da Silva, vd., 2017).



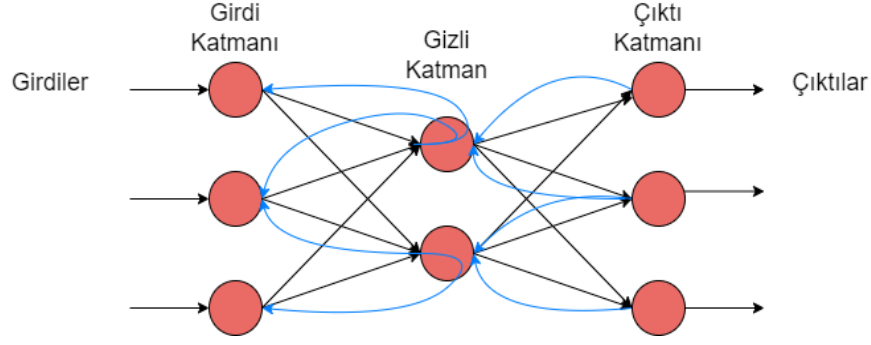
Şekil 3.15. Basit bir yapay sinir ağına ait katmanlar (Jawad, vd., 2021)

Yapay sinir ağları ileri beslemeli ve geri beslemeli olmak üzere ikiye ayrılmaktadır. İleri beslemeli sinir ağlarında Şekil 3.15'te görüldüğü gibi, bilgiler girdi katmanından çıktı katmanına doğru geri döngü olmadan sadece tek yönlü akmaktadır. Katman sayısı ve türü konusunda herhangi bir kısıtı olmayan bu ağların, bir katmanında yer alan her bir birim diğer tüm birimler ile ilişkilidir. İleri beslemeli sinir ağlarına ait görsel aşağıdaki gibidir (Suzuki, 2011; Abiodun, vd., 2018).



Şekil 3.16. İleri beslemeli yapay sinir ağına ait örnek gösterim

Geri beslemeli yapay sinir ağlarında ise bilgi girişten çıkışa ve çıkıştan girişe doğru olmak üzere iki yönlü ilerlemektedir. Bu yöntemde modelin mantığı, modelin yanlıştan ve doğrudan nasıl öğrendiğine dayanmaktadır. Yani model bilgiyi öğrenmek için önce ileri doğru akar, sonrasında ise bilgiyi doğrulamak için geriye doğru akmaktadır. Bu yöntem ağın etkili bir şekilde öğrenmesine neden olarak, kesin ve doğru çıktının üretilmesi sağlamış olur (Abiodun, vd., 2019). Bu yöntemde nöronlar geri besleme (geri bildirim) çıktısı olarak kullanılmaktadır. Bu model zaman serisi tahmini, sistem optimizasyonu, banka kredilerinin değerlendirilmesi gibi alanlarda sıklıkla kullanılmaktadır (Da Silva, 2017).



Şekil 3.17. Geri beslemeli yapay sinir ağına ait örnek gösterim

3.5.9.1. Derin Öğrenme

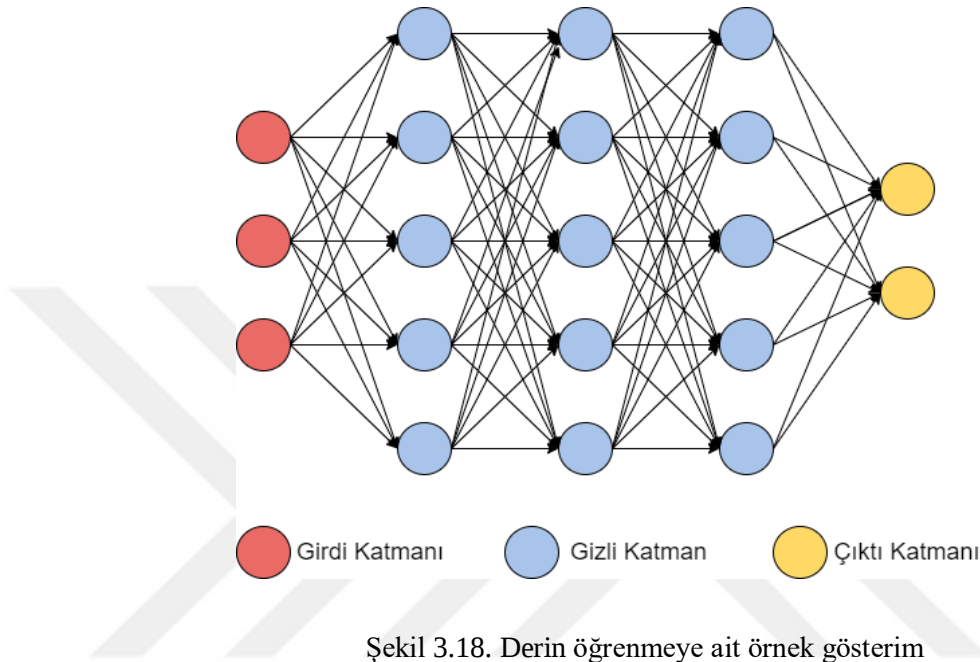
Derin öğrenme, makine öğreniminin bir alt kümesi olup, çok sayıda katman ve parametreye sahip, çoğu yöntemde sinir ağına ait mimarileri kullanan bir algoritmadır. Derin öğrenme; özellik çıkarma ve dönüştürme için doğrusal olmayan işlem birimlerinin çok katmanlı bir dizisini kullanır. Derin öğrenmede alt katmanlar basit özellikleri öğrenirken üst katmanlar alt katmanların özelliklerinden türetilen daha karmaşık özellikleri öğrenir. Bu nedenle derin öğrenme hem büyük miktardaki verileri hem de farklı kaynaklardan toplanan verileri analiz ederek yararlı bilgileri ortaya çıkarmak için kullanılan etkili yöntemlerden biridir (Shide ve Shah, 2018).

Hiyerarşik model açısından bakıldığında makine öğreniminin gelişimi 1980'lerin sonlarından bu yana basit öğrenme ve derin öğrenme olarak iki adımda gerçekleşmeye başlamıştır. 1980'lerin sonlarında geriye yayılım algoritmasının icadı, makine öğreniminde istatistiksel modellere dayalı öğrenme algoritmalarının temelini oluşturmaya başlamıştır. Geri yayılım algoritması, yapay sinir ağı modelinin birçok eğitim verisinden örüntüler öğrenmesine ve bu örüntüye dayalı tahminler yapmasına olanak sağlamıştır.

İnternetin hızla yayılması verilerin analizi ve tahmin edilmesi için büyük bir talebin oluşmasına neden olmuştur. Basit öğrenme, bu anlamda reklam arama sistemleri, spam filtreleme sistemleri ve öneri sistemleri gibi alanlarda büyük başarılar elde etmeye başlamıştır. Geoffrey Hilton tarafından 2006 yılında yayınlanan bir makalede daha fazla girdi katmana sahip bir sinir ağının, daha iyi öğrenme yeteneğine sahip olduğu tespit edilmiştir. Bu anlamda derin öğrenme ile ilgili çalışmaların yapılması devam etmiş ve Google, Microsoft gibi büyük teknoloji şirketleri tarafından

büyük verilerden bilgi zenginliği elde etmek adına derin öğrenmeye ilişkin kaynak ayrılmaya başlanmıştır.

Derin öğrenme pek çok akademik çalışmada ilgiyi üzerine çekmeye başarmış ve endüstriyel uygulamalarda da büyük ilerlemeler kaydedilmiştir. Bu durum da yapay zeka ve insan beyninin etkileşiminin ilerlemesine neden olmuştur (Hao, vd., 2016).



Şekil 3.18. Derin öğrenmeye ait örnek gösterim

Derin öğrenme algoritmasında kullanılan katmanlar aşağıdaki gibidir:

Giriş katmanı: Bu katman ham veri girişinin yapıldığı katmandır. Bu katmanda veri boyutunun büyük olması modelin başarı performansını etkilemektedir. Veri boyutunun büyük olması kurulacak olan ağı derinliğini ve performansını artırmaktadır. Ancak modelin eğitim ve test süresinin uzaması ile beraber yoğun bellek kullanılmasına neden olmaktadır (Özkan ve Ülker, 2017).

Konvolüsyon katmanı: Dönüşüm işleminin yapıldığı katmandır. Bu aşamada araştırmacı tarafından önceden belirlenmiş olan filtre, veri setindeki belirgin özelliklerin ortaya çıkarılmasını sağlar ve ham veriden daha düşük boyutlu bir veri seti oluşturulur. Burada araştırmacının oluşturduğu filtre ağı eğitim ve test süresinin başarısını etkilemektedir (Doğan ve Türkoğlu, 2019).

Aktivasyon fonksiyonu (The Rectified Linear Unit): Bu katman sinir ağları için kullanılan güçlü biyolojik ve matematiksel temellere dayanan bir aktivasyon fonksiyonudur. Modelde bu aktivasyonun kullanılması modelin daha hızlı

öğrenmesine neden olmaktadır (Agarap, 2018).

Havuzlama katmanı (Pooling Layer): Havuzlama katmanı, verilerin küçük bir kayıpla boyutunu indirgeyerek daha iyi doğruluk elde edilmesini sağlar. Bununla birlikte ağ modelinin de daha hızlı çalışmasını sağlayarak hesaplama için gerekli bellek sayısını azaltan önemli bir katmandır (Doğan F. Ve Türkoğlu İ. 2019). Havuzlama katmanı verimli bir eğitim süreci için oldukça önemlidir (Jie ve Wanda, 2020).

Tam bağlı katman (Full Connected): Tam bağlı katmanlarda tüm nöronlar bir önceki katmanlardaki tüm aktivasyonlarla tam bağlantılara sahiptir. Tam bağlı katmanlar modelde iki boyutlu özellik vektörlerini bir boyutlu özellik vektörlerine dönüştürür ve böylece türetilen vektörler ya sınıflandırma için belirli bir kategoriye dönüştürülebilir ya da bir sonraki işlemler için özellik vektörüne dönüştürülebilir (Voulodimos, vd., 2018).

Dropout katmanı: Dropout katmanı modelin aşırı öğrenme problemini önleyen ve ağda ezber yapan düğümlerin silinmesini sağlayarak ağın ezber yapmasını önleyen bir katmandır (Doğan ve Türkoğlu, 2019).

Sınıflandırma katmanı: Sınıflandırma katmanı sınıflandırılması yapılacak öge sayısı kadar sınıf üreten katmandır (Doğan ve Türkoğlu, 2019).

3.6. Değişken Seçimi

Son otuz yılda yüksek boyutlu verilerin makine öğrenimi ve veri madenciliği yöntemleri kullanılarak analiz edilmesi ciddi artış göstermiştir. Verilerin boyutu arttıkça bu durum mevcut öğrenme algoritmalarında da ciddi zorluklar ortaya çıkarmaya başlamıştır. Veri setinde çok fazla özelliğin bulunması öğrenme modelinin performansını olumsuz yönde etkilemesine neden olmaktadır. Bu sorunu çözmek için araştırmacılar boyut indirgeme teknikleri üzerinde çalışmışlardır. Değişken seçimi de araştırmacılar için en çok kullanılan boyut indirgeme tekniklerinden biridir. Değişken seçimi modelin öğrenme performansını daha iyi arttıran ve modelin daha iyi yorumlanabilir olmasını sağlayan tekniklerden biridir. Birçok sınıflandırma probleminde en iyi sınıflandırıcıların tespit edilmesini sağlamak oldukça zordur. Bu nedenle alakasız değişkenlerin sayısını azaltmak, öğrenme algoritmasının süresini azaltacağı gibi sınıflandırıcılar için daha iyi fikirler edinmemize yardımcı olur. Değişken seçimi sınıflandırmada eğitim aşamasını etkilemektedir. Değişkenler

oluşturulduktan sonra tüm değişkenleri modele almak yerine değişken seçimi ile veri setinin alt kümesini modele alarak veriler öğrenme sürecine dahil edilir (Tang, vd. 2014). Değişken seçimi, filtreleme (filter), sarmal (wrapper) ve gömülü (embedded) yöntemler olmaz üzere üç başlıkta toplanabilir. Filtreleme yöntemi her bir girdi değişkeni ile hedef değişken arasındaki ilişkiyi değerlendirerek skorlar ve bunların ilişki düzeyini değerlendirir. Sarmal model girdi değişkenlerinin alt kümelerinden birden fazla model kurar ve en iyi performans gösteren değişkeni belirler. Gömülü model ise değişkenlerin etkileşimlerini dahil eder. Bu yöntem model eğitim sürecinin her yinelenmesini ele alır. Bu çalışmada sarmal yöntemlerden olan boruta yöntemi kullanılarak değişken seçimi yapılmıştır. Boruta yönteminin çalışma prensibi ise aşağıda anlatılmıştır.

3.6.1. Boruta ile Değişken Seçim Yöntemi

Boruta algoritması, R programında, randomForest kütüphanesinde bulunan ve temeli rastgele orman sınıflandırma algoritmasına dayanan bir algoritmadır. Boruta veri setine rastgelelik ekleyerek ve rastgele örnekler topluluğundan sonuçlar toplayarak rastgele dalgalanmaların ve korelasyonların yanıltıcı etkilerini azaltabilen bir algoritmadır. Boruta algoritmasında kullanılan bu ekstra rastgelelik hangi niteliklerin gerçekten önemli olduğu konusunda bilgi verilmesini sağlayan değişken seçimi yöntemlerinden biridir. Boruta algoritmasının çalışma prensibi aşağıdaki gibidir;

- a. Tüm değişkenlerin kopyaları alınır ve veri seti genişletilir.
- b. Korelasyonların kaldırılması için eklenen değişkenler kendi içinde karşılaştırılır.
- c. Genişletilmiş veri setinde rastgele orman algoritması çalıştırılır ve Z skor puanı hesaplanır.
- d. Gölge değişkenlerin içinden maksimum Z skor puanına ait değişken bulunur ve bu değişkenden daha yüksek puana sahip gerçek değişkenler işaretlenir.
- e. Bütün değişkenler, maksimum skorlu gölge değişken ile karşılaştırılır.
- f. Maksimum Z skorlu gölge değişkenden önemli ölçüde daha düşük öneme sahip değişkenler önemsiz olarak kabul edilir ve veri setinden kalcı olarak kaldırılır.
- g. Maksimum Z skoruna sahip gölge değişkenden önemli ölçüde daha yüksek öneme sahip değişkenler önemli olarak kabul edilir.
- h. Tüm gölge değişkenler veri setinden silinir.

i. Tüm değişkenler için önem düzeyi artıncaya kadar veya önceden belirlenmiş rastgele orman sınırlarına ulaşıncaya kadar prosedür tekrarlanır (Kursa ve Rudnicki, 2010).

3.7. Sınıflandırma Algoritmalarında Performans Ölçütleri

Sınıflandırma algoritmalarında, bir algoritmayı değerlendirmek veya farklı algoritmaların performanslarını karşılaştırmak için uygulanabilecek birtakım performans ölçütleri bulunmaktadır. İkili ve çok sınıflı sınıflandırma problemlerinde kullanılan hata matrisi Şekil 3.19’ da gösterildiği gibidir. Matrisin her sütunu, tahmin edilen bir sınıfın örneklerini temsil ederken; her bir satır ise gerçek bir sınıfın örneklerini temsil etmektedir (Markoulidakis, vd., 2021). Sınıflandırma problemlerinde sınıf sayısına göre iki tür sınıflandırma problemi bulunmaktadır. Bunlar sadece iki sınıfın olduğu ikili sınıflandırma ve sınıf sayısının ikiden fazla olduğu durumlarda ise çok sınıflı sınıflandırmadır. Bu çalışmada iki sınıf kullanıldığı için ikili sınıflandırma üzerinden ilerleyecek olursak; ikili sınıflandırmada pozitif sınıflandırma için P , negatif sınıflandırma için N ; bilinmeyen bir örnek P veya N olarak sınıflandırılır. Hata matrisinin öğelerini temsil eden dört adet olası çıktı bulunmaktadır. Burada mavi köşegen doğru tahminleri, kırmızı köşegen ise yanlış tahminleri temsil etmektedir.

	Pozitif (P)	Negatif (N)
Doğru (D)	Doğru Pozitif (DP)	Doğru Negatif (DN)
Yanlış (Y)	Yanlış Pozitif (YP)	Yanlış Negatif (YN)
	$P = DP + YN$	$N = YP + YN$

Şekil 3.19. Performans ölçütlerinin şekil üzerinde gösterimi

Yukarıda görüldüğü üzere örnek pozitifse ve pozitif olarak sınıflandırılmışsa, doğru sınıflandırılmış pozitif örnek söz konusu olur. Gerçek pozitif ise negatif sınıflandırılmışsa, yanlış negatif veya 2. Tip hata olarak kabul edilir. Örnek negatif ve negatif sınıflandırılmışsa gerçek negatif olarak sınıflandırılırken; pozitif olarak sınıflandırılmışsa yanlış pozitif veya 1. Tip hata olarak kabul edilmektedir (Tharwat, 2020).

Doğruluk oranı: Modelin başarısını ölçmek için en çok kullanılan doğruluk ölçütüdür. Bu değer doğru tahmin edilen değerlerin tüm durumlara bölümünden ortaya çıkar (Sokolova, vd., 2006).

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

Duyarlılık (TPR) ve belirginlik (FPR) (sensitivity ve specificity): Bir sınıflandırıcının farklı sınıflardaki performansını ayrı ayrı tahmin eden ölçüttür.

$$\text{Duyarlılık} = \frac{DP}{DP + YN'}$$

$$\text{Belirginlik} = \frac{DN}{YP + DN}$$

Kesinlik (Pr): Tahmin edilen pozitif bir sınıf örneğinin doğru bir pozitif olma olasılığıdır.

$$\text{Kesinlik} = \frac{DP}{DP + YP}$$

F skoru: Duyarlılık ve kesinlik değerlerinin harmonik ortalamasıdır. β değeri ne kadar yüksekse duyarlılık değeri de o kadar kesindir (Maratea vd., 2014).

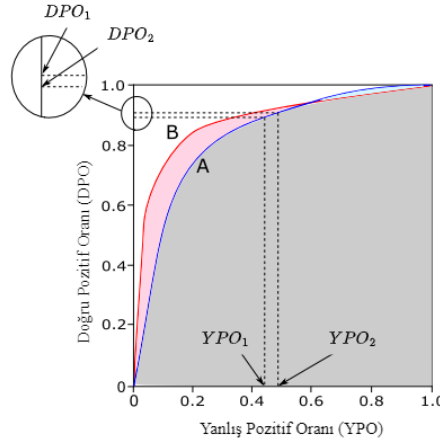
$$F_{\beta} = (1 + \beta^2) \frac{Pr \times TPR}{(\beta^2 \times Pr) + TPR}$$

Matthews Korelasyon Katsayısı (MKK): Gözlenen ve tahmin edilen sınıflar arasındaki korelasyonu temsil eder. Çıkan katsayı değeri +1 e yakınsa tahminin güçlü olduğunu temsil ederken -1 ise tahmin edilen değer ile gerçek değer arasındaki uyumsuzluğu temsil etmektedir (Tharwat, 2020).

$$MKK = \frac{DP \times DN - YP \times YN}{\sqrt{(DP + YP)(DP + YN)(DN + YP)(DN + YN)}}$$

ROC eğrisi altında kalan alan (Area Under The ROC Curve, AUC): ROC eğrisi altında kalan alanlar için kullanılan bir metriktir. AUC değeri her zaman 0 ile 1 arasında sınıflandırılmış bir değer olup bu değer 0.5 den büyük olması beklenmektedir. Aşağıda şekilde gösterildiği gibi B sınıflandırıcısının AUC değeri A sınıflandırıcısından büyüktür. Bu nedenle A değeri daha performanslıdır. Kırmızı ile

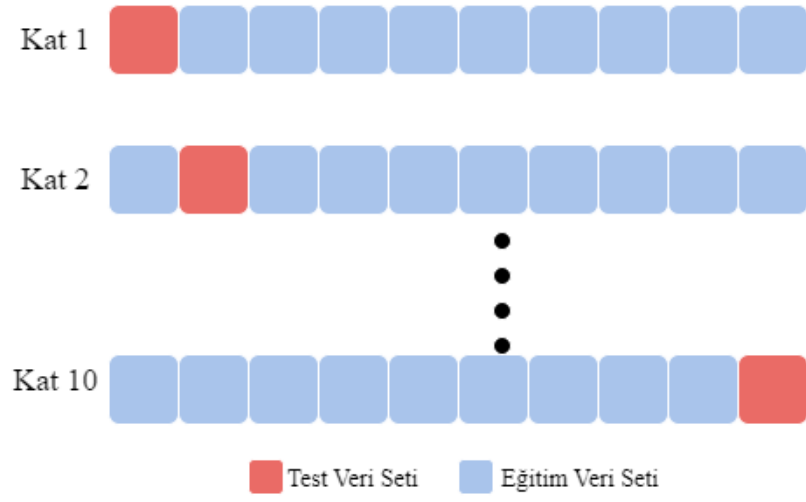
taralı alan ise B sınıflandırıcısının A sınıflandırıcısından daha iyi performans gösterdiğini temsil etmektedir (Tharwat, 2020).



Şekil 3.20. AUC grafiğinin gösterimi

3.8. Çapraz Geçerlilik

Çapraz geçerlilik, model kalibrasyonunun sağlanamadığı durumlarda kullanılan, bir modelin sağlamlık derecesini, modelin doğruluğunu, sınıflandırma başarısını ortaya çıkaran ve aynı zamanda bir modelin ne derece uygun olduğunun belirlenmesine yarayan yöntemlerden biridir. Model doğrulama uygulamalarında daha çok k katlı çapraz geçerlilik yöntemi kullanılmaktadır. Bu yöntemde n gözlemden oluşan bir veri seti oluşturulur, bu veri seti daha sonra rastgele seçilerek eşit boyutta k adet gruba bölünür. Burada belirtilmiş olan k alt küme sayısını belirtmektedir. Bu kümelerden biri test seti olarak atanırken, geri kalanlar ise eğitim seti olarak kullanılır. Model eğitim setini temsil eden $k-1$ alt kümeleri kullanarak modeli eğitir. Daha sonra model, doğrulama seti olarak adlandırılan kalan alt kümeye uygulanır ve modelin performansı ölçülür. Bu süreç k alt kümenin her biri doğrulama seti olarak hizmet edene kadar tekrarlanır. Şekil 3.21’de 10 katlı çapraz geçerlilik örnek verilmiştir. Yani $k = 10$ olduğunda, ilk kat doğrulama kümesini nitelerken diğer katlar eğitim kümesi olarak nitelendirilir. İkinci kat doğrulanan alt küme iken diğer katlar ise yine eğitim kümesidir ve bu işlemin sonucunda k performans ölçümünün ortalaması, çapraz doğrulanmış performans ölçütü olarak nitelendirilir (Berrar, 2019; Marcot ve Hanea, 2021).



Şekil 3.21. Çapraz geçerliliğe ait örnek gösterim (Berrar, 2019)

4. UYGULAMALAR

Bu bölümde öncelikle, kullanılmış olan veri setinin nasıl elde edildiğine değinilmiştir. İkinci adımda veri setindeki kategorik ve nümerik değişkenlerin düzenlenmesinden bahsedilmiş ve veri setinde yer alan değişkenlerin açıklamaları ile beraber dağılımları incelenmiştir. Üçüncü adımda veri setinin modele atılmadan önce veri ön işleme adımlarından bahsedilmiştir. Dördüncü adımda ise makine öğrenmesi ile sınıflandırma işlemi gerçekleştirilmiş, modele ait performans metrikleri ile beraber kurulan algoritmaların hiper parametreleri incelenmiştir. Beşinci ve son adımda ise doğruluk oranı en iyi olan algoritmanın değişken önem seviyeleri incelenmiştir.

4.1. Veri Setinin Elde Edilmesi ve Düzenlenmesi

Çalışmada kullanılan veri seti İBB'nin iştirak şirketi olan BELBİM'den temin edilmiştir. Veri seti araştırmacı tarafından kuruma ait veri tabanından sorgulanmış ve veri setinde kişisel ya da bir kişiyi temsil edebilecek herhangi bir bilginin bulunmamasına büyük önem gösterilerek tasarlanmıştır.

Veri seti belirlenirken ilk olarak bağımlı değişkenin ne olduğuna karar verilmiş ve bu bağımlı değişkeni açıklayacak bağımsız değişkenlerin neler olabileceğine karar verilmiştir. Bu çalışmada yer alan bağımlı değişken açık kart ve kapalı kart olmak üzere iki kategoriden oluşan bir kategorik değişkendir. Bu değişkende yer alan açık kartlar ulaşım ya da ulaşım dışında herhangi bir sistimali olmayan kartları temsil ederken, kapalı kartlar ulaşım ya da ulaşım dışında kurum tarafından belirlenmiş birtakım sistimal senaryolarına takılarak kapatılmış kartları temsil etmektedir. Veri setinde yer alan bağımsız değişkenler ise bu kullanıcılara ait demografik bilgilerden, kart bilgilerinden, ulaşım ve ulaşım dışında yapmış olduğu işlemlerden ve kartlara yapılan yükleme bilgilerinden oluşmaktadır.

Bağımsız değişkenler oluşturulurken, İstanbulkart kullanıcısının 1 Ocak 2021 tarihi ile 1 Kasım 2022 tarihleri arasında en fazla hangi noktalarda ulaşım ve ulaşım dışı harcamalar yaptığı baz alınmış, kullanıcının yapmış olduğu yüklemelerde de bu kriter baz alınarak veri tabanında sorgulama yapılmıştır. Bunu örneklendirmek gerekirse, açık veya kapalı kart olma durumu gözetmeksizin bir kartın en fazla geçiş yaptığı hat türünün toplamda 10 adet metrobüs, 8 adet metro ve 5 adet otobüs olduğunu varsayarsak, bu durumda bu kişinin en fazla geçiş yaptığı hat türü metrobüs olarak gruplandırılır. Bu veri setinde bu kuralın uygulandığı değişkenler sırasıyla: en çok

yüklemenin yapıldığı kanal, en çok geçiş yapılan hat tipi, en çok geçiş yapılan hat, en çok geçiş yapılan yol tipi, en çok geçişin yapıldığı bölge ve ulaşım dışında en çok harcamanın yapıldığı kanallardan oluşmaktadır. Bununla birlikte veri setinde yer alan birtakım nümerik değişkenlere de buna benzer bir kural oluşturulmuştur. Örneğin bir kullanıcının kartına yüklediği bakiyelerin toplamı alınmış ve bu değişken toplam yüklenen bakiye değerini oluşturmuştur. Bu kuralın uygulandığı değişkenler ise sırasıyla: aktif kart sayısı, pasif kart sayısı, ulaşım yapılan işlem adedi, ulaşım yapılan işlem tutarı, ulaşım dışında yapılan işlem adedi, ulaşım dışında yapılan işlem tutarı ve yüklenen bakiyelerden oluşmaktadır.

Veri setinde yer alan birtakım kategorik değişkenlere ait sınıf sayısı fazla olduğundan ve bu durum sınıflandırma performansını düşürdüğünden bu değişkenlere birtakım dönüşümler uygulanmıştır. Bu değişkenlerden biri en çok geçiş yapılan hat değişkenidir. Bu değişken veri setinin ilk halinde 358 adet kategoriden oluşmaktadır. Kategori sayısını azaltmak için en çok geçiş yapılan hatlar, en çok geçiş yapılan hat tipi değişkeni altında sınıflandırılmış ve beş adet kategori altında birleştirilmiştir. Bu değişkene ait beş kategori aşağıdaki kural setine göre oluşturulmuştur.

Ana Hatlar: İstanbul şehrinde güzergah olarak bir yakadan diğer yakaya, aynı zamanda yakalardan Adalar'a ulaşım hizmeti veren veya bir yaka içinde ilçeler arası ulaşım sağlayan hatlardır. Örneğin Metrobüs, Metro, Marmaray, Şehir Hatları ve 500T gibi hatlar ana hat olarak nitelendirilir.

Besleyici Hat: Ana hatları besleyen toplu taşıma araçları besleyici hat olarak nitelendirilmiştir. Besleyici hatlar Metro ve Metrobüs'e ücretsiz ulaşım imkanı sağlayan hatlardır.

Taşıyıcı ve Toplayıcı Hatlar: İlçelerin kendi içinde ya da bir ilçeden başka bir ilçeye ulaşım imkanı sağlayan hatlar taşıyıcı ve toplayıcı hatlar olarak nitelendirilmiştir.

Özel Amaçlı Hatlar: Kısa mesafede hizmet sağlayan hatlar özel amaçlı hat olarak nitelendirilmiştir. Bu hatları örneklendirecek olursak Ada taksi, Ada Otobüs, Teleferik ve Füniküler gibi hatlar özel amaçlı hat olarak tanımlanmıştır.

Tanımsız Hatlar: Ulaşım hattı içinde yer almasına rağmen herhangi bir hat olarak tanımlanamayan hatlar tanımsız hatlardır. Buna en uygun örnek Metro hatları içinde bulunan lavabolardır.

Buna benzer bir diğ er d on    m ise en  ok ge i  yapılan durak de i keni  zerinde uygulanmı tır. En  ok ge i in yapıldı ı durak de i keni toplamda 454 adet kategoriden olu maktadır. Bu de i ken be  kategori altında en  ok ge i  yapılan b lge de i keni altında olu turulmu tur. Bu b lgeler olu turulurken veri setinde yer alan 454 adet dura ın son bir yıldıki ge i  adetlerinin ortalaması alınmı  ve ge i in en yo un oldu u 92 adet durak 1.B lge olarak nitelendirilmi tir. Di er duraklar yo unluklarına g re 2,3,4 ve 5.B lge olarak nitelendirilmi tir.

Di er bir d on    m n uygulandı ı kategorik de i ken ise kart tipleridir. Bu kart tipleri insanların mesleki gruplarına veya sosyal durumlarına g re  e itlili i bulunmaktadır.  alı mada kullanılan veri setinde toplamda 46 adet kart tipinin oldu u g zlenmi tir. Bu kart tipleri kullanım alanlarına g re kart t rleri de i keni altında toplanarak toplamda altı kategoriden olu an de i ken haline getirilmi tir.   retmen, sa lık  alı anı ve kolluk kuvvetleri gibi belirli meslek grubuna ait olan kartlar kamu kartı olarak nitelendirilirken; sosyal kart,  ehit aileleri kartı ve anne kart gibi kartlar imtiyazlı kart olarak nitelendirilmi tir. Di er kart t rlerinde ise herhangi bir de i iklik yapılmamı  ve bu de i kene dahil edilmi tir.

Veri setinde sorgulama i lemi yapılırken de i ken bazlı kayıp verinin olmamasına  zen g sterilmi  ve bu  alı mada kullanılan veri setinin hi bir de i keninde kayıp veri bulunmamaktadır.

4.2. Veri Setinde Yer Alan De i kenlerin Da ılımları ve A ıklamaları

Bu b l mde veri setinde kullanılan t m de i kenlerin da ılımları ve de i kenlerin veri seti i indeki a ıklamalarına de inilmi tir. Tablo 4.1’de demografik de i kenlerin da ılımları ve tanımlayıcı istatistikleri, Tablo 4.2’de kitlenin kullandı ı kart tipleri ve kart hareketlerine kategorik de i kenlerin da ılımları ve a ıklamaları, Tablo 4.3’te kitlenin kart hareketlerinin n merik de i kenlerinin tanımlayıcı istatistiklerine de inilmi tir.

Tablo 4.1. Kitleye ait demografik özellikler

Değişkenler		Adet(n)	%
Cinsiyet	Erkek	3736	66,18
	Kadın	1909	33,82
Medeni durum	Bekar	3298	58,42
	Evli	2347	41,58
	$\bar{X} \pm SS$	Min-Mak	Medyan (Ortanca)
Yaş	35,39 $\pm$ 13,90	10-100	31

Çalışmaya katılan kitlenin %66,18'i ($n = 3736$) erkek, %33,82'si ($n = 1909$) kadından oluşmaktadır. Kitleye ait yaş dağılımı ortalama $35,39 \pm 13,90$ olup; en düşük yaş 10 en yüksek 100 olduğu gözlenmiştir. Kitlenin %58,42'si ($n = 3298$) bekar, %41,58'nin ($n = 2347$) evli olduğu gözlenmiştir.

Tablo 4.2. Kitlenin kart bilgileri ve hareketleri

Değişkenler	Kategoriler	Adet(n)	%
Kart durumu	Açık Kart	3440	60,94
	Kapalı Kart	2205	39,06
Kartlara ait kart türleri	65 Yaş Üstü Kart	331	5,86
	Engelli Kart	381	6,75
	İmtiyazlı Kart	1035	18,33
	Kamu Kart	1953	34,60
	Öğrenci Kart	528	9,35
	Tam Kart	1417	25,10
Kartlara ait geçiş tipleri	İndirimli Geçiş Tip 1	868	15,38
	İndirimli Geçiş Tip 2	634	11,23
	Tam Geçiş	1417	25,10
	Ücretsiz Geçiş	2726	48,29
En çok yüklemenin yapıldığı kanal	Akdom POS	563	9,97
	Biletmatikler	3731	66,09
	Bilinmeyen	182	3,22
	Mobil	683	12,10
	Online Kanallar	486	8,61
En çok geçiş yapılan hat tipi	Ana Hatlar	4268	75,61
	Besleyici Hatlar	30	0,53
	Özel Amaçlı Hatlar	56	0,99
	Tanımsız	168	2,98

	Taşıyıcı ve Toplayıcı Hat	1123	19,89
	Elektrikli Araçlar	56	0,99
	Metrobüs	990	17,54
	Otobüs	1194	21,15
En çok geçiş yapılan hat	Tanımsız	168	2,98
	Marmaray	1037	18,37
	Metro Hattı	1891	33,50
	Deniz Araçları	309	5,47
	Deniz	309	5,47
	Elektrikli Araçlar	56	0,99
En çok geçiş yapılan yol tipi	Otoyol	2184	38,69
	Raylı	2928	51,87
	Tanımsız	168	2,98
	1.Bölge	2884	51,09
	2.Bölge	1258	22,29
En çok geçişin yapıldığı bölge	3.Bölge	665	11,78
	4.Bölge	623	11,04
	5.Bölge	215	3,81
	Belediyeye Ait Sosyal Tesisler	5313	94,12
	Cafe ve Restoranlar	69	1,22
Ulaşım dışında en çok harcama yapılan kanallar	Elektrikli Araçlar	39	0,69
	Kültür ve Sanat Noktaları	71	1,26
	Marketler	153	2,71

Kart durumu değişkeni açık ve kapalı kartlar olmak üzere iki kategoriden oluşmaktadır. Açık kartlar ulaşım veya ulaşım dışında herhangi bir suistimali bulunmayan kartlardan oluşurken, kapalı kartlar ulaşım ya da ulaşım dışında yapılan herhangi bir suistimalden dolayı kapatılmış kartlardan oluşmaktadır. Kart durumu değişkeni veri setinde içinde bağımlı bir değişken olup %60,94'ü ($n = 3440$) açık kartlardan, %39,06'sı ($n = 2205$) kapalı kartlardan oluşmaktadır.

Kartlara ait kart türleri kategorik değişkeni veri seti içinde yer alan kartların kart çeşidini temsil etmektedir. 65 yaş üstü kart, 65 yaşını doldurmuş kişilerin kullandığı kart tipini temsil etmekte olup değişken içindeki dağılımı %5,86'dır ($n = 331$); engelli kart ise %40 ya da üzerinde engeli bulunan kişilerin kullanmış olduğu kart tipini temsil etmekte olup değişken içindeki dağılımı %6,75'dir ($n = 381$). İmtiyazlı kart ise kurum tarafından belirlenmiş özel imtiyaza sahip olan kartları temsil etmekte ve değişken içindeki dağılımı %18,33'dür ($n = 1035$). Kamu kartı ise, kamu

sektöründe çalışan meslek gruplarına göre kişilere verilen kart tipi olarak nitelendirilmektedir. Bu kart tipinin değişken içindeki dağılımı ise %34,60'dır ($n = 1953$). Öğrenci kartı ise ilk okul, lise ve üniversite öğrenimi gören kişilere verilen kart tipi olarak nitelendirilen kartlar olup değişken içindeki dağılımı %9,35'dir ($n = 528$). Tam kart, herhangi bir imtiyaz ya da özelliğe sahip olmayan kartlardan oluşan kartlardır. Bu kartların değişken içindeki dağılımı %25,10'dur ($n = 1417$).

Geçiş tipi değişkeni, kartların sahip olduğu geçiş tiplerini temsil etmektedir. Kartların ulaşım terminallerinden yaptığı geçiş tipi indirimli tip 1, indirimli tip 2, tam ve ücretsiz olmak üzere dört çeşide ayrılmaktadır. İndirimli tip 1 geçiş tipi, imtiyazlı ve kamu kartı içinde yer alan bazı kartların indirimli olarak geçiş yaptığı kartları temsil etmektedir. Bu kategorinin değişken içindeki dağılımı %15,38'dir ($n = 868$). İndirimli tip 2 geçiş tipi, indirimli tip 1 geçiş tipine benzer olup imtiyazlı ve kamu kartı içinde yer alan bazı kartların indirimli geçiş yapabildiği geçiş tipi olarak nitelendirilmektedir. Bu kategorinin değişken içindeki dağılımı %11,23'dür ($n = 634$). Ücretsiz geçiş tipi; 65 yaş üstü kartların, engelli kartların ve belirli bir imtiyaza sahip birtakım kartların ulaşımında ücretsiz geçiş yaptığı geçiş tipini temsil etmekte olup, bu kategorinin değişken içindeki dağılımı %48,29'dur ($n = 2726$). Tam geçiş ise herhangi bir imtiyazı bulunmayan kartların tam ücret ödeyerek yaptığı geçiş tipi olarak nitelendirilmiş ve bu kategorinin değişken içindeki dağılımı ise %25,10 ($n = 1417$) olduğu gözlenmiştir.

Yüklemenin yapıldığı kanal değişkeni, kart sahiplerinin karta bakiye yükleme işlemi gerçekleştirilirken en çok hangi terminalleri kullandığı bilgisini içermektedir. Bu değişkene ait kategoriler Mobil, Online kanallar, Akdom POS, Biletmatik ve bilinmeyen olmak üzere toplamda beş kategoriden oluşmaktadır. Akdom POS kanalı, kuruma ait bireysel ve kurumsal temsilciliklerden bakiye yükleme işlemlerinin gerçekleştirildiği kanallardır. Bu kategorinin değişken içindeki dağılımı %9,97'dir ($n = 563$). Biletmatikler kurum tarafından üretilmiş ve kart yüklemelerinin gerçekleştirildiği makinelerdir. Biletmatiklerin değişken içindeki dağılımı %66,09'dur ($n = 3731$). Mobil uygulamadan yapılan yüklemelerin değişken içindeki dağılımı %12,20'dir ($n = 683$). Online kanallardan yapılan yüklemenin değişken içindeki dağılımı ise %8,61 ($n = 486$) olup, yüklemenin yapıldığı kanalı bilinmeyenlerin değişken içindeki dağılımı ise %3,22'dir ($n = 182$).

En çok geiş yapılan hat türü deęişkenini incelediğimizde, ana hatların deęişken içindeki dağılımı %75,61 ($n = 4268$), besleyici hatların deęişken içindeki dağılımı %0,53 ($n = 30$), özel amaçlı hatların deęişken içindeki dağılımı %0,99 ($n = 56$), taşıyıcı ve toplayıcı hatların deęişken içindeki dağılımı %19,89 ($n = 1123$) ve tanımsız olarak nitelendirilen hatların deęişken içindeki dağılımı %2,98 ($n = 168$) olarak gözlenmiştir.

Kitlenin en çok geiş yaptığı hatlar incelendiğinde, Metro hattının ulaşım hatları içindeki dağılımı %33,50 ($n = 1891$), otobüs hatlarının ulaşım içindeki dağılımı %21,15 ($n = 1194$), Marmaray hattının ulaşım içindeki dağılımı %18,37 ($n = 1037$), Metrobüs hattının ulaşım içindeki dağılımı %17,54 ($n = 990$), deniz araçlarının ulaşım içindeki dağılımı %5,47 ($n = 309$), tanımsızların (lavabo) ulaşım içindeki dağılımı %2,98 ($n = 168$) ve elektrikli araçların ulaşım içindeki dağılımı %0,99 ($n = 56$) olarak gözlenmiştir.

Kitlenin ulaşımında en çok tercih ettiği yol tipi incelendiğinde, raylı sistemleri tercih edenlerin oranı %51,87 ($n = 2928$), lastik tekerlekli araçların tercih edilme oranı %38,69 ($n = 2184$), deniz yolunu tercih edenlerin oranı %5,47 ($n = 309$), tanımsız olarak tercih edilenlerin oranı %2,98 ($n = 168$) ve elektrikli araçların tercih edilme oranı %0,99 ($n = 56$) olarak gözlenmiştir.

Kitlenin en çok geiş yaptığı bölgeyi incelediğimizde; 1.bölgeden geiş yapanların deęişken içindeki dağılımı %51,09 ($n = 2884$), 2.bölgenin deęişken içindeki dağılımı %22,29 ($n = 1258$), 3.bölgenin deęişken içindeki dağılımı %11,78 ($n = 665$), 4.bölgenin deęişken içindeki dağılımı %11,04 ($n = 623$) ve 5.bölgenin deęişken içindeki dağılımı %3,81 ($n = 215$) olarak gözlenmiştir.

Kitlenin ulaşım dışındaki en çok harcama yaptığı kanallar incelendiğinde %94,12'nin ($n = 5313$) belediyeye ait sosyal tesislerde, %2,71'nin ($n = 153$) marketlerde, %1,26'nın ($n = 71$) kültür ve sanat noktalarında, %1,22'nin ($n = 69$) Cafe ve Restonlar'da, %0,69'nun ise elektrikli araçlarda kullanıldığı gözlenmiştir.

Tablo 4.3. Kitlelerin kart bilgilerinin ve kart hareketlerinin nümerik dağılımları

Değişkenler	$\bar{X} \pm SS$	Medyan (Min-Maks)
Toplam kart sayısı	4,47 $\pm$ 2,93	4 (1-57)
Aktif kart sayısı	2,58 $\pm$ 2,18	2 (0-58)
Pasif kart sayısı	1,90 $\pm$ 1,96	1 (0-21)
Ulaşım dışında yapılan işlem adedi	6,09 $\pm$ 15,68	2 (1-408)
Ulaşım dışı yapılan harcama tutarı (₺)	222,39 $\pm$ 818,64	19,00 (0-15583,50)
Ulaşımında yapılan harcama sayısı	472,92 $\pm$ 489,70	314 (1-4008)
Ulaşımında yapılan harcama tutarı (₺)	2505,09 $\pm$ 4105,96	933,47 (0-60048,08)
Toplam bakiye (₺)	1867,42 $\pm$ 3115,70	680,00 (0-45637,55)

Kitleye ait nümerik değişkenlerin tanımlayıcı istatistikleri Tablo 4.3'te verilmiştir. Kitlenin ortalama kart sayısı 4,47 $\pm$ 2,93 olup; aktif kart sayısı ortalama 2,58 $\pm$ 2,18 ve pasif kart sayısı ortalama 1,90 $\pm$ 1,96 olarak gözlenmiştir.

Kitlenin ulaşımında yaptığı toplam harcama adedi ortalama 472,92 $\pm$ 489,70 olup; ulaşımında yaptığı toplam harcama tutarı ortalama 2505,09 $\pm$ 4105,96 olarak gözlenmiştir.

Kitlenin ulaşım dışında yapmış olduğu toplam harcama adedi ortalama 6,09 $\pm$ 15,68 olup, ulaşım dışında yapılan toplam harcama tutarı 222,39 $\pm$ 818,64 olarak gözlenmiştir. Buradan sonuçla kitlenin ulaşımında yapmış olduğu işlem adedi ve işlem hacminin ulaşım dışında yapılan işlem adedi ve işlem hacminden daha büyük olduğu gözlenmiştir.

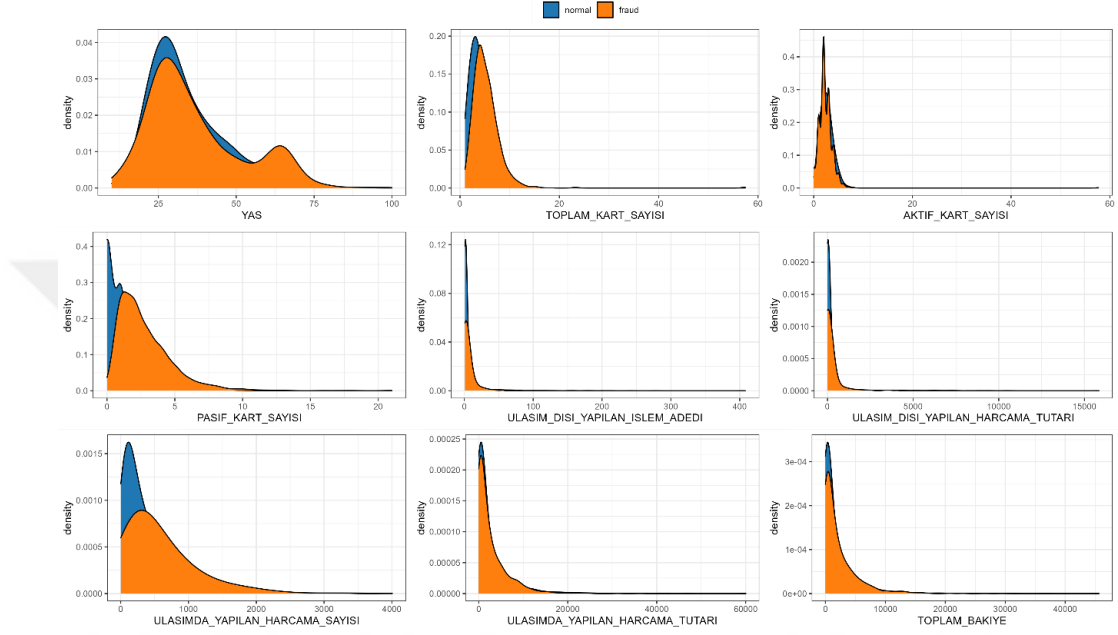
Kitlenin karta yükledikleri toplam bakiye tutarlarını incelediğimizde ise ortalama 1867,42 $\pm$ 3115,70 TL olarak gözlenmiştir.

4.3. Veri Ön İşleme

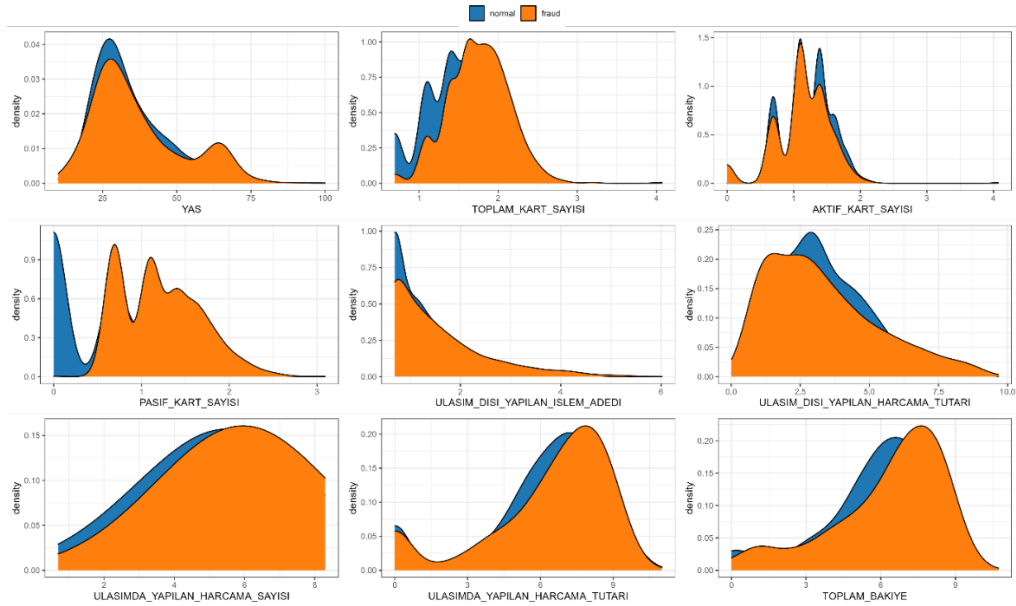
Bu bölümde veri seti içinde yer alan nümerik değişkenlerin normallik dağılımları incelenmiş ve modele dahil edilecek olan tüm değişkenlere değişken seçimi yöntemi uygulanmıştır.

Şekil 4.1'de gösterildiği gibi kartların açık ve kapalı olma durumlarına göre yaş, toplam kart sayısı, aktif kart sayısı, pasif kart sayısı, ulaşım dışında yapılan işlem adedi, ulaşım dışında yapılan harcama tutarı, ulaşımında yapılan harcama sayısı, ulaşımında yapılan harcama tutarı ve toplam yüklenen bakiye değişkenlerinin dağılımları incelenmiştir. Şekil 4.1'e göre yaş değişkeni haricindeki diğer

değişkenlerin normallik varsayımını sağlamadığı görülmektedir. Yani bu değişkenlerin simetrik bir dağılım göstermediği ve hepsinde pozitif çarpıklığın olduğu gözlenmektedir. Bu nedenle Şekil 4.2’de verilerin simetrik bir görünüme sahip olabilmesi için yaş haricindeki diğer nümerik değişkenlere logaritmik dönüşüm uygulanmış ve verilerin pozitif çarpıklık durumu kaldırılarak verilerin simetrik dağılımı sağlanmıştır.

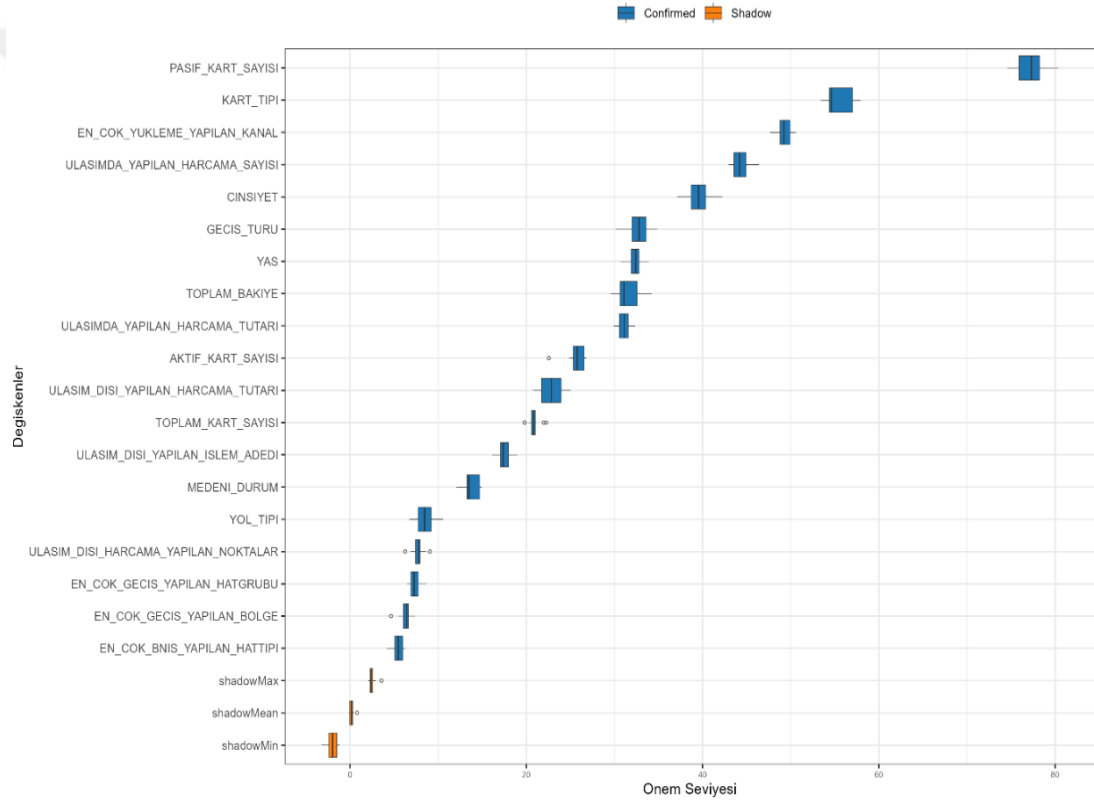


Şekil 4.1. Nümerik değişkenlere ait yoğunluk grafikleri



Şekil 4.2. Logaritmlar alındıktan sonra nümerik değişkenlere ait yoğunluk grafikleri

Veri seti içinde yer alan değişkenlerin önem seviyeleri boruta algoritması ile değerlendirilmiş ve Şekil 4.3’de görüldüğü gibi kategorik değişkenlerin önem seviyeleri verilmiştir. Bu sonuca göre veri seti içinde yer alan pasif kart sayısı değişkeninin anlamlılık düzeyi en yüksek olduğu gözlenirken, anlamlılık düzeyi en yüksek ikinci değişkenin kart tipi değişkeni olduğu gözlenmiştir. En çok geçiş yapılan bölge ve en çok geçiş yapılan hat tipi değişkenlerinin önem seviyesi her ne kadar düşük olsa da boruta algoritmasına göre veri setinde yer alan tüm değişkenler, en etkili shadow değişken olan shadowMaxdan anlamlı derecede daha etkili oldukları için önemli olduğu kabul edilmiş ve herhangi bir değişken silme işlemi gerçekleştirilmeden tüm değişkenler sınıflandırma algoritmalarına dahil edilmiştir.



Şekil 4.3. Boruta değişken seçimi ile değişken önem seviyeleri

4.4. Sınıflandırma Algoritmalarının Veri Seti Üzerindeki Uygulaması

Bu bölümde kullanılan veri setinde makine öğreniminde kullanılan birtakım sınıflandırma algoritmalarının sınıflandırma performansları karşılaştırılmış ve algoritmalarda kullanılan hiper parametreler incelenmiştir.

Tablo 4.4. Sınıflandırma algoritmalarına ait performans değerleri

Sınıflayıcı	ACC	MCC	F1 SKOR	AUC
CART algoritması	0,834	0,657	0,827	0,902
Rastgele ormanlar	0,863	0,714	0,857	0,942
Adaboost	0,871	0,730	0,865	0,947
Lojistik regresyon	0,802	0,583	0,791	0,885
K - en yakın komşuluk	0,794	0,567	0,783	0,869
DVM (Lineer çekirdek)	0,804	0,588	0,793	0,885
XGBoost	0,881	0,750	0,875	0,953
Naif Bayes	0,721	0,546	0,720	0,874
Derin öğrenme	0,845	0,681	0,840	0,925

Tablo 4.4'e göre sınıflandırma doğruluk dereceleri incelendiğinde doğruluk derecesi en yüksek olan algoritma 0,881 ile XGBoost algoritmasıdır. Adaboost algoritması 0,871 doğruluk derecesi ile en iyi ikinci algoritma olurken, 0,863 ile Rastgele Orman algoritması doğruluk derecesi en iyi üçüncü algoritmadır. Doğruluk oranı en düşük olan algoritma ise 0,721 ile Naif Bayes algoritmasıdır.

Tablo 4.5. Sınıflandırma algoritmalarının hiper parametreleri

Sınıflayıcı	Seçilen hiperparametreler
CART algoritması	split = "gini", minsplit = 30, minbucket = 2, maxdepth = 30, cp = 0.002
Rastgele ormanlar	max.depth = 30, num.trees = 250, sample.fraction = 1, mtry = 13, splitrule = "gini", min.node.size = 1
Adaboost	mfinal = 100, maxdepth = 30
Lojistik regresyon	Hiperparametresi bulunmamaktadır.
K - en yakın komşuluk	K = 10
DVM (Lineer çekirdek)	C = 1
XGBoost	nrounds = 200, max_depth = 10, eta = 0.1, colsample_bytree = 1, min_child_weight = 2, subsample = 0.9
Naif Bayes	laplace = 0
Derin öğrenme	hiddenlayers = 2, activation = {"selu", "sele", "softmax"}, layer nodes = {32, 16, 2}, optimizer = "adam", epoch = 500, batch_size = 64

Sınıflandırma algoritmalarının en yüksek doğruluk derecesinin bulunabilmesi için algoritmalara ait en iyi hiper-parametreler denenmiş ve algoritmalara ait en yüksek

doğruluk derecesini veren hiper-parametreler Tablo 4.5'te verilmiştir.

Karar ağaçlarında en yüksek doğruluk derecesinin bulunabilmesi için aşağıdaki hiper-parametreler denenmiştir;

- En iyi bölünme kriteri (split): Gini İndeksi
- Bir düğümün yeniden bölünmesi için minimum veri sayısı (minsplit): 20
- Herhangi bir yaprak düğümdeki minimum gözlem sayısı (minbucket): 2
- Karar ağacının maksimum derinliği (maxdepth): 30
- Bölünme için gereken minimum kök düğümün sapma oranı (cp): 0.002

Rastgele Orman'da en yüksek doğruluk derecesinin bulunması için aşağıdaki hiper-parametreler denenmiştir;

- Ağacın maksimum derinliği (maxdepth): 30
- Oluşturulan ağaç sayısı (num.trees): 250
- Örnekleme alınan gözlemlerin oranı (sample.fraction): 1
- Her bir ağaçta kullanılacak rastgele değişken sayısı: 13
- En iyi bölünme kriteri (splitrule): Gini indeksi
- Minimum düğüm boyutu (min.node.size):1

Adaboost algoritmasında en yüksek doğruluk derecesinin bulunması için aşağıdaki hiper-parametreler denenmiştir;

- Kullanılan ağaç sayısı (mfinal): 100
- Ağacın maksimum derinliği (maxdepth): 30

K – en yakın komşuluk algoritmasında en yüksek doğruluk derecesinin bulunması için en iyi hiper-parametre olarak k-en yakın komşu sayısı 10 olarak verilmiştir.

Destek vektör makinelerinde en yüksek doğruluk derecesi için en iyi hiper-parametre değeri için c parametresi kullanılmıştır. Bu parametre XGBoost algoritmasında en yüksek doğruluk derecesinin bulunması için aşağıdaki hiper-parametreler denenmiştir;

- Final modele ait ağaç sayısı (nrounds): 200
- Ağacın maksimum derinliği (max_depth): 10
- Öğrenme katsayısı (eta): 0.1
- Ağaç oluşturmak için değişkenlerin alt örnek oranı (colsample_bytree):1
- Minimum gözlem ağırlığının toplamı (min_child_weight): 2

- Ağacın yapısını oluşturan satırların oranı (subsample): 0.9

Naif Bayes algoritmasında en yüksek doğruluk derecesinin bulunması için sıfır olasılık sorununun düzeltilmesini sağlayan Laplace dönüşümü kullanılmış ve bu parametre 0 olarak alınmıştır.

Derin öğrenme algoritmasında en yüksek doğruluk derecesinin bulunması için aşağıdaki hiper-parametreler denenmiştir;

- Gizli katman sayısı (hiddenlayers): 2
- Aktivasyon fonksiyonları (activation) : selu, selu, softmax
- Eğitim adım sayısı (epoch): 500
- Optimizasyon algoritması (optimizer): adam
- Ağa verilen alt örnek sayısı (batch_size): 64

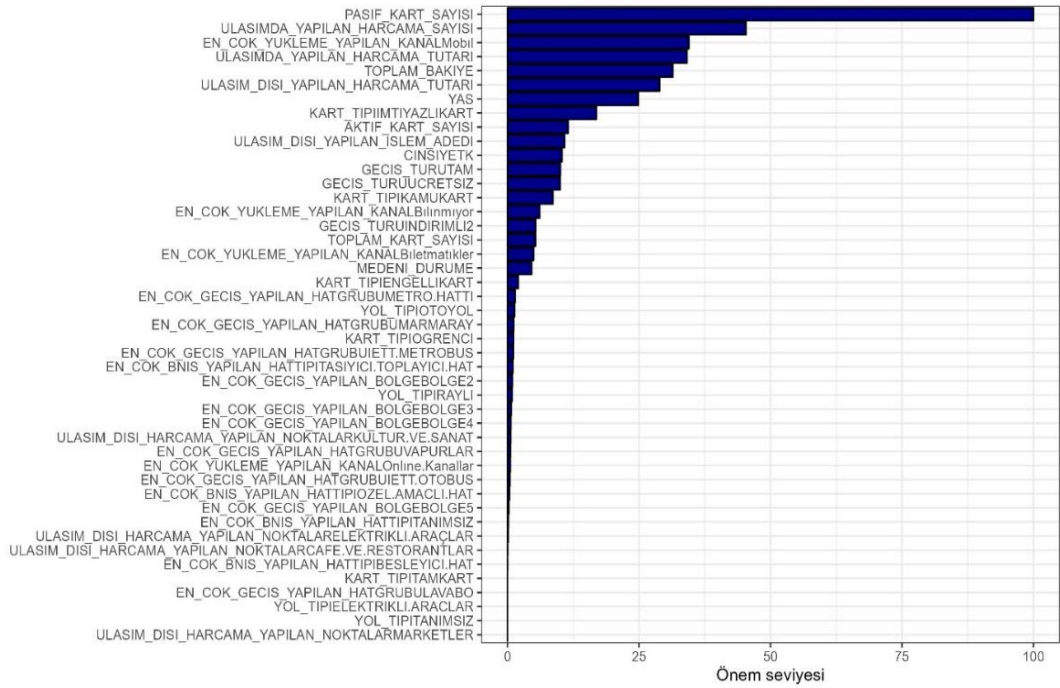
4.5. En İyi Modele Göre Değişkenlerin ve Kategorilerin Önem Seviyeleri

Bir önceki bölümde doğruluk oranı en yüksek olan algoritmanın XGBoost (0,881) algoritması olduğu görülmüş ve bu bölümde de XGBoost algoritması kullanılarak bir kartın suistimal yapmasını açıklayan en iyi değişkenler belirlenmiştir. XGBoost algoritmasına göre bir kart kullanıcısının suistimal yapmasını en iyi açıklayan bu değişkenler sırasıyla; kullanıcıya ait toplam pasif kart sayısı, ulaşımda yapılan toplam harcama sayısı, en çok yüklemenin yapıldığı kanal, ulaşımda yapılan toplam harcama tutarı, toplam yüklenen bakiye, ulaşım dışında yapılan toplam harcama tutarı, yaş, kart tipi, kullanıcıya ait toplam aktif kart sayısı, ulaşım dışında yapılan toplam işlem adedi, cinsiyet, geçiş türü, toplam kart sayısı ve medeni durum değişkeninin bir kullanıcının kart suistimali yapmasını açıklayan değişkenlerdir. Kart kullanıcısının en çok geçiş yaptığı hat tipi, en çok geçiş yaptığı hat grubu, yol tipi, en çok geçiş yapılan bölge ve ulaşım dışında en çok harcamanın yapıldığı nokta değişkenlerinin önem seviyelerinin düşük olduğu gözlenmiştir.

Bir kart kullanıcısının suistimal yapmasını açıklayan en iyi değişkenler, değişken türüne göre incelendiğinde nümerik değişkenlerin bir kullanıcının suistimal yapmasını en iyi açıklayan değişken türü olduğu gözlenmiştir. Kategorik değişkenlerden ise bir kullanıcının suistimal yapmasını açıklayan en iyi değişkenler şu şekilde değerlendirilmiştir. En çok yüklemenin yapıldığı kanal değişkeninin içinde; mobil, biletmatik ve yüklemenin yapıldığı kanalın bilinmiyor olmasının bir kullanıcının suistimal yapmasını açıklayan en iyi kategorik değişken olduğu

görülmüştür. Bir kullanıcıya ait kart tipinin imtiyazlı, kamu kartı ve engelli kartı olması bir kullanıcının suistimal yapmasını açıklayan en iyi ikinci kategorik değişken olduğu görülmüştür. Bir kullanıcının kart ile yapmış olduğu geçiş tipi değişkenine ait tam geçiş, indirimli tip 2 geçiş ve ücretsiz geçişin olması bir kullanıcının suistimal yapmasını açıklayan en iyi üçüncü kategorik değişken olduğu görülmüştür.

Buradan elde edilen sonuca göre bir kart kullanıcısının suistimal yapmasını açıklayan en iyi değişken türünün nümerik değişkenler olduğu görülmüştür.



Şekil 4.4. XGBOOST modeline göre değişken önem seviyeleri

5. SONUÇ VE DEĞERLENDİRME

Bu çalışmada ulaşım kartlarında yapılan suistimallerin makine öğreniminin sınıflandırma algoritmaları kullanılarak tahmin edilmesi amaçlanmış ve yapılan literatür taramasında buna benzer bir çalışmanın olmadığı görülmüştür. Bu yüzden literatür taraması yapılırken kredi kartlarında yapılan dolandırıcılıkların makine öğrenimi yöntemleri ile tahmin edilmesine yönelik çalışmalar incelenmiştir.

Makine öğreniminin hayatımıza girmesi ve bu alanda yapılan çalışmaların çeşitlilik kazanması yeni neden sonuç ilişkilerinin ortaya çıkmasına yardımcı olduğu gibi günlük hayatta karşılaşılan birtakım problemlerin çözümüne hızlı ve somut katkılar sağlamıştır. Yapılan bu çalışmada ulaşım kartlarında yapılan suistimler bir problem olarak görülmüş ve bu problemin çözülmesi için makine öğreniminde sınıflandırma algoritmaları kullanılarak, sınıflandırma algoritmalarının performans metrikleri karşılaştırılmış ve bir kart kullanıcısının suistimal yapmasını açıklayan en iyi değişkenlerin hangisi olduğuna odaklanılmıştır.

Çalışmayı oluşturan veri seti nümerik ve kategorik değişkenlerden oluşmaktadır. Nümerik değişkenlerin içinden yaş değişkeni hariç diğer tüm nümerik değişkenlerin normal dağılım göstermediği görülmüş ve logaritmik dönüşüm yapılarak normal dağılıma sahip olması sağlanmıştır. Normallik varsayımı sağlandıktan sonra tüm değişkenler için değişken seçim yöntemi uygulanmış ve burada Boruta yöntemi ile değişken seçim işlemi gerçekleştirilmiştir. Bu yöntemle göre veri seti içinde yer alan tüm değişkenlerin önem seviyelerinin kabul edilebilir olduğu görülmüş ve modelleme işlemleri veri setindeki tüm değişkenler ile yapılmıştır.

Çalışma kapsamında değerlendirilen sınıflandırma algoritmalarının performans metriklerine göre en yüksek doğruluk derecesini veren algoritmanın XGBoost algoritması olduğu görülmüş en düşük doğruluk derecesini veren algoritmanın ise Naif Bayes algoritmasının olduğu tespit edilmiştir. Bu sonuca göre çalışma XGboost algoritmasına göre değişkenlerin önem seviyelerine odaklanmış ve çalışmada önem seviyeleri en yüksek ve en düşük değişkenler incelenmiştir. XGboost algoritmasının sonuçlarına göre nümerik değişkenlerin önem seviyelerinin kategorik değişkenlere göre daha yüksek olduğu sonucuna varılmıştır. Dolayısıyla bu ve buna benzer veri setlerinde birtakım kategorik değişkenler nümerik değişken yapısına dönüştürülebilir. Bu çalışma nezdinde değerlendirildiğinde; en çok biniş yapılan hat tipi değişkeninin

fazlalaştırılması ve bu algoritmalarda kullanılacak olan hiper parametre kombinasyonlarının değerlendirilmesi önerilir.

KAYNAKLAR

- Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., and Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11), e00938.
- Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., ... and Kiru, M. U. (2019). Comprehensive review of artificial neural network applications to pattern recognition. *IEEE Access*, 7, 158820-158846.
- Akar, Ö., ve Güngör, O. (2013). Rastgele Orman Algoritması Kullanılarak Çok Bantlı Görüntülerin Sınıflandırılması. *Journal of Geodesy and Geoinformation*, 1(2), 139–146. <https://doi.org/10.9733/jgg.241212.1t>
- Akman, M., Genç, Y. ve Ankaralı, H. (2011). Random forests yöntemi ve sağlık alanında bir uygulama/Random forests methods and an application in health science. *Türkiye Klinikleri Biyoistatistik*, 3(1), 36.
- Akyol, K., Şen, B., and ÇALIK, E. (2012). Biyokimya ve Hemogram Laboratuvar Test Sonuçlarının Lojistik Regresyon Yöntemiyle Analizi. *Karabük Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Karabük*.
- Asha, R. B., and KR, S. K. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, 2(1), 35-41.
- Awoyemi, J. O., Adetunmbi, A. O., and Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference on computing networking and informatics (ICCNi)* (pp. 1-9). IEEE.
- Ayhan, S., and Erdoğan, Ş. (2014). Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 9(1), 175-201.
- Balaban, M.E, and Kartal, E. (2015). *Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları* (1 st Edition). Beyoğlu, İstanbul: Çağlayan Kitabevi.
- Bardina, M., Vaganov, D., and Guleva, V. (2020). Sociodemographic features meet interests: on subscription patterns and attention distribution in online social media. *Procedia Computer Science*, 178, 162–171, <https://doi.org/10.1016/j.procs.2020.11.018>
- Batra, M., and Agrawal, R. (2018). Comparative analysis of decision tree algorithms. In *Nature inspired computing* (pp. 31-36). Springer, Singapore.
- Belattar, K., and Mostefai, S. (2015, November). Similarity measures for content-based dermoscopic image retrieval: A comparative study. In *2015 First international conference on new technologies of information and communication (NTIC)* (pp. 1-6). IEEE.
- BELBİM (2022). Faaliyet Raporu. Erişim: 20.11.2022, https://webkurumsalapi.belbim.istanbul/Media/pdf/2021_Faaliyet_Raporu.pdf
- BELBİM (2022). Ürün ve Hizmet Kataloğu. Erişim: 19.11.2022, https://webkurumsalapi.belbim.istanbul/Media/resimler/products-and-services/Belbim_urun_hizmet_katalogu.pdf
- Berrar, D. (2019). Cross-Validation. *Encyclopedia of Bioinformatics and Computational Biology*, 542–545. <https://doi.org/10.1016/b978-0-12-809633-8.20349-x>

- Bhagya Raj, G. V. S., and Dash, K. K. (2022). Comprehensive study on applications of artificial neural network in food process modeling. *Critical Reviews in Food Science and Nutrition*, 62(10), 2756-2783.
- Bhanusri, A., Valli, K. R. S., Jyothi, P., Sai, G. V., and Rohith, R. (2020). Credit card fraud detection using Machine learning algorithms. *Journal of Research in Humanities and Social Science*, 8(2), 04-11.
- Breiman, L. (2001). Random forests. *Machine learning*, 45:1, 5-32.
- Burmaoğlu, S., Oktay, E., and Üstün, Ö. Z. E. N. (2009). Birleşmiş milletler kalkınma programı beşeri kalkınma endeksi verilerini kullanarak diskriminant analizi ve lojistik regresyon analizinin sınıflandırma performanslarının karşılaştırılması. *Savunma Bilimleri Dergisi*, 8(2), 23-49.
- Campbell, C. and Ying, Y., *Learning with Support Vector Machines*, Morgan and Claypool Publishers, 2011
- Cao, J., Kwong, S., and Wang, R. (2012). A noise-detection based AdaBoost algorithm for mislabeled data. *Pattern Recognition*, 45(12), 4451-4465.
- Chaoji, V., Rastogi, R., and Roy, G. (2016). Machine learning in the real world. *Proceedings of the VLDB Endowment*, 9(13), 1597-1600.
- Charbuty, B., and Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), 20-28.
- Chen, Y. J., Lin, J. A., Chen, Y. M., and Wu, J. H. (2019). Financial Forecasting With Multivariate Adaptive Regression Splines and Queen Genetic Algorithm-Support Vector Regression. *IEEE Access*, 7, 112931–112938. <https://doi.org/10.1109/access.2019.2927277>
- Cheng, W. C., and Jhan, D. M. (2012). Triaxial accelerometer-based fall detection method using a self-constructing cascade-AdaBoost-SVM classifier. *IEEE journal of biomedical and health informatics*, 17(2), 411-419.
- Çavuşoğlu, Ü., and Kaçar, S. (2019). Anormal trafik tespiti için veri madenciliği algoritmalarının performans analizi. *Academic Platform Journal of Engineering and Science*, 7(2), 205-216.
- Çırak, G. (2012). Yüksek öğretimde öğrenci başarılarının sınıflandırılmasında yapay sinir ağları ve lojistik regresyon yöntemlerinin kullanılması. Yayınlanmış yüksek lisans tezi, Eğitim Bilimleri Enstitüsü, Ankara Üniversitesi, Ankara.
- Da Silva, I. N., Spatti, D. H., Flauzino, R. A., Liboni, L. H. B., and dos Reis Alves, S. F. (2017). Artificial neural networks. *Cham: Springer International Publishing*, 39.
- Daneshvar, D., and Behnood, A. (2022). Estimation of the dynamic modulus of asphalt concretes using random forests algorithm. *International Journal of Pavement Engineering*, 23(2), 250-260.
- Delamaire, L., Abdou, H., and Pointon, J. (2009). Credit card fraud and detection techniques: a review. *Banks and Bank systems*, 4(2), 57-68.
- Dhankhad, S., Mohammed, E., and Far, B. (2018, July). Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In *2018 IEEE international conference on information reuse and integration (IRI)* (pp. 122-125). IEEE.
- Dilki, G., and Başar, Ö. D. (2020). İŞLETMELERİN İFLAS TAHMİNİNDE K-EN YAKIN KOMŞU ALGORİTMASI ÜZERİNDEN UZAKLIK ÖLÇÜTLERİNİN KARŞILAŞTIRILMASI. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 19(38), 224-233.

- Dornadula, V. N., and Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. *Procedia computer science*, 165, 631-641.
- Ebru, A. K. Ö. S. T. (2011). Endeks getirisi yönünün ikili sınıflandırma yöntemiyle tahmin edilmesi: İMKB 100 endeksi örneği. *Niğde Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 4(2), 45-59.
- Ersöz, F., and ÇINAR, Y. (2021). Veri Madenciliği ve Makine Öğrenimi Yaklaşımlarının Karşılaştırılması: Tekstil Sektöründe bir Uygulama. *Avrupa Bilim ve Teknoloji Dergisi*, (29), 397-414.
- Filiz, E. Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasına yönelik tahmininin destek vektör makineleri ve k-en yakın komşu algoritmaları ile gerçekleştirilmesi. *Gümüşhane Üniversitesi Fen Bilimleri Dergisi*, 12(1), 370-379.
- Friedl, M. A., and Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3), 399-409.
- Gupta, B., Rawat, A., Jain, A., Arora, A., and Dharni, N. (2017). Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications*, 163(8), 15-19.
- Güner N., Çomak E., (2011), Mühendislik Öğrencilerinin Matematik I Derslerindeki Başarısının Destek Vektör Makineleri Kullanılarak Tahmin Edilmesi, 17(2), 87-96.
- Güran A., Uysal M., Doğrusöz Ö., (2014) Destek Vektör Makineleri Parametre Optimizasyonunun Duygu Analizi Üzerindeki Etkisi, 16(48), 86-93.
- Ito, F., and Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, 13(4), 1503-1511.
- Jadhav, S. D., and Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR)*, 5(1), 1842-1845.
- Jain, V., Agrawal, M., and Kumar, A. (2020, June). Performance analysis of machine learning algorithms in credit cards fraud detection. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 86-88). IEEE.
- Jain, Y., Tiwari, N., Dubey, S., and Jain, S. (2019). A comparative analysis of various credit card fraud detection techniques. *Int J Recent Technol Eng*, 7(5S2), 402-407.
- Jawad, J., Hawari, A. H., and Zaidi, S. J. (2021). Artificial neural network modeling of wastewater treatment and desalination using membrane processes: A review. *Chemical Engineering Journal*, 419, 129540.
- Jin y., Liu x., Chen Y., and Liang X., Land-cover mapping using Random Forest classification and incorporating NDVI time-series and texture: a case study of central Shandong 1-21, Land-cover mapping using Random Forest classification and incorporating NDVI time-series and texture: a case study of central Shandong: International Journal of Remote Sensing: Vol 39, No 23 (tandfonline.com)
- Karagül, K. (2014). İstanbul Menkul Kıymetler Borsası'nda İşlem Gören Firmaların Destek Vektör Makineleri Kullanılarak Sınıflandırılması. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 20(5), 174-178.
- Karakulah, M., DANACI, M., and Ciritci, İ. H. (2011). BİYOMETRİK PARMAK İZİNİN AKILLI KARTLARLA KULLANIMI VE UYGULAMASI. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 10(4), 13-16.

- Kavzoğlu, T., and Çölkesen, İ. (2010). Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi. *Harita Dergisi*, 144(7), 73-82.
- Kazancı (2014), Yargıtay 13. Ceza Dairesi'nin 15.04.2014 tarihli 2013/13961 esas ve 2014/14051 karar sayılı kararı, Erişim: 12.05.2022, <https://www.kazanci.com.tr/>
- Kazancı (2019), Yargıtay 15.Ceza Dairesi'nin 19.09.2019 tarihli, 2017/11313 esas ve 2019/8602 karar sayılı kararı, Erişim: 12.05.2022, <https://www.kazanci.com.tr/>
- Kazancı (2019), Yargıtay Yargıtay 5.Ceza Dairesi'nin 08.10.2019 tarihli, 2015/5933 esas ve 2019/9440 karar sayılı kararı, Erişim: 12.05.2022, <https://www.kazanci.com.tr/>
- Kellie J.Archer ve Ryan V.Kimes, "Empirical characterization of random forest variable importance measures", *Computational Statistics and Data Analysis*, 52(3),2249-2260
- Keskin, M. V., and YILDIZ, D. (2018). BÜYÜK VERİ ARAÇLARI VE R KULLANARAK AMERİKAN HAVAYOLU FİRMALARININ SORUNLARININ KEŞFEDİLMESİ. *Uygulamalı Sosyal Bilimler Dergisi*, 2(2), 1-19.
- Khan, A., Baharudin, B., Lee, L. H., and Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1), 4-20.
- Kolluri, J., Kotte, V. K., Phridviraj, M. S. B., and Razia, S. (2020, June). Reducing overfitting problem in machine learning using novel L1/4 regularization method. In *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)*(48184) (pp. 934-938). IEEE.
- Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1), 3-24.
- Kursa, M. B., and Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of statistical software*, 36, 1-13.
- Küçüksille, E. U., ve Ateş, N. (2013). Destek Vektör Makineleri ile Yaramaz Elektronik Postaların Filtrelenmesi. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 6(1).
- Lakshmi, S. V. S. S., and Kavilla, S. D. (2018). Machine learning for credit card fraud detection system. *International Journal of Applied Engineering Research*, 13(24), 16819-16824.
- Lexpera (2014), Yargıtay 15. Ceza Dairesi'nin 19.02.2014 tarihli 2012/9345 esas ve 2014/3022 karar sayılı kararı, Erişim: 12.05.2022, <https://www.lexpera.com.tr/>.
- Liu, H., Song, D., Rüger, S., Hu, R., and Uren, V. (2008, January). Comparing dissimilarity measures for content-based image retrieval. In *Asia Information Retrieval Symposium* (pp. 44-50). Springer, Berlin, Heidelberg.
- Ma, S., and Du, T. (2010, March). Improved adaboost face detection. In *2010 International Conference on Measuring Technology and Mechatronics Automation* (Vol. 2, pp. 434-437). IEEE.
- Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*. [Internet], 9, 381-386.
- Manlangit, S., Azam, S., Shanmugam, B., Kannoorpatti, K., Jonkman, M., and Balasubramaniam, A. (2017, December). An efficient method for detecting fraudulent transactions using classification algorithms on an anonymized credit card data set. In *International Conference on Intelligent Systems Design and Applications* (pp. 418-429). Springer, Cham.

- Maratea, A., Petrosino, A., and Manzo, M. (2014). Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences*, 257, 331-341.
- Marcot, B. G., and Hanea, A. M. (2021). What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?. *Computational Statistics*, 36(3), 2009-2031.
- Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., and Doulamis, N. (2021). Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem. *Technologies*, 9(4), 81.
- Maxwell, A. E., Warner, T. A., and Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, 39(9), 2784-2817.
- Nguyen, Q. (2008). Machine Learning Algorithms and applications. *University of Bozen-Bolzano*, 2009.
- Nilson Report (2022). *Global Network Cards–Purchase Transactions with Market Shares* Erişim:12.11.2022, https://nilsonreport.com/publication_chart_and_graphs_archive.php
- Nilson Report (2022). *Annual Fraud Report* Erişim: 12.11.2022, https://nilsonreport.com/publication_newsletter_archive_issue.php?issue=1187
- Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., and Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3), 128-138.
- Pal M., (2005), Random forest classifier for remote sensing classification, *International Journal Of Remote Sensing*, 26(1), 217-222
- Pala, M. A., Çimen, M. E., Boyraz, Ö. F., Yildiz, M. Z., and Boz, A. F. (2019). Meme Kanserinin Teşhis Edilmesinde Karar Ağacı Ve KNN Algoritmalarının Karşılaştırmalı Başarım Analizi. *Academic Perspective Procedia*, 2(3), 544-552.
- Potur E., Erginel N., (2021) Kalp Yetmezliği Hastalarının Sağ Kalımlarının Sınıflandırma Algoritmaları ile Tahmin Edilmesi, 24, 112-118
- Rehman Javed, A., Jalil, Z., Atif Moqurrah, S., Abbas, S., and Liu, X. (2020). Ensemble adaboost classifier for accurate and fast detection of botnet attacks in connected vehicles. *Transactions on Emerging Telecommunications Technologies*, e4088.
- Sailusha, R., Gnaneswar, V., Ramesh, R., and Rao, G. R. (2020, May). Credit card fraud detection using machine learning. In *2020 4th international conference on intelligent computing and control systems (ICICCS)* (pp. 1264-1270). IEEE.
- SELİMOĞLU, M., ve YILMAZ, A. (2021). Kredi Kartı Dolandırıcılık Tespitinin Makine Öğrenmesi Yöntemleri ile Tahmin Edilmesi. *Beykent Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi*, 13(2), 28-33.
- Shirgave, S., Awati, C., More, R., and Patil, S. (2019). A Review On Credit Card Fraud Detection Using Machine Learning. *International Journal of Scientific and technology research*, 8(10), 1217-1220.
- Singh, S., and Gupta, P. (2014). Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey. *International Journal of Advanced Information Science and Technology (IJAIST)*, 27(27), 97-103.
- Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation.

- In *Australasian joint conference on artificial intelligence* (pp. 1015-1021). Springer, Berlin, Heidelberg.
- Solmaz R., Günay M., Alkan A., Fonksiyonel Tiroit Hastalığı Tanısında Naive Bayes Sınıflandırıcının Kullanılması, (2014)
- Song, Y. Y., and Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2), 130.
- Sonobe R., Tani H., Wang X., Kobayashi., Shimamura H., (2014), Thus, by using a given combination of features, a decision tree is made to grow up to its maximum depth with no pruning, 5, 157-164 <http://dx.doi.org/10.1080/2150704X.2014.889863>
- Soofi, A. A., and Awan, A. (2017). Classification techniques in machine learning: applications and issues. *Journal of Basic and Applied Sciences*, 13, 459-465.
- Suzuki, K. (Ed.). (2011). *Artificial neural networks: methodological advances and biomedical applications*. BoD–Books on Demand.
- Şenel, S. ve Atlı, B. (2014). Lojistik Regresyon Analizinin Kullanıldığı Makaleler Üzerine Bir İnceleme. *Eğitimde ve Psikolojide Ölçme Ve Değerlendirme Dergisi*, 5(1), 35-52.
- Tang, D., Tang, L., Dai, R., Chen, J., Li, X., and Rodrigues, J. J. (2020). MF-Adaboost: LDoS attack detection based on multi-features and improved Adaboost. *Future Generation Computer Systems*, 106, 347-359.
- Tang, J., Alelyani, S., and Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, 37.
- Tanouz, D., Subramanian, R. R., Eswar, D., Reddy, G. P., Kumar, A. R., and Praneeth, C. V. (2021, May). Credit card fraud detection using machine learning. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 967-972). IEEE.
- Tekin, M. C., and TUNALI, V. (2019). Yazılım geliştirme taleplerinin metin madenciliği yöntemleriyle önceliklendirilmesi. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 25(5), 615-620.
- Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.
- Trivedi, N. K., Simaiya, S., Lilhore, U. K., and 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), 3414-3424.
- Üstüner M., Abdikan S., Bilgin G., Şanlı F., Hafif Gradyan Artırma Makineleri ile Tarımsal Ürünlerin Sınıflandırılması, 1(2), 97-105
- Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., and Anderla, A. (2019, March). Credit card fraud detection-machine learning methods. In *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)* (pp. 1-5). IEEE.
- Wang, H., & Wang, G. (2020). Improving random forest algorithm by Lasso method. *Journal of Statistical Computation and Simulation*, 91(2), 353–367. <https://doi.org/10.1080/00949655.2020.1814776>
- Watts, J. D., Powell, S. L., Lawrence, R. L., & Hilker, T. (2011). Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery. *Remote Sensing of Environment*, 115(1), 66–75. <https://doi.org/10.1016/j.rse.2010.08.005>
- Wikipedia (2022). Biological Neuron Model. Erişim: 13.11.2022. https://en.wikipedia.org/wiki/Biological_neuron_model

Wu, W. T., Li, Y. J., Feng, A. Z., Li, L., Huang, T., Xu, A. D., and Lyu, J. (2021). Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Military Medical Research*, 8(1), 1-12.

Yakut, Y. B. E. T. Y., YAKUT, E., and Yavuz, S. (2014). Yapay Sinir Ağları Ve Destek Vektör Makineleri Yöntemleriyle Borsa Endeksi Tahmini. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 19(1), 139-157.

5237 Sayılı Türk Ceza Kanunu, <https://www.mevzuat.gov.tr/MevzuatMetin/1.5.5237.pdf>



ÖZ GEÇMİŞ

Serhat DEMİRTÜRK, Samsun Recep TANRIVERDİ Lisesi'ni bitirdikten sonra Ondokuz Mayıs Üniversitesi, Fen Edebiyat Fakültesi, İstatistik bölümünden 28.05.2018 yılında mezun oldu. 2020 yılında OMÜ LEE İstatistik Ana Bilim Dalı Yüksek Lisans programına girdi. Mezuniyetinden bu yana İş Zekası ve Raporlama Analisti olarak görev yapan Serhat DEMİRTÜRK, iyi düzeyde İngilizce bilmektedir. Temel ilgili alanları; makine öğrenimi, iş zekası ve uçtan uca veri tabanı mimarisi tasarımıdır.

İletişim Bilgileri

ORCID ID : 0000-0003-3572-8213

Yayınlar:

1. Dünder, E. ve Demirtürk, S. (2022), Detection of Fraud of Transport Cards by Classification Methods in Machine Learning, *New Trends In Computational Statistics*, Duvar Yayınları, Chapter 7, 2022, s. 87-96.