



**T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**



YÜKSEK LİSANS TEZİ

**FİNANS SEKTÖRÜNDE VERİ MADENCİLİĞİ TEKNİKLERİ
KULLANILARAK KAMPANYA MODELLEME**

Arca ÇAKMAKOĞLU

Endüstri Mühendisliği Anabilim Dalı

Endüstri Mühendisliği Programı

**DANIŞMAN
Doç. Dr. Tarık KÜÇÜKDENİZ**

Haziran, 2018

İSTANBUL

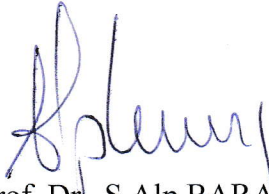
Bu çalışma, 7.06.2018 tarihinde aşağıdaki jüri tarafından Endüstri Mühendisliği Anabilim Dalı

, Endüstri Mühendisliği Programında Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi



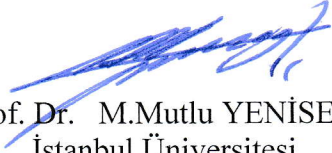
Doç. Dr. Tarık KÜÇÜKDENİZ(Danışman)
İstanbul Üniversitesi
Mühendislik Fakültesi



Prof. Dr. Ş.Alp BARAY
İstanbul Üniversitesi
Mühendislik Fakültesi



Prof. Dr. Şakir ESNAF
İstanbul Üniversitesi
Mühendislik Fakültesi



Prof. Dr. M.Mutlu YENİSEY
İstanbul Üniversitesi
Mühendislik Fakültesi



Dr. Öğr. Üyesi S.Kerem AYTULUN
Beykent Üniversitesi
Mühendislik-Mimarlık Fakültesi



20.04.2016 tarihli Resmi Gazete’de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi’nin aboneli olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü’nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

ÖNSÖZ

Tez çalışmalarım boyunca, sürecin her aşamasında bilgi ve tecrübelerini paylaşan, değerli zamanını benden esirgemedi bana her zaman destek olan değerli danışman hocam Doç. Dr. Tarık KÜÇÜKDENİZ' e teşekkürlerimi sunarım.

Yüksek lisans öğrenimim boyunca bilgi ve birikimlerini paylaşan Endüstri Mühendisliği bölümündeki hocalarıma teşekkür ederim.

Tez sürecinde desteğini esirgemeyen Selma KAYACAN' a ve hayatımın her döneminde benden desteklerini esirgemeyen her zaman yanımda olan aileme teşekkürlerimi sunarım.

Haziran, 2018

Arca ÇAKMAKOĞLU

İÇİNDEKİLER

Sayfa No

ÖNSÖZ	iv
İÇİNDEKİLER.....	v
ŞEKİL LİSTESİ	vii
TABLO LİSTESİ.....	ix
SİMGE VE KISALTMA LİSTESİ	xi
ÖZET	xiv
SUMMARY	xvi
1. GİRİŞ	1
2. GENEL KISIMLAR.....	3
2.1. PAZARLAMA	3
2.1.1. Pazarlama Karması	6
2.1.2. Doğrudan Pazarlama	7
2.2. VERİ MADENCİLİĞİ.....	11
2.2.2. Veri Madenciliği Süreci.....	13
2.2.2.1. Problemin tanımlanması.....	14
2.2.2.2. Verinin anlaşılması	14
2.2.2.3. Veri işleme süreci	14
2.2.2.4. Modelin kurulması	16
2.2.2.5. Modelin değerlendirilmesi ve seçilmesi.....	16
2.2.2.6. Modelin uygulanması ve yorumlanması	21
2.2.3. Veri Madenciliği Yöntemleri	21
2.2.3.1. Sınıflandırma	23
2.2.3.2. Regresyon	24
2.2.3.3. Kümeleme	24
2.2.3.4. Birliktelik kuralları	25
2.2.3.5. Ardışık zaman örüntüleri	25
2.2.4. Yapay Sinir Ağları.....	26
2.2.5. Aşırı Öğrenme Makinesi	29
2.2.6. Genetik Algoritma	36
2.2.7. Karar Ağacı Sınıflandırıcısı.....	38

2.2.8. K En Yakın Komşu Algoritması	41
2.2.9. Naive Bayes.....	41
2.2.10. Gradyen Artırma Modeli	42
2.2.11. Rasgele orman	42
2.2.12. Lojistik Regresyon.....	42
2.2.13. Destek Vetör Makineleri	43
2.2.14. Lineer Diskriminant Analizi.....	44
2.2.15. Model Performansının Topluluk Metotlarıyla Geliştirilmesi.....	44
2.2.15.1. Oylama.....	46
2.2.15.2. Torbalama.....	47
2.2.15.3. Artırma.....	48
3. MALZEME VE YÖNTEM.....	50
3.1. PROBLEMİN TANIMLANMASI	50
3.2. KULLANILAN YAZILIM GELİŞTİRME ORTAMI	52
3.3. VERİNİN ANLAŞILMASI	53
3.4. VERİ İŞLEME	55
3.4.1. Nitelik Seçimi.....	58
3.5. MODELİN KURULMASI.....	59
3.6. MODEL PERFORMANS DEĞERLENDİRME YÖNTEMLERİ.....	60
3.7. MODELİN UYGULANMASI VE YORUMLANMASI	61
4. BULGULAR.....	62
4.1. VERİLER ARASINDAKİ İLİŞKİLERİN İNCELENMESİ.....	62
4.2. UYGULAMA	69
4.2.1. Parametre Değerleriyle Sınıflandırıcıların Değerlendirilmesi.....	70
4.2.2. Nitelik Seçimi Sonrasında Varsayılan Parametre Değerleriyle Sınıflandırıcıların Değerlendirilmesi	75
4.2.3. Aşırı Öğrenme Makinesi Sınıflandırma Algoritması ile Uygulama.....	77
4.2.3.1. Genetik algoritma ile AÖM parametreleri optimizasyonu	80
4.2.3.2. Torbalama yöntemiyle modelin geliştirilmesi	81
5. TARTIŞMA VE SONUÇ	83
KAYNAKLAR.....	86
ÖZGEÇMİŞ	94

ŞEKİL LİSTESİ

	Sayfa No
Şekil 2.1: Bilgi keşfi sürecinde veri madenciliği (Fayyad ve diğ. 1996).....	14
Şekil 2.2: Model uyumu.....	16
Şekil 2.3: Çapraz geçerleme gösterimi.....	18
Şekil 2.4: ROC eğrisi örneği.	21
Şekil 2.5: Tanımlayıcı ve tahminleyici veri madenciliği yöntemleri.....	23
Şekil 2.6: Yapay sinir ağlarının katmanları.	27
Şekil 2.7: Aşırı öğrenme makinesi.....	33
Şekil 2.8: Karar ağacı gösterimi.....	38
Şekil 2.9: Oylama yöntemi.....	47
Şekil 2.10: Torbalama yöntemi.....	48
Şekil 2.11: Topluluk algoritması uygulanmadan önce ve sonra model gösterimi.	48
Şekil 3.1: Veri analizi süreci akış şeması.....	51
Şekil 3.2: Programa ait ekran görüntüsü (bankacılık verisi ile).....	52
Şekil 3.3: Veri setinin programda gösterimi.	54
Şekil 3.3 (devam): Veri setinin programda gösterimi.	55
Şekil 3.3 (devam): Veri setinin programda gösterimi.	55
Şekil 3.4: Nitelik seçimi.....	59
Şekil 4.1: Veri dağılımlarının incelenmesi.	62
Şekil 4.2: Kampanyaya katılımın aylara göre dağılımı.....	63
Şekil 4.3: Aylara göre kampanya başarısı.....	64
Şekil 4.4: Mevsimlere göre kampanya başarısı.....	64
Şekil 4.5: Yaşlara göre müşteri sayısı dağılımı.....	65
Şekil 4.6: Kampanyaya katılım durumunun yaş ve mesleklerle ilişkisi.	66

Şekil 4.7: Potansiyel müşterilerin meslek dağılımları.....	67
Şekil 4.8: Korelasyon matrisi.....	69



TABLO LİSTESİ

	Sayfa No
Tablo 2.1: Model değerlendirme ölçütleri.	19
Tablo 2.2: Model değerlendirme ölçütleri.	20
Tablo 2.3: AÖM’ de kullanılabilen aktivasyon fonksiyonları.	36
Tablo 3.1: Ön işleme öncesi veri.	53
Tablo 3.1 (devam): Ön işleme öncesi veri.	54
Tablo 3.2: Ön işlem sonrası veri seti.	56
Tablo 3.3: Veri seti özeti (Nitelik seçimi öncesi).	57
Tablo 3.4: Nitelik ilişkileri.	59
Tablo 3.5: Python’ da kullanılan algoritmalar.	60
Tablo 4.1: Yaş kategorisine göre kampanya katılım oranları.	65
Tablo 4.2: Meslek gruplarının kampanyaya katılım oranları.	67
Tablo 4.3: Ev kredisi olan müşterilerin kampanyaya katılım oranları.	68
Tablo 4.4: Sınıflandırıcı ve regresör parametreleri.	70
Tablo 4.4 (devam): Sınıflandırıcı ve regresör parametreleri.	71
Tablo 4.4 (devam): Sınıflandırıcı ve regresör parametreleri.	72
Tablo 4.4 (devam): Sınıflandırıcı ve regresör parametreleri.	73
Tablo 4.5: Sınıflandırma performans sonuçları.	74
Tablo 4.5: Karmaşıklık matrisi sonuçları.	74
Tablo 4.7: Sınıflandırma algoritmalarına göre eğitim süreleri.	75
Tablo 4.8: Sınıflandırma performans sonuçları.	76
Tablo 4.9: Doğruluk oranları değişimi.	76
Tablo 4.10: Gizli nöron sayısı değişikliklerinin sınıflandırmaya etkisi.	78
Tablo 4.11: Alpha değeri değişikliklerinin sınıflandırmaya etkisi.	79

Tablo 4.12: Aktivasyon fonksiyonu deęişikliklerinin sınıflandırmaya etkisi.	79
Tablo 4.13: Alt aktivasyon fonksiyonu deęişikliklerinin sınıflandırmaya etkisi.	80
Tablo 4.14: Regresör deęişiklięinin sınıflandırmaya etkisi.....	80
Tablo 4.15: Topluluk algoritmaları sonuçları.....	82
Tablo 5.1: Kullanılan veri seti ile yapılan dięer alıřmaların sonuçları.	84
Tablo 5.2: Uygulamada ulařılan sonuçların doęruluk oranları.	84



SİMGE VE KISALTMA LİSTESİ

Simgeler	Açıklama
A_{kj}	: Girdi Elemanını j Ara Katman Elemanına Bağlayan Bağlantının Ağırlık Değeri
j	: Ara Katman Elemanın Çıktısı
N	: Rasgele Seçilen Eğitim Verisi Seti Örnekleri
$A_{kj} \cdot \zeta_k^i$	: Her Bir Girişin İlgili Nöronla Bağlantısını
w_i	: i . Giriş Sinir Hücresi ve i . Gizli Sinir Hücresine Bağlı Ağırlık Vektörü
δ	: Hata Kareleri
ρ	: Adım Uzunluğunu
β	: i . Gizli Sinir Hücresine ve Çıkış Sinir Hücrelerine Bağlı Olan Ağırlık Vektörü
$\hat{\beta}$	: β 'nin Tahmini Değeri
b_i	: i . Gizli Düğümün Rasgele Seçilen Eşik Değeri
$g(x)$	: Aktivasyon Fonksiyonu Vektörü
H	: Gizli Katman Çıktı Matrisi
H^+	: H Matrisinin Moore–Penrose Genelleştirilmiş Tersidir
H	: Entropi
$p(x)$	: Belirli Bir Sınıfa Ait Grubun Yüzdesi
$Gain(S, D)$	: Bilgi Kazancı
S	: Orijinal Veri Kümesidir
D	: Veri Kümesinin Bölünmüş Bir Parçası
V	: S Kümesinin Alt Kümeleri

Kısaltmalar	Açıklama
ADWIN	: Adaptif Kayan Pencere
ALIFT	: LIFT Eğrisi Altında Kalan Alan
AÖM	: Aşırı Öğrenme Makinesi
ARCH	: Otoregresif Koşullu Heteroskedastik
AUC	: Eğri Altında Kalan Alan
CART	: Classification and Regression Trees
ÇKA	: Çok Katmanlı Algılayıcı
EOS-ELM	: Çevrimiçi Ardışık Aşırı Öğrenme Makinesi
FP	: Yanlış Pozitif
FN	: Yanlış Negatif
GARCH	: Genelleştirilmiş Otoregresif Koşullu Heteroskedastik
GFF	: Genel İleri Beslemeli
HI	: Hibrit Zeka
I-ELM	: Artışlı Aşırı Öğrenme Makinesi
İGT-AÖM	: İleri-Geri Takip Algoritması Aşırı Öğrenme Makinesi
KNN	: K-en Yakın Komşu
LDA	: Doğrusal Diskriminant Analizi
MIDSO	: Karışık ve Eksik Benzerlik Operatörü
MLP	: Çok Katmanlı Algılayıcı
OP-ELM	: Çok Parçalı Aşırı Öğrenme Makinesi
OS-ELM	: Çevrimiçi Ardışık Aşırı Öğrenme Makinesi
RBF	: Radyal Tabanlı Yapay Sinir Ağları
RELIEF	: Kabartma Tekniği
RIPPER	: Artımlı Budama İçin Hata Azaltma
ROC	: Alıcı Çalışma Eğrisi
SLFN	: Tek Gizli Katmanlı İleri Beslemeli Sinir Ağları
SMOTE	: Sentetik Azınlık Aşırı Öğrenme Tekniği
SVM	: Karar Destek Makineleri
TP	: Doğru Pozitif
TN	: Doğru Negatif
UCI	: Kaliforniya Irvine Üniversitesi

YSA : Yapay Sinir Ağları



ÖZET

YÜKSEK LİSANS TEZİ

FINANS SEKTÖRÜNDE VERİ MADENCİLİĞİ TEKNİKLERİ KULLANILARAK KAMPANYA MODELLEME

Arca ÇAKMAKOĞLU

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Endüstri Mühendisliği Anabilim Dalı

Danışman : Doç. Dr. Tarık KÜÇÜKDENİZ

İçinde bulunduğumuz bilgi çağında, teknolojinin getirdiği bilgi toplumu unsurları hayatın her alanında kendini göstermektedir. Her geçen gün yeni kavramlar ve çalışma alanları ortaya çıkmaktadır. Gelecek dönemlerde en önemli teknolojik pazarın veri üzerinde oluşması beklenmektedir. Gelişen iletişim ve bilgi teknolojileri ile web sayfaları, bloglar, kameralar, sosyal medya uygulamaları, mobil uygulamalar, pek çok akıllı cihaz ve internet tabanlı uygulamalar tarafından sürekli olarak veri toplanmaktadır. Teknolojiye bağlı olarak farklı kaynaklar tarafından üretilen çok boyutlu ve büyük alan kaplayan veri ortaya çıkmıştır. Veri boyutundaki hızlı büyüme, temeli istatistiğe dayanan veri madenciliğinin önemini ortaya çıkarmıştır. Biriken büyük miktarda ve çeşitlilikteki verinin saklanması, işlenmesi ve veri üzerinden değerli bilgi elde etme için geleneksel yöntemler yetersiz kalmaktadır. Büyük veri teknolojileri sayesinde daha hızlı, daha şeffaf, daha etkin çözümlerle birçok sektörde iş süreçlerinde iyileştirme noktaları, fırsatlar, verimli iş modelleri, sunulan hizmetlerin iyileştirilmesi ile müşteri memnuniyeti, yüksek kârlılık ve rekabet avantajı sağlanmaktadır. Veri madenciliği alanında yapılan çalışmaların, hemen hemen her sektörde olduğu gibi bankacılık sektöründe de kullanımının artması ve daha büyük kazanımlar elde edilmesi beklenmektedir.

Tez çalışmasında bir bankanın geçmiş pazarlama verisiyle yapılan analizlerde veri madenciliği yöntemlerinden Aşırı Öğrenme Makinesini kullanarak, vadeli para yatırma kampanyasına katılabilecek müşteriler ile yapılan telefon görüşmeleri için sınıflandırma yapılmıştır. Böylece tele pazarlama faaliyetlerinde kaynakların doğru müşteri kitlesine ulaşmak için kullanılarak satış fırsatını artırmak, aramalardan kaynaklı müşteri memnuniyetsizliğini azaltmak ve arama maliyetlerini düşürmek amaçlanmıştır. Bu kapsamda, Aşırı Öğrenme Makinesi sınıflandırıcısı parametreleri model ayarlama metotları olarak ızgara arama ve genetik algoritma ile optimizasyon yapılmış, en yüksek doğrulukta sınıflandırma oranına yaklaşılmaya çalışılmıştır. Son olarak, topluluk algoritmalarından torbalama yöntemiyle model geliştirilmiş, sonuçları incelenmiştir.

Bu tez çalışması, finans sektöründe veri madenciliği metotlarının uygulanmasında Aşırı Öğrenme Makinesi algoritmasının kullanılmasının yaygınlaşması ve nasıl uygulanabileceği konusunda örnek bir çalışma ortaya koymuştur.

Haziran 2018, 111 sayfa.

Anahtar kelimeler: Veri madenciliği, sınıflandırma yöntemleri, kampanya modelleri, aşırı öğrenme makinesi.

SUMMARY

M.Sc. THESIS

CAMPAIGN MODELING IN FINANCE SECTORE BY USING

DATA MINING TECHNIQUES

Arca ÇAKMAKOĞLU

İstanbul University

Institute of Graduate Studies in Science and Engineering

Department of Industrial Engineering

Supervisor : Assoc. Prof. Dr. Tarık KÜÇÜKDENİZ

Today, with the developing technology, the elements of information society stands out in every area. Every day new concepts and new working areas arise. In the next periods, it is expected that the technology sector will be focused on data. With the developing technology of information and communication, websites, blogs, cameras, social media applications, mobile applications and any other internet based applications and smart devices collect these data. With the effects of technology, there is a huge amount of data that is produced by different sources. The increase in the size of data leads the importance of data mining that is based on the statistics. Traditional methods are inadequate in processing data and gathering important data because of storing different types and the size of data. With the big data technologies that are more effective, transparent and quicker; it is possible to maintain effective business models, get opportunities and business process improvements. On the other hand, with the service improvements, it is possible to maintain customer satisfaction, competitive edge, and a high profit. As it is expected in every sector in data mining, the usage in the banking sector will increase and higher gains will be obtained.

In the thesis study, the classification for phone calls of the customers that could join to a bank term deposit has been studied. For that, a data mining method, extreme Learning Machine, has

been used as a classification method on the previous marketing data of a bank. With this study, it is aimed to reduce customer dissatisfaction caused by faulty calls, reduce operational costs and increase the sales opportunities by reaching the right customers' cluster for telemarketing activities. In this context, it is aimed to get the highest classification rate by applying model tuning which is grid search and genetic algorithm, on the classification parameters of Extreme Learning Machine. Finally, the model has been developed with the bagging method which is an ensemble algorithm and the results have been studied.

This study shows an example of how Extreme Learning Machine as a data mining methods can be used in the finance sector and how it can be applied.

June 2018, 111 pages.

Keywords: Data mining, classification methods, campaign models, extreme learning machine.



1. GİRİŞ

İçinde bulunduğumuz bilgi çağında, teknolojinin getirdiği bilgi toplumu unsurları hayatın her alanında kendini göstermektedir. Her geçen gün yeni kavramlar ve çalışma alanları ortaya çıkmaktadır. Gelişen iletişim ve bilgi teknolojileri ile web sayfaları, bloglar, akıllı sistem sensörleri, sosyal medya uygulamaları, mobil uygulamalar, pek çok akıllı cihaz ve internet tabanlı uygulamalar sayesinde sürekli olarak veri toplanmaktadır. Gelecek yıllarda en önemli teknolojik pazarın veri üzerinde oluşması beklenmektedir. Teknolojiye bağlı olarak farklı kaynaklar tarafından üretilen, çok boyutlu ve büyük alan kaplayan veri ortaya çıkmıştır. Veri boyutundaki hızlı büyüme, temeli istatistiğe dayanan veri madenciliğinin önemini ortaya koymaktadır.

Veri madenciliği, verileri farklı yöntemlerle analiz etme ve o verilerden anlamlı, değerli bilgiler elde etme süreci olarak tanımlanabilir. Veri madenciliği yöntemleriyle büyük veri seti içerisinde anlamlı ilişkiler ve yeni ilişkiler keşfedilmeye çalışılır ve değerli bilgiye ulaşılabilir.

Elektronik ortamda toplanan ve saklanan veriler üzerinde yapılan çalışmalar için klasik istatistiksel yöntemler yetersiz kalmaya başlamış, toplanan veri analiz edilemez boyutlara ulaşmıştır. Bu verilerden değerli bilgilere ulaşmak için anlamlı ve yararlı kurallar tespit edebilmek amacıyla veri madenciliği alanında birçok teknik geliştirilmiş ve bu tekniklerin uygulanabileceği farklı bilgi sistemi çözümleri üretilmiştir. Veri madenciliği sürecinde veri toplama, veri işleme, veriyi kullanmaya hazır hale getirme, veriyi saklama ve analiz etme gibi adımlar bulunmaktadır.

Pazarlama, güvenlik, halkla ilişkiler, sağlık, bankacılık, uygulama yönetimi, finans gibi pek çok sektörde veri madenciliği ile ilgili çalışmalar yapılmaktadır. Pazarlama açısından eldeki müşteri sadakatini devam ettirme ve yükseltme, satış ve kampanya fırsatları yaratma, yeni müşteri kazanma gibi konularda yoğun olarak veri madenciliği yöntemlerinden faydalanılmaktadır. Pazarlamada kullanılan yöntemler araştırıldığında, bunların genel olarak “Kitlesel Pazarlama” ve “Doğrudan Pazarlama” olmak üzere iki başlıkta incelendiği görülmüştür. Bilgi çağının getirdiği yeniliklerden bir diğeri ise kişiye özel ürün yaklaşımıdır. Doğrudan pazarlama, işletmeleri müşterileri için kişisel çözümler üretmeye yönlendirilmiştir. Veri madenciliği bu aşamada geçmiş dönem verileri ile hedef müşteri setinin belirlenmesi, mevcut müşterilere daha

iyi hizmet sağlama, işletme kaynak yönetiminde verimlilik, müşteri ihtiyaçlarını tahminleme gibi doğrudan pazarlamayı ilgilendiren birçok konu ile iç içedir. Doğrudan pazarlama için birçok kanal kullanılmaktadır. En etkin olanlardan biri tele pazarlamadır.

Bu çalışmada veri madenciliğinin pazarlama ve bankacılık alanında kullanımı incelenmiş ve tele pazarlama üzerine bir uygulama yapılması amaçlanmıştır. Kaliforniya Üniversitesi Irvine Makine Öğrenimi veri tabanındaki (UCI) banka pazarlama veri seti güncel yazılım dillerinden biri olan Python kullanılarak Portekiz’de bir bankanın pazarlama kampanyalarıyla ilgili analizler yapılmıştır.

Tez çalışmasında, veri madenciliği yöntemleri kullanılarak bankanın geçmiş pazarlama verisiyle yapılan analizlerde sınıflandırma yöntemi olarak Aşırı Öğrenme Makinesi sınıflandırıcısı kullanılıp, tele pazarlama kampanyasına katılacak müşterileri en iyi doğruluk oranı ile sınıflandırmak amaçlanmıştır. Doğrudan yapılan pazarlama aramaları müşteri memnuniyetsizliği oluşturabilir. Önerilen yöntemle arama yapılacak müşterilerin doğru olarak sınıflandırılması, doğru müşteri grubuna ulaşarak müşteri memnuniyetsizliğini azaltacak ve arama maliyetlerini düşürecektir. Ayrıca kaynakların verimli kullanımından doğan ek bütçe çalışanların eğitimi ya da teknoloji yatırımı gibi banka süreçlerini iyileştirecek alanlarda ek yatırım yapma imkanı sağlayacaktır.

Bu kapsamda Aşırı Öğrenme Makinesi sınıflandırıcısı parametreleri üzerinde yapılan değişikliklerle model ayarlama çalışması yapılmış, en yüksek doğrulukta sınıflandırma oranına ulaşmak için genetik algoritma yaklaşımı ile parametre seçimi gerçekleştirilmiştir. Ayrıca literatürde bu veri setiyle yapılan çalışmalarda kullanılan sınıflandırma yöntemleri karşılaştırılarak AÖM’ nin farkları değerlendirilmiştir.

2. GENEL KISIMLAR

2.1. PAZARLAMA

Birçok kiři pazarlamayı sadece satıř ve reklamcılık olarak düşünmektedir. Bu iki terim her ne kadar önemli olsa da pazarlama fonksiyonlarından sadece ikisidir ve çoęu zaman en önemlileri değildir. Günümüzde pazarlama eski anlamı ile anlatmak ve satmak olarak değil, yeni anlamı ile müşteri ihtiyaçlarını karşılamak olarak anlaşılmalıdır. Satıř sadece bir ürün üretildikten sonra gerçekleşir. Aksine pazarlama şirketin bir ürüne sahip olmasından çok daha önce başlar. Pazarlama, yöneticilerin, ihtiyaçları değerlendirmeyi, kapsamlarını ve yoğunluęunu ölçmeyi ve karlı bir fırsat olup olmadığını belirlemeyi üstlenen ev ödevidir. Pazarlama, ürün yaşamı boyunca devam etmekte, yeni müşteriler bulmaya çalışmakta ve ürün cazibesini ve performansını geliştirerek, ürün satıř sonuçlarından öğrenerek ve tekrar performanslarını yöneterek mevcut müşterileri korumaktadır (Kotler ve dię., 2005).

Bunlar göz önüne alındığında Kotler ve dię. (2018) tarafından pazarlama, şirketlerin müşteri ile güçlü ilişkiler kurduęu ve karşılığında müşteriden değer elde etmek için müşteri değeri yaratan süreç olarak tanımlanır.

Pazarlama, hedeflenen müşteri kitlesine değer sunacak hizmet ve ürünlerin üretilmesi ve pazara sunulmasını sağlayarak, ürünün pazardaki etkisi ile ilgili bilgilerin işletme içerisinde paylaşımı, müşteri ilişkilerinin yönetilebilmesi için ileri vadeli pazarlama stratejileri geliştirme ve tüm bu amaçları içerecek şekilde toplumsal kaynakların devamlılıęını sağlayarak işletmenin paydařlarını kapsayan değer temelli faaliyetlerin tümüdür.

Pazarlama günlük yaşamda pek çok alanda karşımıza çıkmaktadır ve bunula ilgili pek çok tanım yapılmıştır. Amerikan Pazarlama Derneęi 1985 yılında geleneksel pazarlama tanımını bırakıp, pazarlamanın kapsamını genişleterek yeniden tanımlamıştır. Buna göre: Pazarlama, kişisel ve örgütsel amaçlara ulaşmayı sağlayacak ilişkileri kurmak üzere, fikirlerin, malların ve hizmetlerin geliştirilmesi, fiyatlandırılması, tutundurulması ve dağıtılmasına ilişkin planlama ve uygulama sürecidir (Emgin ve Süngü, 2004).

Pazarlama artık, mal ve hizmetler gibi, fikirlerin de geliştirilmesiyle hedef kitlelere ulaşılmasını ve anlaşılmasını kolaylařtıran bir faaliyetler sistemi olarak görölmektedir. Ayrıca, bu faaliyetler sistemi, sadece kar amacı güden kuruluşlar olarak işletmeler tarafından değil, amaçlarına

ulaşmak için her türlü örgüt veya kuruluş tarafından yürütölmektedir. Diğer önemli husus da pazarlamanın, hedeflere ulaşmak için planlar yapma ve bu planları uygulama süreci ile, yani strateji ile yakından ilgili olmasıdır (Mucuk, 1997). Özel sektör ve akademik alanlarda veri madenciliğinin uygulandığı birçok çalışma bulunmaktadır. Bugüne kadar pazarlama alanında veri madenciliğı uygulamalarıyla ilgili yapılan çalışmalarda genel olarak müşteri ilişkileri yönetimi, satış tahminleme, kampanya yönetimi, müşteri sınıflandırma, kampanya uygulamaları konuları incelenmiştir.

Birant ve diğ. (2010) zincir mağazaları olan bir firmanın ürün ve satış verileri üzerinden ay, gün, şubeye göre hangi ürünlerin satıldığına dair analizler yapmıştır. Çalışmada birliktelik kuralları ile hangi ürünlerin bir arada satılabildiğini incelemiştir.

Ahi (2015), veri madenciliğı yöntemleriyle ana harcama gruplarının paylarının tahmini üzerine çalışma yapmıştır. Çalışmada k-en yakın komşu algoritması kullanılarak sınıflandırma tamamlanmıştır. Türkiye’ de bu alanda yapılan ilk analiz çalışmalarındandır.

Çil (2010), bir bankanın yatırım fonu müşteri hareketlerinin belirlenmesine yönelik bir çalışma ile daha önce bankanın müşterisi olup daha sonra fon sahibi olmayı bırakan müşterilerin davranışsal analizlerini veri madenciliğı ile incelemiştir. Sonuçlara göre fon hesabını kapatabilecek potansiyel müşterileri tahminlemeye çalışmıştır.

Zakaryad ve Duman (2016), optimizasyon teknikleri kullanarak yapay sinir ağı sınıflandırma tekniğini geliştirmek için kar odaklı bir yaklaşımı ele almıştır. Çalışmada 5 farklı yapay sinir ağı versiyonu ile kar odaklı yaklaşım karşılaştırılmış ve diğer yöntemlere göre daha yüksek doğruluk değerleri ile sınıflandırma sağlanmıştır. Analizde Weka yazılım programı kullanılmıştır.

Parlar ve Acaravcı (2017), yapılan çalışmada veri madenciliğı teknikleri kullanılarak kampanya etkinliği artırmıştır. Çalışmada analizin yapıldığı nitelik sayısı azaltılarak daha yüksek verim alınması hedeflenmiştir. Gain ile ki kare metotları kullanılarak nitelik seçimi üzerinde çalışma yapılmıştır. Naive Bayes algoritması kullanılarak ulaşılan sınıflandırma değerleri karşılaştırılmıştır.

Apampa (2016), banka pazarlama verisiyle yaptığı çalışmada rasgele orman kümeleme yöntemiyle kampanyaya katılımı tahminlemiştir. Rasgele orman algoritması dışında; lojistik regresyon, Naïve Bayes ve karar ağaçları algoritmalarıyla da karşılaştırma yapılmıştır. Deneysel yöntemlerden elde edilen sonuçlar, veriler dengelendiğinde rasgele orman algoritması performansının iyileştiğini ortaya koymuştur. Performans ölçüm yöntemlerinden sınıflandırma doğruluğu, AUC, F1 skor, geri çağırma ve hassaslık değerlerini kullanmıştır.

Ruangthong ve Jaiyen' in (2015) yaptığı çalışmada SMOTE ve Rotasyon Orman algoritmaları kullanılarak sınıflandırma yapılan bir metod önerilmiştir. Yöntem; bankacılık pazarlama veri seti kullanılarak satış tahminin doğruluğunu iyileştirmek için kullanılmıştır. Önerilen yöntemin performansı Karar Ağacı, Rotasyon Ormanı, Navie Bayes, Bayes Net, Çok Katmanlı Algılayıcı Sinir Ağı, RBF Sinir Ağı ile karşılaştırılmıştır. Yazılım programı olarak Weka kullanılmıştır.

Asare-Frempong ve Jayabalan'ın (2017) çalışmasında Sinir Ağları, Karar Ağaçları, Lojistik Regresyon ve Rasgele Orman sınıflandırma yöntemleri kullanılarak müşterilerin kampanyaya katılımı tahminlenmiştir. Çalışmada en verimli sınıflandırıcının Rasgele Orman sınıflandırıcısı olduğu görülmüş, ROC eğrisi ile performansı kontrol edilmiştir. Ayrıca kümeleme analizi ile en önemli özellikler belirlenmiş, görüşme uzunluğunun en önemli özellik olduğu ortaya çıkmıştır. Weka yazılım programı kullanılmıştır.

Villuende-Rey ve diğ. (2017), yeni bir denetimli öğrenme modeli olan Naive Çağrışımsal Sınıflandırıcısı ile üç farklı veri seti üzerinde çalışmıştır. Yapılan çalışma sonunda Karışık ve Eksik Benzerlik Operatörü (MIDSO) tasarımı ile yeni bir sınıflandırıcı oluşturulmuştur. Yeni model üç farklı veri seti ile ölçülmüş, sonuçları incelenmiştir.

Moro ve diğ. (2014), nitelik önceliklendirme sonrasında dört sınıflandırma algoritmasıyla veri madenciliği çalışması yapmıştır. Veri setinin UCI kütüphanesine eklenmesi bu çalışma ile sağlanmıştır. Orijinal veri setinde 150 nitelik ve 52.944 kayıt ile çalışma yapılmıştır. UCI' ye daha kısıtlı veri seti yüklenmiştir. Çalışma sonucunda AUC ve ALIFT değerleri karşılaştırılmış en yüksek performans değeri sinir ağlarında görülmüştür. Çalışma R yazılım programında yapılmıştır.

2.1.1. Pazarlama Karması

Pazarlama karması, seçilen hedef pazardaki tüketicilerin istek ve ihtiyaçları doğrultusunda müşteri tatminini sağlamak amacıyla işletmenin kontrolünde olan değişkenlerden oluşan karmadır (Özmen, 2013).

Pazarlama karması, pazarda rekabet üstünlüğü için işletmeler tarafından en büyük silah olarak kullanılmaktadır. James Culliton, pazarlama karmasını, pazarlama maliyetlerini açıklayan girdilerin karışık olarak ilk defa 1948 yılında ortaya çıkarmıştır. Bu açıklamadan sonra Neil Borden, 1953 senesinde pazarlama karmasında işletme yöneticilerini çeşitli unsurları birleştiren kişiler olarak belirtmiştir. Neil Borden pazarlama karmasını etraflıca incelemiş ve 12 bileşen (ürün planlama, ücretlendirme, markalandırma, dağıtım kanalı, kişisel satış, reklâm, tanıtımlar, paketleme, gösteri, hizmet verme, insan gücü ve analiz) olarak ortaya koymuştur. 1962 yılında pazarlama karması, pazarlama için kullanılan önemli terimlerden biri olmuştur. Jerome McCarthy tarafından 1964 yılında formüle edilen pazarlama karması, kavram içinde yer alan dört temel bileşeni (ürün, yer, fiyat, tanıtım) anlatan sözcüklerin İngilizce karşılıklarının (product, place, price, promotion) ilk harflerinden yola çıkılarak '4P' şeklinde sembolleştirmiştir (Öndoğan, 2010).

Bazı araştırmacılar tarafından klasik 4P anlayışının günümüzdeki modern pazarlama anlayışını karşılamakta yetersiz olduğu düşünülmüş ve 4P'ye üç P ilave edilmiştir. Böylece 7P kavramı ortaya çıkmıştır. İnsan (People), Süreçler (Process) ve Fiziksel Öğeler (Physical Evidences), 3P'yi oluşturur. Diğer taraftan işletme ya da pazarlamacı yönlü bir bakış açısına sahip olduğundan 4P'nin bileşenlerinin tüketici tarafı karşılıkları belirlenmiştir. Buna göre, ürüne karşılık olarak tüketiciye sağlanan değer, fiyata karşılık tüketiciye maliyet, promosyona karşılık iletişim, dağıtımına karşılık kolaylık kavramları geliştirilmiştir (Özmen, 2013).

Pazarlama kampanyaları, müşteri ihtiyaçlarına ve genel memnuniyetine odaklanır. Bununla birlikte, bir pazarlama kampanyasının başarılı olup olmayacağını belirleyen farklı değişkenler vardır. Bir pazarlama kampanyası yapılırken dikkate alınması gereken belirli değişkenler vardır. Bu değişkenlerden en önemlileri:

- **Ürün:** Bir ihtiyaç ve isteği karşılama özelliğine sahip ve değişime konu olan her şey ürün olarak isimlendirilir. Ürün somut bir nesne, bir hizmet veya fikir olabilir. Pazarlama karmasının ana ögesi üründür. Ürün, kalite, tasarım ve marka gibi önemli

bazı elemanlarından oluşan ve firma tarafından pazara sunulan nesnedir (Yurdakul ve Kiracı, 2008).

- Kampanyaların yapılacağı dağıtım kanalı: Pazarlama kampanyaları için en iyi kanalın, telefon, TV, sosyal medya, radyo, web siteleri vb. gibi, seçimi önemlidir. Ürünler, çok çeşitli yollardan geçer ve üretim yerinden tüketicilere ulaşır. Ürünlerin tüketicilere ulaştığı bu yol dağıtım kanalı olarak isimlendirilir. Toptancı ve perakendeci gibi pek çok aracı dağıtım kanalında bulunmaktadır.
- Fiyat: Bir mal ya da hizmetin değişim veya birim değerinin para ile gösterilmesidir. Fiyatın pazarlamadaki anlamı ise, alıcıların bir mal ya da hizmete sahip olabilmek için ödemeleri gereken para olarak tanımlanır (Eser ve Ilgaz Sümer, 2006).
- Promosyon Stratejisi: Stratejinin uygulanma şekli ile potansiyel müşterilerin nasıl ele alınacağını belirler. İşletmenin, doğru ilginin hedef kitle tarafından kabul edilebilirliğinin artırılması, istenilen tepkiyi yaratması veya başka bir tepkiyi ortaya çıkaracak kadar ikna edici bir şekilde, belirlenmiş olan kitle ile iletişimidir. Daha önce yapılan hataların öğrenilerek, pazarlama kampanyalarını daha etkili hale getirebilmek için önceki pazarlama kampanyaları analiz edilmelidir. Promosyon stratejisi pazarlama kampanyası analizinin son kısmı olmalıdır.

2.1.2. Doğrudan Pazarlama

Doğrudan pazarlama birebir seçilmiş hedef bir alıcı grubuna ulaşılabilen, ölçülebilen, marka bağlılığı oluşturan, somut, iletişim sağlayan, geri dönüşümü kesin bir şekilde raporlanabilen etkinliklerdir. Yani doğrudan pazarlama ile ürün veya hizmetlerin hangi tüketici grubu için uygun olduğu belirlenebilir. Buna örnek olarak, sosyal medya yardımı ile tüketicinin profiline uyan reklamlar bu kişinin sayfasına ulaştırılabilmektedir (Geçer, 2016).

Tüketiciler ile iletişim sağlamak için günümüzde pek çok yöntem kullanılmaktadır. En fazla benimsenen iletişim araçları dergi, televizyon, gazete ve açık hava mecralarıdır. Fakat bunlar kişisel değil kitle iletişim araçları olduğundan tek bir tipe indirgenmiş mesajlar verirler. Doğrudan pazarlamada tüketiciler ile direkt iletişim kurulur. Hedeflenen iki durum vardır, bunlardan biri tüketicilerin iletişim kurulan konu ile ilgili olarak hemen harekete geçmeleri, diğeri ise firma ile uzun süreli ilişki geliştirmeleridir (Zengin, 2010).

Doğrudan pazarlama; müşteriye doğrudan ulaşan postayla satış, postayla reklamdan farklıdır. Aynı zamanda bu kavramları da içeren daha kapsamlı bir uygulamaya sahiptir. Doğrudan pazarlama tek başına; satış, dağıtım veya promosyonu ifade etmez, tüm bunların bileşimidir (İslamoğlu, 2002).

Doğrudan pazarlama yapmanın çeşitli yolları vardır. Bunlardan en önemlileri şunlardır:

- Doğrudan posta
- Tele pazarlama
- E-mail pazarlama
- Sesli posta pazarlama
- Kuponlar
- Televizyon üzerinden doğrudan geri dönüş ile pazarlama
- Doğrudan satış
- Kioks makineleri
- Mobil cihazlar
- Elektronik pazarlama

Doğrudan pazarlama ve klasik pazarlamanın konuları arasındaki farklılıklar aşağıdaki şekilde özetlenebilir (Kobs, 1991).

- Klasik pazarlamada müşteriler demografik özellikler, coğrafi özellikler ve davranışsal özellikler olarak birbirine benzeyenler olarak gruplara ayrılmaya çalışılır. Ulaşılmaya çalışılan müşteriler doğrudan pazarlamanın aksine kitleseldir. Doğrudan pazarlamada müşterilerin işlem hareketleri, satın alma işlemleri, kişisel bilgileri, alışkanlıklar gibi nitelikler değerlendirilerek analizler sonucunda müşterilere birebir ulaşmak hedeflenir.
- Doğrudan pazarlama faaliyetleri için daha az kaynak harcanabilir. Reklam maliyetleri klasik pazarlamaya göre daha düşüktür. Her boyuttaki işletme tarafından kullanılabilir.
- Klasik pazarlamada kullanılan kitle iletişim araçları tek yönlüdür. Doğrudan pazarlamada çift yönlü iletişim sağlanır.
- Klasik pazarlamada kullanılan reklamlar marka imajını yaratma, markayı hatırlatıcı özelliktedir. Doğrudan pazarlamada kullanılan reklamlar, müşteriyi yönlendirerek hızlı karar almasını sağlama amaçlıdır.

Hızlı, güvenilir ve etkin bir şekilde pazara ulaşma ihtiyacını karşılayan en önemli yollarından biri telefonun doğrudan pazarlama aracı olarak kullanılmasıdır. Telefonun bu şekilde kullanılması ile tele pazarlama terimi ortaya çıkmıştır. Tele pazarlama, iletişim teknolojisi ve tele pazarlama konusunda eğitilmiş personel yardımıyla, hedef tüketicilere yönelik etkili bir iletişim sistemidir (Varinli ,2010). Kişisel satış, müşterinin araması veya müşterinin aranması ile gerçekleşir (Geçer, 2016).

Firmalar, müşterilerden gelen aramalar için gelen veya içe doğru tele pazarlama, müşterilere yaptıkları aramalara ise giden veya dışa doğru tele pazarlama kavramlarını kullanırlar (Kotler ve diğ., 2012).

Tele pazarlama zamanla müşterileri rahatsız edecek seviyelere ulaşmıştır. Tele pazarlama tarihi olarak en temel doğrudan pazarlama araçlarından biri olarak kullanılmasına rağmen; izinsiz kullanıma açık doğası nedeniyle, Amerika Federal Ticaret Komisyonu tarafından 2003 yılında “Ulusal Arama Engelleme Kayıt Merkezi”nin kurulmasına sebep olmuştur. Yaklaşık 191 milyon kişi, evlerinin tele pazarlamacılar tarafından aranmasını istemediğinden, 2009 yılı itibarıyla bu merkeze kayıt olmuştur (Kotler P ve diğ., 2012).

Şirketlerin tele pazarlamaya yönelik yaptığı aramalar üçe ayrılabilir. Bunlar soğuk arama, ılık arama ve sıcak aramadır. Soğuk aramalarda firmalar rasgele arama gerçekleştirir. Arama öncesinde herhangi bir strateji geliştirilmez. Ilık aramalarda hedef müşteri kitlesine yönelik aramalar yapılır. Seçilen müşterilerin pazarlanmak istenen ürünle ilişkileri olduğu varsayılarak aranır. Sıcak aramalarda ise arama yapılan müşteriler daha önce ürünle etkileşime geçmişlerdir. Daha önce ürün ile ilgilenmiş, ürünü veya hizmeti satın almış ya da kampanyaya olan ilgisini belirtmiştir.

Telefona dayalı teknolojiler sayesinde, hizmetin düşük maliyetli olarak ulaştırılmasını sağlama ve yüksek getiri şansı nedeniyle çağrı merkezleri global olarak büyüme göstermiştir. Çok sayıda işletme için tele pazarlama, pazarlama stratejileri için en önemli yoldur. Örnek olarak, 1890 yılında İngiliz markası olarak kurulan Raleigh Bisikletleri, tele pazarlama sayesinde bireysel satış maliyetlerinde %50 azalma, satışlarda %34 oranında artış elde etmiştir (Gelibolu ve Özsoy, 2013).

Çağrı merkezleri hem satışı destekleyen hem de pazarlama stratejileri için tüketici bilgilerinin toplandığı yerler olarak önemlidir. Çağrı merkezlerinin tüketicilere sağladığı kolaylıklar şunlardır:

- Şubeye gidilmediği ve kuyrukta bekleme olmadığı için zaman kaybı söz konusu değildir.
- Tüketici ülkenin hatta dünyanın herhangi bir yerinde işlemlerini yapabilir.
- Tüketici hatta bekletilmediğinden memnuniyeti artar.
- Tüketici ihtiyacı olduğu anda hizmete ulaşabilir. Bankaya sadece mesai saatleri içinde ulaşıldığı için bu durum bir avantajdır.
- Banka çağrı merkezi ile gerçekleştirilen işlemlerin maliyeti daha düşüktür.
- Müşteri işlemlerini kişiselleştirebilir. Örneğin istediği dilde hizmet alabilir veya sık kullandığı işlemler için kendine menü oluşturabilir.
- Çağrı merkezleri tüketici bilgilerini depoladığından müşteriler her seferinde özlük bilgilerini girmek zorunda kalmazlar; sadece banka hesap numarası veya müşteri numarası işlemleri yapmak için yeterli olur. Böylece işletme ile müşteriler arasında uzun süreli ilişkiler sağlanır (Sarıyer, 2007).

Tele pazarlama kampanyalarında pazarlama faaliyeti yapılmak istenen müşteri bilgileri son derece önemlidir. Müşterilerin kişisel özellikleri, demografik bilgileri, coğrafik bilgileri, alışkanlıkları, hizmet veya ürün ile etkileşim noktaları, hizmet veya ürüne verdikleri tepkiler, geçmiş işlemleri gibi bilgiler tele pazarlama stratejilerinde kullanmak üzere etkilidir. Müşteri bilgileri dışında ilgili hizmetin veya ürünün pazardaki genel durumu, geçmiş ekonomik veriler, finansal göstergeler gibi pazarın ekosisteminde olan veri kaynakları da doğrudan pazarlamada dikkate alınması gereken bilgilerdir. Doğrudan pazarlamanın etkin şekilde yapılması veri tabanlı pazarlama ile sağlanır. Veri tabanlı pazarlama müşteri bilgileri ve diğer etmenlerin bir arada değerlendirilmesini sağlar.

İçinde bulunduğumuz dönemde yapılan sonuç odaklı pazarlama faaliyetleri, pazarlama sürecinin yanında müşterilerin tepkilerine de önem vermektedir. Veri tabanlı pazarlama faaliyetleri ile geri dönüşü de değerlendirerek verimliliği artırıcı etkisini dikkate almak gerekir. Veri tabanlı pazarlama müşteri yönetimini sağlayan bir yöntemdir. Yeni müşterilere ulaşmak,

müşterileri elde tutmak, müşteri satışlarını arttırmak müşteri ilişkileri yönetiminin esaslarıdır. Müşteri ilişkileri yönetimi için en çok kullanılan yöntem veri tabanlı pazarlamadır.

Tez kapsamında UCI makine öğrenimi web kütüphanesinde yer alan bankacılık veri seti kullanılmıştır. Veri setinde yer alan müşteri demografik bilgileri, müşterinin bankacılık işlemleri, daha önce yapılan kampanya bilgileri gibi bilgilerinden faydalanılarak sonraki tele pazarlama kampanyaları için bankaya yüksek kar sağlayacak en doğru müşteri sınıflandırmasını sağlayacak modeli geliştirmek amacıyla veri madenciliği tekniklerinden yararlanılmıştır. Veri seti ön işlemleri tamamlandıktan sonra, farklı sınıflandırma ve regresyon modelleri ile ulaşılan sonuçlar değerlendirilmiştir. Daha iyi sınıflandırma sonuçları elde etmek ve Aşırı Öğrenme makinesi sınıflandırıcısının diğer sınıflandırıcılara göre performansını incelemek için çalışmalar yapılmıştır. Öncelikle parametre ayarlama teknikleri ile en iyi sınıflandırıcı parametrelerinin tespiti incelenmiş sonrasında daha iyi sınıflandırma sonuçları elde etmek için genetik algoritma tekniğinden yararlanılmıştır. Son olarak topluluk algoritmalarından faydalanarak en iyi sınıflandırma performansını sağlayacak yöntem belirlenmiştir.

2.2. VERİ MADENCİLİĞİ

Veri madenciliği büyük miktardaki veriyi analiz etme ve tahminleme yapmaya yardımcı olacak ilişkiler kurma, veriler arasındaki kuralların bilgisayar programlarıyla aranması ve analiz edilmesi tekniğidir (Kalikov, 2006). Veri Madenciliği İngilizce Türkçe Ansiklopedik Bilişim Sözlüğü'nde, "Doğal Dillerin Semantik yapısına dayanarak elektronik metin belgeleri içinde saklı kalmış ilintileri, örüntüleri, stratejik bilgileri, modelleri vb. bulup ortaya çıkarmayı amaçlayan araştırma tekniği" (Sankur, 2004) olarak tanımlanmaktadır.

Veri madenciliği için veriden bilgi keşfetme süreci de denebilir. Veri madenciliğinde otomatik ve yarı otomatik metotlar kullanılarak büyük veri içerisinde değerli bilgi edinme amaçlanır. Veriden modeller geliştirmek, ilişkiler incelenerek ilginç özellikleri belirlemek veya tekrarlanan veri seti kümelerini yakalamak ve bunları yapabilmek için istatistik, yapay zekâ, bilgisayar bilimi gibi çeşitli bilim dallarında geliştirilen algoritmalar kullanmak veri madenciliğidir. Veri madenciliği, veri seti içerisindeki kullanışlı bilgilerin anlamlı ve kullanışlı ilişkilerin bulunabilmesi için veri tabanı enformasyonunu analiz edebilen bütün teknolojileri bir araya getirmektedir (Monino ve Sedkaoui, 2016).

Akpınar'a (2000) göre veri madenciliği büyük veri içerisinde istatistiksel ve matematiksel yöntemlerle veri desenini anlamayı sağlayan teknolojilerin kullanılarak anlamlı yeni ilişkiler, yeni veri desenleri, yeni kurallar keşfetme sürecidir.

Prytherc (2005), "İlişkileri ve modelleri analiz etme amacı ile gizli olan bilgileri keşfetmek için ayrıntılı teknikler kullanarak, bilginin veri tabanından ve veri kümelerinden çıkarılması işlemidir" şeklinde tanımlanmaktadır.

Veri madenciliği büyük veri setlerini yorumlayarak tahminleme yapabilmemizi sağlayabilecek bağlantıların bilgisayar programları ile aranmasıdır. (Doğan ve Türkoğlu., 2007)

Veri madenciliği tanımlarının büyük kısmı birbirine benzerdir. Genel olarak değerli bilgiyi ortaya çıkarma amacıyla yapılan analiz çalışmaları ve bu çalışmaların yapılabilmesi için matematiksel modellerden yararlanma sürecidir.

Veri madenciliği birçok disiplini kapsayan bir kavramdır. Çok disiplinli yapısının aşağıdaki alanlardan etkilendiği söylenebilir. Bu alanlar tamamen birbirinden bağımsız düşünülemez. Veri madenciliğinde bu alanlardaki çalışma ve yöntemlerden faydalanılmaktadır (Hand ve diğ., 2001).

- Makine Öğrenmesi
- Örüntü Tanıma
- Görselleştirme
- Algoritmalar
- Yüksek Performanslı Hesaplama
- Uygulamalar
- Bilgi Getirimi
- Veri Ambarı
- Veri Tabanı Sistemleri
- İstatistik

Veri madenciliğinin birçok uygulama alanı vardır. En önemli uygulama alanları aşağıda belirtilmiştir.

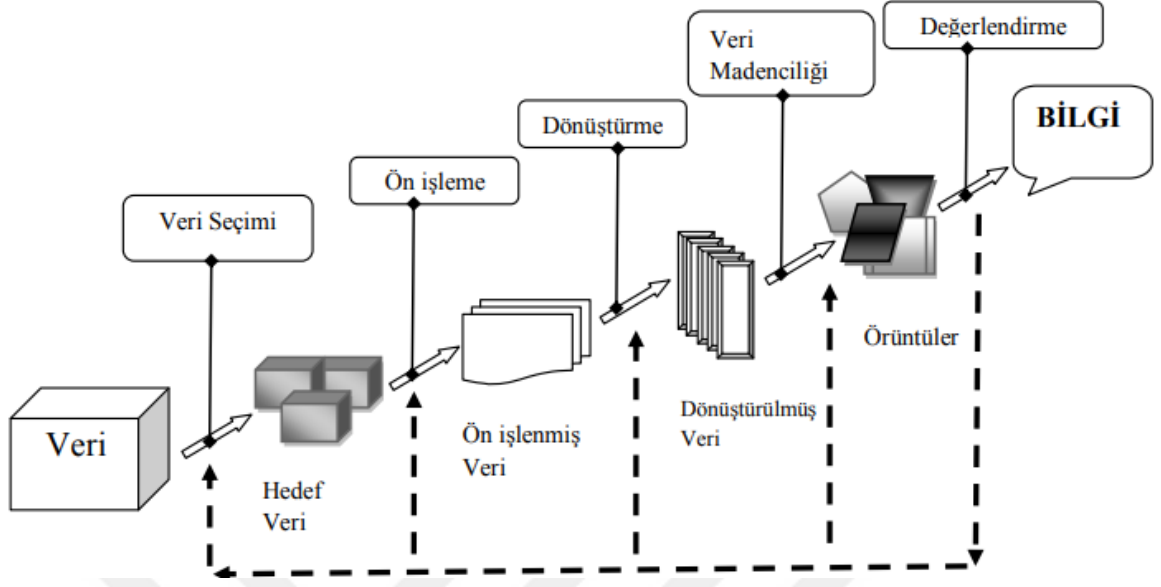
- Pazarlama
- Finans
- Sigortacılık
- Perakendecilik
- Sağlık
- Mühendislik
- Arama Motorları
- Mobil Uygulamalar

2.2.2. Veri Madenciliği Süreci

Veri madenciliği bir süreci ifade eder. Veri üzerinde çalışarak değerli bilgiye ulaşmanın yanında, bilgi keşfi sürecinde ilişki keşifleri yaparak veriyi ayırmak ve sonraki adımlara hazır hale getirmek sürecin parçasıdır. İncelemeye konu olan veri setindeki verilerin özelliklerinin bilinmemesi algoritmaların çalışmasını etkileyecektir. Veri madenciliği sürecine başlamadan önce veri analizinin detaylı şekilde yapılması çok önemlidir.

Fayyad ve diğ. (1996) bilgi keşif sürecini Şekil 2.1'deki gibi göstermiştir. Bilgiye ulaşmak belirtilen beş adım tamamlanmalıdır. Veri madenciliği bu adımlardan biridir.

- Problemin Tanımlanması,
- Veri İşleme Süreci,
- Modelin Kurulması,
- Modelin Değerlendirilmesi ve Seçilmesi
- Modelin Uygulanması



Şekil 2.1: Bilgi keşfi sürecinde veri madenciliği (Fayyad ve diğ. 1996).

2.2.2.1. Problemin tanımlanması

Sürecin en önemli adımıdır. Verinin seçilmesidir. Problemin tanımlanması proje amacının belirlendiği adımdır. Bu adımda elde edilen sonuçların nasıl ölçüleceği belirlenecektir. Bilgi keşfi ile ulaşılmak istenen hedefin belirlendiği adımdır.

2.2.2.2. Verinin anlaşılması

Verilerin anlaşılması verilerin toplanmasıyla başlar. Proje amacına göre iç ve dış veriler toplanır. İç veriler ürüne dair bilgilerin bulunduğu kurumun veri tabanında yer alan veridir. Dış veriler ise projenin etki alanına göre dışarıdan sağlanan nüfus, araştırma firmalarının verileri, devlet veri tabanları gibi kaynaklardır.

Verinin problemle ilişkilendirilerek yorumlanması ve probleme farklı bakış açıları kazanarak yaklaşılması son derece önemlidir.

2.2.2.3. Veri işleme süreci

Verinin temizliğinin yapıp eksik verinin ele alındığı, alakasız verileri ile sınır değerlerin veri setinden çıkarıldığı aşamadır. Modelin doğruluk ölçümlerine göre sık sık bu adıma geri dönüp düzenleme yapmak gerekebilir. Toplam analiz zamanı içerisinde, verilerin hazırlanması ve modelin kurulması, büyük zaman almaktadır. Veri işleme süreci verilerin hazırlanması, veri temizleme, veri bütünleştirme, veri seçimi, veri dönüştürme aşamalarından oluşmaktadır.

1- Veri temizleme: Eksik verilerin tamamlandığı, tutarsız ve kirli verinin çıkarıldığı, uç değerlerin tespit edildiği, verinin kalitesini artırma aşamasıdır (Han ve diğ., 2012). Eksik verileri çeşitli yöntemlerle doldurmak gerekir. Kirli verilerin, gürültüye neden olanların tespit edilmesi ve gürültünün azaltılması hedeflenmelidir.

2- Veri bütünleştirme: Analiz birden çok veri kaynağı kullanılarak yapılabilir. Bu kaynakların bütünleştirilmesi ve ilişkilere göre verinin birleştirildiği aşamadır. Veri bütünleştirmede kullanılan kaynaklarda aynı verinin farklı şekilde isimlendirme ya da gösterimleri olması sorun oluşturabilir. Veri bütünleştirme öncesinde ortak ifadeler tüm kaynaklarda aynı hale getirilmelidir.

3- Veri seçimi: Veri tabanından analizle alakalı olan niteliklerin seçildiği süreçtir. Çözümlemelerin belli boyutlar için yapılması, bazı niteliklerin veri setinden çıkarılması veya büyük veri toplulukları yerine onu temsil eden daha küçük veri kümelerinin oluşturulmasıyla veri seçimine gidilebilir. Veri seçimi yapılmadan yapılan veri madenciliği çalışmaları maliyetli olabilir. Sonucu etkilemeyecek niteliklerin çıkarılması performansı artıracaktır.

Nitelik seçimi, veri analizi makine öğrenimi aşamasında sıklıkla kullanılan bir yöntemdir. Esas özelliklerin seçimidir. Nitelik seçiminde nitelik sayısı belirlenen değerlendirme ölçütüne göre en uygun şekilde azaltılır. Nitelik seçimi 1970'li yıllardan itibaren veri analizi süreci için bir araştırma ve geliştirme alanı olmuştur (Blum,1997).

Veri seti içerisinde yer alan niteliklerin bir kısmı sonucu etkilemeyebilir, konuyla ilişkisi olmayabilir. Ayrıca birçok algoritmanın performans açısından çok sayıda nitelik üzerinde çalışması verimsiz olabilir. Bu nedenle, veri madenciliği algoritmaları uygulanmadan önce, özellik seçimi tekniklerinin uygulanması faydalı olabilir. Özellik seçiminin amacı, mümkün olduğunca küçük bir özellik alt kümesini belirleyerek model performansının aşırı öğrenmesini engellemek ve daha hızlı, maliyet açısından daha uygun modeller sunmaktır. Nitelik seçimi için sınıflandırma modelleri kullanılarak amaç fonksiyonunu etkileyen nitelikler sınıflandırılır.

Azaltılmış özellik alt kümesine madencilik görevlerinin uygulanmasıyla, orijinal yüksek boyutlu veri kümesiyle aynı sonuç elde edilir. Özellik seçimi, depolama gereksinimlerini azaltmak, aşırı uyumu önlemek, veri görselleştirmesini kolaylaştırmak, madencilik

algoritmalarının yürütülmesini hızlandırmak ve eğitim sürelerini kısaltmak gibi avantajlar sunar (Das, 2001).

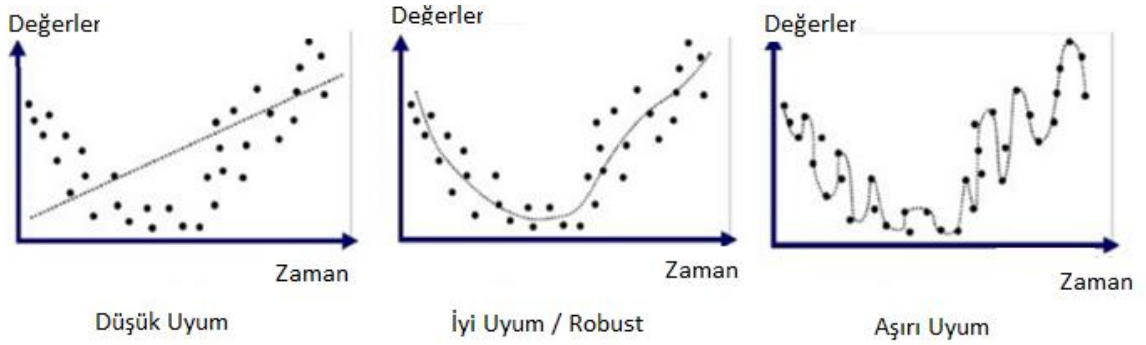
4- Veri dönüştürme: Veriler arasındaki veri tipi farklılıkları hatalı analize yol açabilir. Veri dönüştürme, analize uygun hale gelmesi için verinin standartlaştırılması ve normalleştirilmesi, veri tiplerinin birbirleri arasında dönüştürülmesidir. Bu dönüştürme sonucunda 1-5, 6-10, 11-15 gibi aralık etiketleri veya düşük, normal, yüksek gibi etiketlemeler yapılmaktadır (Han ve diğ., 2012).

2.2.2.4. Modelin kurulması

Veri madenciliği sürecinin bu aşamasında önceki aşamalarda hazırlanan veri kullanılarak amaca uygun veri madenciliği yöntemleri uygulanır. Veri madenciliği yöntemleri farklı modellerle sağlanır. Veri madenciliği için en uygun modelin bulunabilmesi, mümkün olan en çok sayıda model kurularak sağlanır. Veri hazırlama fazı ve model kurma fazı en iyi modele ulaşılan kadar tekrarlanmalıdır. Model kuruluş çalışmalarına başlamadan önce en uygun modele karar verilmesi zordur. Bahsedilecek model değerlendirme ve seçim aşamaları ile en doğru modele karar verilebilir.

2.2.2.5. Modelin değerlendirilmesi ve seçilmesi

Model eğitim veri seti üzerinde çok kompleks kurgulanırsa aşırı uyum durumu oluşur. Bu model çalıştığında yüksek sınıflandırma oranı verir. Fakat daha önce karşılaşılmayan yeni veri setlerinde başarısız olma olasılığı yüksektir. Tam tersi şekilde model aşırı basit kurgulandığında düşük uyum durumu oluşur. Düşük uyum, iyi uyum ve aşırı uyum durumlarında görülen modelin görüntüsü Şekil 2.2’de gösterilmiştir.



Şekil 2.2: Model uyumu.

Modelin aşırı öğrenmesini (ezberlemesini) veya tersi durumunu (genellemesini) engellemek için model değerlendirme yöntemlerine başvurulur.

Model performans değerlendirme yöntemleri:

Dışarda Tutma

Bu yöntemde ilk olarak test ve eğitim verisi ayrılır. Verinin seçimi rasgele gerçekleşir. Bu sebeple farklı sınıflara ait veriler test ile eğitim verisinde tutarsız oranlarda dağılabilir. Test setinde kullanılan veri, eğitim setinin dışındaki verilerden oluşur. Aynı sınıftan çok fazla verinin eğitim veya test seti içerisinde olması engellenmelidir. Bu durumu engellemek için her sınıfın eğitim ve test veri setlerinde eşit oranda dağıldığından emin olmak amacıyla tabakalı örnekleme yöntemi tercih edilebilir. Bu yöntemde iki dezavantaj görülür.

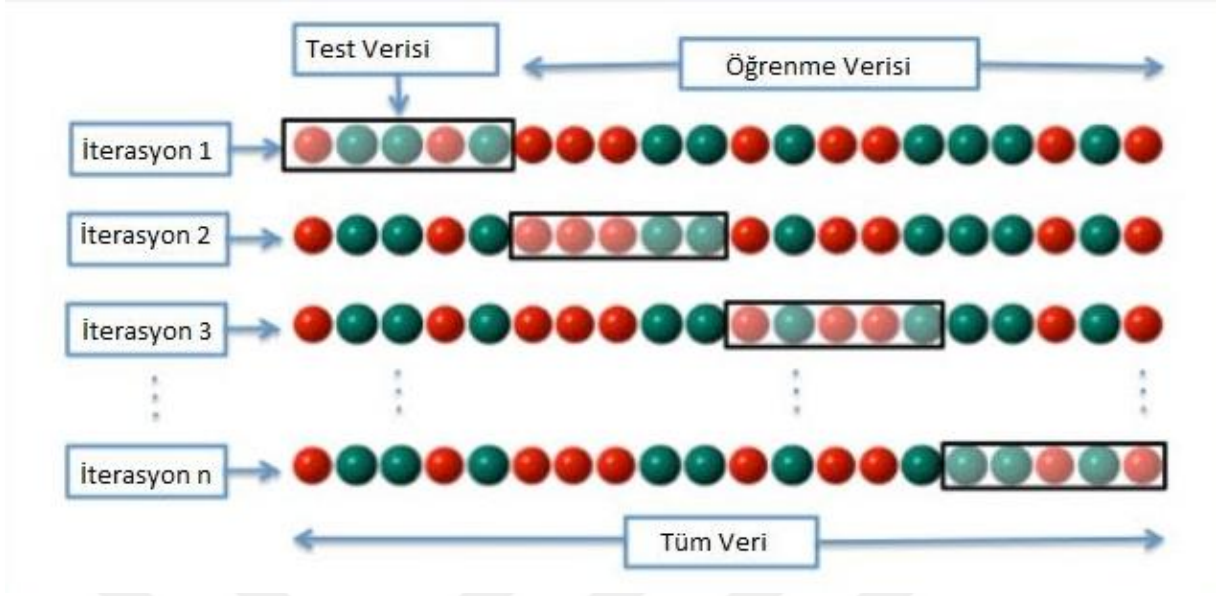
- Veri setinde bulunan örnekleme sayısının az olması durumunda test ile eğitim verisi ayrımının yeterince uygun yapılamaması,
- Veri setinin eğitim ve test olarak bir defa ayrılması ve verilerin tamamının modelde kullanılamaması test veri setinde yer alan modelin gelişimine katkı sağlayamaz.

Tekrarlı Dışarda Tutma

Dışarda tutma yönteminin birkaç defa tekrarlanmasıdır. Bu yöntemde seçim işlemi her ne kadar rasgele yapılsa da farklı test veri setleri üst üste binebileceğinden elverişli görülmemektedir. Çok yaygın bir şekilde kullanılmamaktadır.

Çapraz Geçerleme

Bu yöntemde veri kümesi eşit boyutta k adet veri grubuna ayrılmaktadır. Her zaman bir grup test için ayrılırken geriye kalan gruplarla eğitilecek veri için model kurulur ve test için ayrılan verilerle doğruluk oranı hesaplanır. Bu işlem k defa tekrarlanarak ölçümler yapılır. Bunun sonucunda her bir veri grubu eğitim ve test için aynı sayıda kullanılmış olur (Han ve Kamber, 2006). Çapraz geçerleme yönteminde model doğruluğu her bir gruba ayrılma işlemi sonucunda elde edilen doğruluk değerlerinin ortalaması ile hesaplanır. Şekil 2.3'de çapraz geçerleme gösterimi verilmiştir.



Şekil 2.3: Çapraz geçerleme gösterimi.

Model performans değerlendirme ölçütleri:

Model kurulduktan sonra eğitim ve test verileriyle yapılan analizlerin sonuçlarının değerlendirilmesiyle model performansı ölçülür. Model eğitimi için eğitim verisi kullanılarak sınıflandırmak ve performans değerlendirme için öğrenilmiş modelin doğruluğunu tahmin etmek, öğrenme algoritmasının çalıştırılan verilerle aşırı öğrenmesine sebep olur ve fazla iyimser tahminlemelere yol açar. Bunun yerine, sınıflandırmayı test verisiyle ölçmek daha doğru olur (Han ve diğ., 2012).

Model performansı değerlendirilirken belirli ölçütlere göre yapılan karşılaştırma ile en iyi performansı veren modelin seçilmesi sağlanmalıdır. Bu değerlerin hesaplanabilmesi için doğru pozitif, yanlış pozitif, doğru negatif, yanlış negatif değerlerinin kullanılması gerekmektedir (Balaban ve Kartal, 2015).

Birçok model değerlendirme ölçütünde kullanılan temel yapı blokları olarak, bilinmesi gereken dört terim vardır. Bu kavramların iyi anlaşılması modele dair önlemlerin alınmasını ve değerlendirilmesini sağlar (Han ve diğ., 2012).

- Doğru Pozitif (True Positive): Pozitif sınıfına ait olup, modelde pozitif olarak tahminlenen örneklerin sayısıdır. Tahmin doğru yapılmıştır. Veri pozitif sınıfa aittir.
- Doğru Negatif (True Negative): Negatif sınıfına ait olup, modelde negatif olarak tahminlenen örneklerin sayısıdır. Tahmin doğru yapılmıştır. Veri negatif sınıfa aittir.

- Yanlış Pozitif (False Positive): Negatif sınıfına ait olup, modelde pozitif olarak tahminlenen örneklerin sayısıdır. Tahmin edilen veriler gerçekte negatif sınıftadır ve sınıflandırma yanlış yapılmıştır.
- Yanlış Negatif (False Negative): Pozitif sınıfına ait olup, modelde negatif olarak tahminlenen örneklerin sayısıdır. Tahmin edilen veriler gerçekte pozitif sınıftadır ve sınıflandırma yanlış yapılmıştır.

Karmaşıklık matrisi, sınıflandırıcıların farklı sınıflardaki örneklemeleri ne kadar iyi tahminleyebildiğini analiz etmek için kullanılır. Doğru Pozitif (TP) ve Doğru Negatif (TN), sınıflandırıcının örneklemeleri doğru tahminleyebildiğini gösterir. Yanlış Pozitif (FP) ve Yanlış Negatif (FN) sınıflandırıcının örneklemeleri yanlış tahminlediğini gösterir. Tablo 2.1’de model değerlendirme ölçütleri gösterilmiştir.

Tablo 2.1: Model değerlendirme ölçütleri.

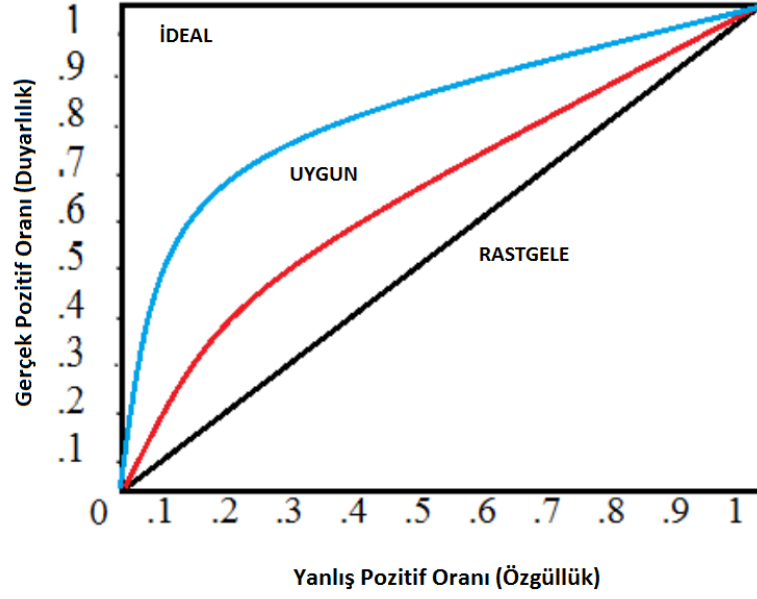
Karmaşıklık Matrisi		Tahmin		Toplam
		Pozitif	Negatif	
Gerçek	Pozitif	TP	TN	P
	Negatif	FP	FN	N
	Toplam	P'	N'	P + N

Hesaplanan doğru pozitif, doğru negatif, yanlış pozitif ve yanlış negatif değerleri kullanılarak daha detaylı model değerlendirme ölçütleri hesaplanır. Bu değerler 1 ile 0 arasında hesaplanıp 1 değerine yakın sonuçlar en iyi tahmin sonuçlarını gösterir. Ölçütler ve ölçütlerin İngilizce karşılıkları, açıklamaları ve formülleri Tablo 2.2’de gösterilmiştir.

Tablo 2.2: Model değerlendirme ölçütleri.

Ölçüt	Açıklama	Formül
Doğruluk (<i>accuracy, recognition rate</i>)	Doğru tahmin edilen örneklerin, tüm örneklerin sayısına oranıdır.	$\frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{P + N}$
Hata oranı (<i>error rate, missclassification rate</i>)	Doğruluk değeri ile toplamı bire eşittir.	$\frac{FP + FN}{TP + FP + TN + FN} = \frac{FP + FN}{P + N}$
Duyarlılık (<i>sensitivity, recall, true positive rate-TPR</i>)	Doğru tahmin edilen pozitif sınıftaki örneklerin gerçekte pozitif sınıftaki örneklerin sayısına oranıdır.	$\frac{TP}{TP + FN} = \frac{TP}{P}$
Belirleyicilik (<i>specificity, true negative rate-TNR</i>)	Doğru tahmin edilen negatif sınıftaki örneklerin, gerçekte negatif sınıftaki örneklerin sayısına oranıdır.	$\frac{TN}{TN + FP} = \frac{TN}{N}$
Pozitif öngörü değeri/Kesinlik (<i>positive predictive value, precision</i>)	Doğru tahmin edilen pozitif sınıftaki örneklerin, pozitif sınıfta tahmin edilen örneklerin sayısına oranıdır.	$\frac{TP}{TP + FP} = \frac{TP}{P'}$
Negatif öngörü değeri (<i>negative predictive value</i>)	Doğru tahmin edilen negatif sınıftaki örneklerin, negatif sınıfta tahmin edilen örneklerin sayısına oranıdır.	$\frac{TN}{TN + FN} = \frac{TN}{N'}$
Yanlış pozitif oranı (<i>false positive rate-FPR</i>)	Yanlış tahmin edilen negatif sınıftaki örneklerin, negatif sınıftaki örneklerin sayısına oranıdır.	$\frac{FP}{FP + FN} = \frac{FP}{N}$ $1 - \text{belirleyicilik}$
Yanlış negatif oranı (<i>false negative rate-FNR</i>)	Yanlış tahmin edilen pozitif sınıftaki örneklerin, pozitif sınıftaki örneklerin sayısına oranıdır.	$\frac{FN}{FN + TP} = \frac{FN}{P}$ $1 - TPR$
F-ölçütü (<i>F-score</i>)	Duyarlılık ve kesinliğin birlikte değerlendirildiği ölçüttür.	$\frac{2 \times \text{kesinlik} \times \text{duyarlılık}}{\text{kesinlik} + \text{duyarlılık}}$

ROC eğrisi (Receiver Operating Curve), modellerin doğrulunu ölçmek için sıkça kullanılır. Farklı sınıflandırıcıları karşılaştırmak için kullanılmaktadır. ROC eğrisi üzerindeki çizginin altında kalan her bir alan ilgili sınıflandırıcı performansını ifade eder. ROC eğrisi, y ekseninde hassaslık ve x ekseninde ise özgüllük oranlarını kıyaslayarak birimleri sınıflarına ayıran en uygun kesişim noktasını belirler. Sınıflandırma doğruluğu bu eğri altında kalan alana bağlıdır. Şekil 2.4'te ROC eğrisi örneği gösterilmiştir.



Şekil 2.4: ROC eğrisi örneği.

2.2.2.6. Modelin uygulanması ve yorumlanması

Üretilen bilgiler problemin amacı doğrultusunda kullanılır. Önceki her bir aşamanın sonucu değerlendirilir, yorumlanır ve elde edilen bulgularla birlikte seçilen stratejiler veriye uygulanır. Karar verilen model uygulama olarak kullanılabileceği gibi farklı uygulamaların alt uygulaması olarak ta kullanılabilir. Bu aşamada çalışmanın raporunun hazırlanması ve projenin en baştan değerlendirilmesi de yapılmalıdır. Sonuçların sonraki çalışmalarda kaynak olarak kullanılabilmesi ve farklı modellerle yeniden değerlendirilmesine olanak sağlayacak şekilde bilgi içermelidir.

2.2.3. Veri Madenciliği Yöntemleri

Veri madenciliği denetimli ve denetimsiz öğrenme yöntemleri olarak ikiye ayrılır. Kesin bir hedefe ulaşmak için tanımlanan yöntemler için denetimli öğrenme, hedeflerde belirsizlik varsa kesin bir hedef yoksa denetimsiz öğrenme söz konusudur.

Denetimli öğrenme, veri üzerinden her bir niteliğin birbiri ile ilişkisi analiz edilerek örneklerin sınıflarının bulunması ve bu sınıflara ulaşan kuralların belirlenmesidir. Denetimli öğrenmeye yönlendirilmiş yöntem de denir.

Denetimsiz öğrenmede ise sonuç net değildir. Bir amaç belirlenmemiştir. Denetimli yöntemlerde olduğu gibi veriler incelenerek sınıfları tespit edilmeye çalışılır. Denetimliden

farklı olarak hedef niteliği olmadığından kümeleme analizinde olduğu gibi veriler arası ilişkiler kullanılarak sınıfların tanımlanması amaçlanır. Veri madenciliğinde bu iki yöntem temel alınarak tahminleyici ve tanımlayıcı modeller oluşturulmuştur (Canan, 2014).

Tanımlayıcı yöntemler, bir veri setindeki ortak özelliklerin ortaya çıkarılması ve anlaşılmasını sağlayan yöntemlerdir. Tahminleyici yöntemler ise veri seti üzerinden tahmin yapılması için kullanılan yöntemlerdir.

Tahminlemeye dayalı olan yöntemler, eldeki veriyi kullanarak belirlenen veri deseni üzerinden geleceğe yönelik tahminler yapmayı hedefler. Sınıflandırma modellerinde amaç en baştan bellidir, veri seti analiz edilerek sınıflara ayrılır ve gelecek ile ilgili bir amaç tahminlenir. Sınıflandırma ve Regresyon tahminlemeye dayalı yöntemlerdendir.

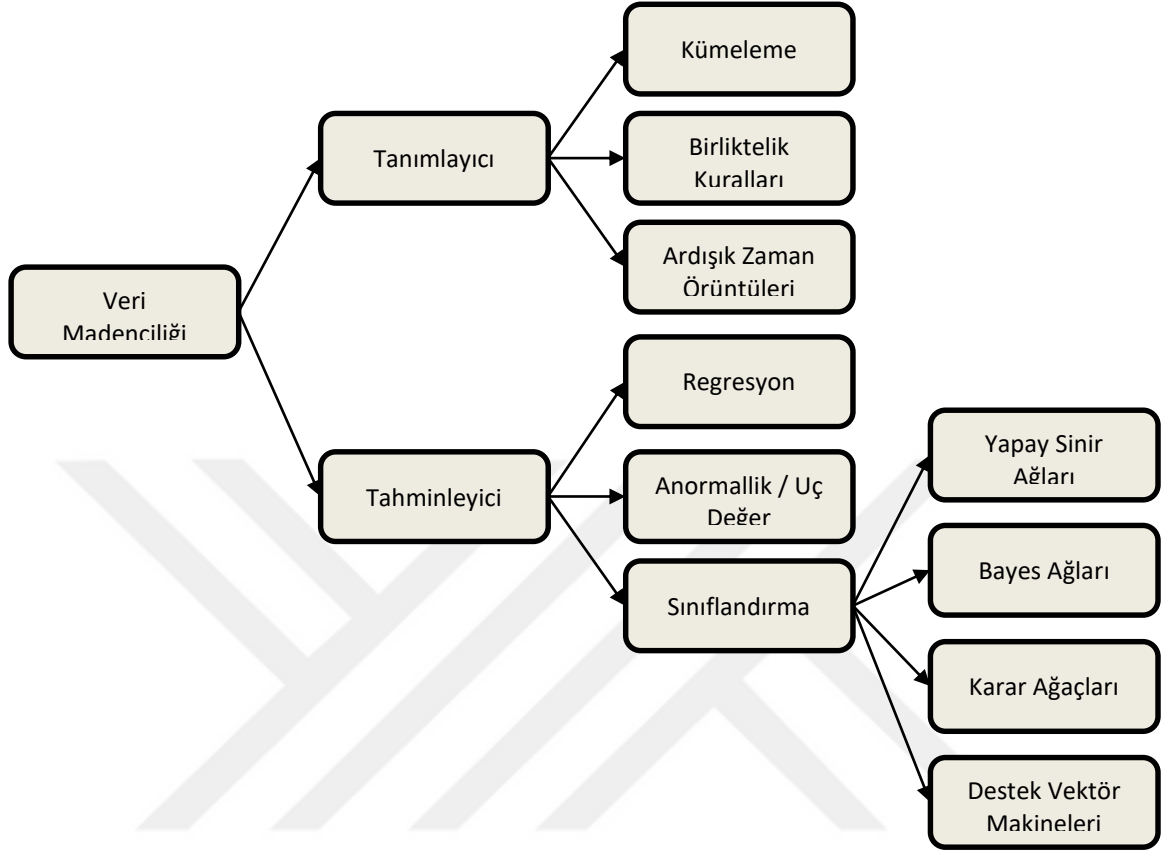
Tanımlayıcı modellerde ise veri seti içerisindeki ilişki tanımlamaları yaparak amaç doğrultusunda aksiyon alma için sonuçlar çıkarılmaktadır. Kümeleme modellerinde ise veri seti analiz edildiğinde doğal olarak verilerin kümелendiği görülür. Ardışık zamanlı örüntülerde zaman içerisinde oluşan ilişkiler dikkate alınır. Birliktelik kurallarında ise veri setindeki farklılıklar arasında ilişkiler belirlenir. Veri setindeki ürün bilgileri, davranışları, durum bilgileri gibi bilgiler arasında ortak bir eğilim belirlenir.

Kümeleme, Birliktelik Kuralları ve Sıralı Diziler tanımlayıcı modeller arasında yer almaktadır. Özkan (2008) veri madenciliği yöntemleri aşağıdaki başlıklar altında toplamaktadır:

- Sınıflandırma
- Regresyon
- Kümeleme
- Birliktelik kuralları
- Ardışık zamanlı örüntüler

Maimon ve Rokach (2010), veri madenciliği yöntemlerini daha genel olarak tanımlayıcı ve tahminleyici modeller olmak üzere iki başlık altında incelemiştir. Tanımlayıcı ve tahminleyici veri madenciliği yöntemleri şekil 2.5’de gösterilmiştir.

Şekil 2.5: Tanımlayıcı ve tahminleyici veri madenciliği yöntemleri.



2.2.3.1. Sınıflandırma

Sınıflandırma, her bir kaydı veri setinde yer alan çeşitli gruplara dağıtma anlamına gelir. Sınıflandırmada yeni örneklemin özellikleri incelenerek, mevcut gruplardan hangisine ait olduğunu belirleme amacı vardır. Sınıflandırma, hangi gruba ait olduğu bilinmeyen verilerin sınıflarının belirlenmesi işlemidir (Özkan, 2008).

Sıklıkla karşılaşılan sınıflandırma yöntemlerinden bazıları şunlardır:

- Lojistik regresyon
- Karar ağaçları
- Yapay sinir ağları
- Sade Bayes sınıflandırıcı
- Destek vektör makineleri
- k-en yakın komşuluk

2.2.3.2. Regresyon

Regresyon analizi iki veya daha fazla deęişken arasındaki ilişkinin göstergesidir. Regresyonda amaç girdiler ile çıktıları arasındaki ilişkiyi kurmada en doğru tahmini yapabilecek modeli oluşturmaktır. Regresyon analizinin amacı girdi ve çıktıları arasındaki ilişkiyi en doğru şekilde tahminlemeyi sağlayacak regresyon modeli veya tahmin denklemi kurmaktır. Sınıflandırma ve regresyon arasındaki fark ise sınıflandırmada çıktı deęişkeni kategorik iken regresyonda sayısalıdır.

Regresyon analizinde girdiler bağımsız deęişkenler, çıktıları ise bağımlı deęişkenler olarak isimlendirilir. Regresyon analizi modelleri, bir bağımlı deęişken ile bir veya birden çok bağımsız deęişkendenden oluşmaktadır. Deęişkenler nümerik olmalı, sayılabilir ve dolayısıyla ölçülebilir olmalıdır, Regresyon analizi fiyat karşılaştırmaları, saęlık sektöründe ilaç reaksiyonları, finansal tahminler, talep tahminleri gibi alanlarda kullanılmaktadır (Argüden ve Erşahin, 2008).

2.2.3.3. Kümeleme

Kümeleme, örnekler arasındaki yakınlık veya benzerliğin uygun niteliğine göre, örnek kümesinin belirli parça veya kümelerle gruplanması süreci olarak tanımlanır (Ye, 2003). Amaç aynı grup içerisindeki örneklerin birbirleriyle benzerlikleri ve dięer gruplardan ne kadar farklı olduğunu bulmaktır. Sınıflandırmadan farklı olarak kümelemede örneklemin hangi kümeye ait olduęu önceden bilinmez. Kümeleri belirlemek için örneklerin birbirine olan benzerlikleri deęerlendirilmelidir.

Kümeler belirlendikten sonra veri kümesi grup olarak ayrılmalıdır. Böylece bu kümeler daha sonra yeni verileri sınıflandırmak için kullanılabilir (Kulluk, 2009).

Kümelemede, benzerlikleri bulmak için en çok kullanılan yöntemler Manhattan ve Öklit uzaklık fonksiyonlarıdır. Uzaklık fonksiyonlarının iki örnek üzerinde uygulanması sonrasında çıkan deęer düşükse bu örnekler birbirine yakın kümelerle dahil olmalı, çıkan deęer yüksekse örnekler birbirine uzak kümelerle dağılmalıdır.

Bazı kümeleme yöntemler aşağıda yer almaktadır.

- Bölmeli Kümeleme
- Hiyerarşik Kümeleme
- Özel Olmayan Kümeleme
- Bulanık Kümeleme
- Yoğunluk Tabanlı Kümeleme

2.2.3.4. Birliktelik kuralları

Birliktelik kuralları, veri tabanında bulunan veriler arasındaki bağlantıyı bulmak ve veriler arasındaki ilişkilerin sorgulanması için kullanılmaktadır. İlişkilendirme kuralı, veri tabanı içerisinde bulunan belirli nesnelerin eş zamanlı olarak, birlikte görülme olasılığının ifadesidir (Hand ve diğ, 2001).

Birliktelik kuralları yaygın olarak sepet analizi olarak bilinen yöntemle verilerin biriktirilmesi ve analiz edilmesi sürecidir (Alpaydın, 2004).

Birliktelik kuralları veriler arasında net olarak görülemeyen ilişkileri ortaya çıkarmaktadır. E-ticaret, perakende, ekonomi ve pazarlama sektörlerinde etkin olarak kullanılmaktadır. Özellikle E-ticaret alanında kullanılan birliktelik kuralları müşterilere ürün alma anında benzer ürün önerileri, müşteriye kişisel beğenebilecekleri farklı ürünleri de pazarlama amacıyla sıkça kullanılmaktadır.

2.2.3.5. Ardışık zaman örüntüleri

Ardışık zaman örüntüleri birbiri ardına gerçekleşen kayıtları incelemektedir. Veri seti üzerinden verinin eğilimi ve sonraki adımları belirlenir. Sonraki adımlar veri seti içerisinde birbiri ardına olan değişikliklerin frekansı incelenerek tahmin edilir. Ardışık zamanlı örüntülerin zaman içerisinde gerçekleşen olaylara dayanması bu algoritmayı birliktelik kurallarından farklılaştırmaktadır.

Müşterilerin farklı zamanlarda gösterdikleri davranışsal hareketler, satın alma alışkanlıkları gibi faaliyetleri arasında ilişkiler takip edildiğinde müşteri sadakat programları, kişiye özel kampanyalar geliştirme gibi imkanlar elde edilebilir ve sektörde rekabet avantajları sağlanabilir.

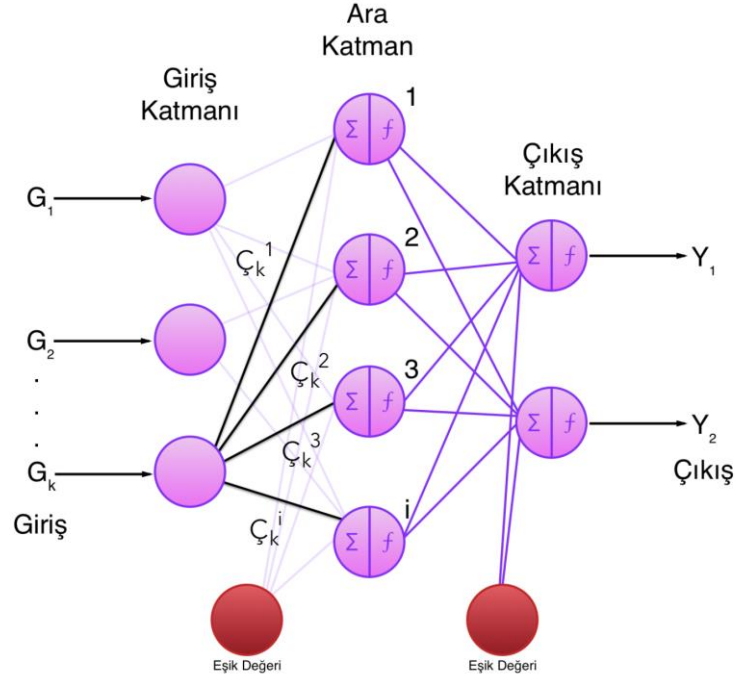
Tez çalışmasında ele alınan veri seti için sınıflandırıcı olarak Aşırı Öğrenme Makinesi (AÖM) kullanıldığından yöntemin temeli olan yapay sinir ağları ve AÖM ile ilgili bilgiler detaylı olarak sonraki kısımlarda açıklanmıştır. Parametre optimizasyonu için genetik algoritma kullanılmıştır. Nitelik seçiminde karar ağacı sınıflandırması kullanılmıştır. Çalışma kapsamında kullanılan diğer algoritmalar da aşağıda açıklanmıştır.

2.2.4. Yapay Sinir Ağları

Sinir Ağları, biyolojik nöronların bir ağını veya devrelerini belirtmek için kullanılır. Yapay nöron veya düğümlerin bileşimidir, dolayısıyla modern kullanımında yapay sinir ağı olarak belirtilir. Yapay sinir ağları, insan beyninin belirli bir görevi nasıl gerçekleştirdiğini modellemek için tasarlanmış, birbirine bağlı basit işlem birimleridir.

Sinir hücreleri aldıkları bilgileri aksonlar aracılığıyla diğer hücelere gönderirler. Sinir hücreleri etkileşimde olduğu diğer sinir hücrelerinden uyarıları alır, hücre gövdesine taşır ve gelen bilgileri toplam fonksiyonu ile toplar. Aktivasyon fonksiyonundan geçirilmesi ile hücre içi aktivasyonun (kararlılık) bozulmasıyla kimyasal süreç başlar. Süreç sonunca oluşan çıktı ağ üzerindeki bağlantılar üzerinden sonrakine gönderilir. Toplama ve aktivasyon fonksiyonları farklı çeşitlerdedir. Yapay sinir ağlarını arasındaki bağlantıların değerlerine ağırlık değerleri denmektedir (Öztemel, 2012).

Bir yapay sinir ağı Şekil 2.6' da görüldüğü üzere girdi katmanı, çıktı katmanı ve gizli katmandan oluşmaktadır. Gizli katmanda birden fazla katmanın olduğu görülebilir. Her bir katmanda nöronlar bulunmaktadır. Girdi katmanında, fonksiyonun değişken sayısı kadar, çıktı katmanında ise çıktı sayısı kadar nöron bulunur. Gizli katmanda yer alacak nöron sayısı belirli değildir. Her bir nöron kendisinden önceki çıktının nöronlarını girdi olarak kullanıp değer üretir. Gizli katmanda ve çıktı katmanında bulunan nöronlar aktivasyon fonksiyonunu kullanarak gelen bilgileri işler ve sonraki katmana iletir. Gizli katmanı bulunan yapay sinir ağları katmanlarında yeterli sayıda nöron bulundurduğunda farklı tipte doğrusal olmayan fonksiyonları modelleyebilir (Fausett, 1994).



Şekil 2.6: Yapay sinir ağlarının katmanları.

$$\zeta_k^i = G_k \quad (2.1)$$

Denklem 2.1’de girdi katmanındaki k .işlem elemanının çıktısının gösterimi verilmiştir.

Ağda bulunan her bir birim önceki düğümden gelen girdi değerini alır, bu değerle ağırlıklı toplamlarını hesaplar. Girdi verileri bağlantı ağırlıkları ile çarpılarak ağ üzerinde bir sonraki hücreye ilerler. Sistemin başlangıç noktasında kilitlenmemesi için ağırlıklar başlangıçta rasgele seçilir (Baxter, 2001).

Ağ katmanındaki her bir işlem önceki girdi katmanından gelen bilgileri bağlantı ağırlıklarının ($A_1, A_2, A_3, \dots$) hesaplanmış hali ile alır. Net girdi Denklem 2.2 kullanılarak önceki ara katmandan gelen bilgilerle hesaplanır. A_{kj} girdi elemanını j ara katman elemanına bağlayan bağlantının ağırlık değerini göstermektedir. j ara katman elemanın çıktısı ise bu net girdinin aktivasyon fonksiyonundan geçirilmesiyle hesaplanır. N bir nöronun bağlantı sayısını ifade eder. $A_{kj} \cdot \zeta_k^i$ her bir girişin ilgili nöronla bağlantısını ifade eder.

$$NET_j^a = \sum_{k=1}^n A_{kj} \zeta_k^i \quad (2.2)$$

Tahmin için en yaygın olarak kullanılan yapay sinir ağı tipi çok katmanlı algılayıcı (MLP) modelidir. Herhangi bir döngü oluşturmeyen ara bağlantılara sahip ağlara ileri beslemeli ağ denir (Catalao, 2007).

Genel ileri beslemeli (GFF) ağlar, ileri beslemeli ağların genel halidir. Genel ileri beslemeli ağlarda bağlantılar, bazı katmanlara uğramadan geçebilir ancak genel olarak problemin etkin şekilde çözümünü sağladığı için tercih edilir (Rumelhart, 1986).

Yapay sinir ağlarında genel olarak Sigmoid aktivasyon fonksiyonu kullanılır (Bambang et al., 2002). Girdi değerleri dışındaki her bir u_i birimi, yeni aktivasyon birimini, kendisine doğrudan bağlı girdilerin ağırlıklı toplam fonksiyonunu kullanarak hesaplar. İşlemler ileri doğru Denklem 2.3 ve 2.4'teki gibi tanımlanır.

$$S_i = \sum_{j=0}^n w_{i,j} * u_j \quad (2.3)$$

$$u_i = f(S_i) \text{ öyle ki } f(x) = \frac{1}{(1+e^{-x})} \quad (2.4)$$

Yapay sinir ağlarında geriye yayılım:

Yapay sinir ağlarının geriye yayılımı sırasında ağ her girdi deseni ile hedef çıktıyı karşılaştırır ve öğrenir. Hatayı hesaplar ve hata fonksiyonunu ağ üzerinde geriye doğru (çıktıdan girdiye) çalıştırır. Bu tip ağların en önemli avantajı kolay kullanım ve girdi/çıktı haritası için yakınsamadır. Düşük öğrenme oranlarına karşılık iyi bir performans ile tahminleme sergiler. Yavaş ve öğrenme süreci için çok fazla veri gereksinimi olması dezavantaj oluşturmaktadır.

Geriye yayılım ağı aktivasyon fonksiyonlarının türevleri kullanılmaktadır. Burada amaç seçilen amaç fonksiyonunu en aza indirmektir. Geri dönüş işlemlerinde her birimin çıktısı Denklem 2.5'teki gibi hesaplanır.

$$f'(S_i) = u_i * (1 - u_i) \quad (2.5)$$

Sonrasında w ağırlık değeri, Denklem 2.6' daki formül ile hesaplanır. Burada ε hata kareleri ortalaması, p adım uzunluğunu ifade eder.

$$\delta_i = -\left(\frac{\partial \varepsilon}{\partial S_i}\right) \quad (2.6)$$

$$w_{*,i,j} = w_{i,j} + p\delta_i u_j \quad (2.7)$$

Yapay sinir ağlarında öğrenme süreci; model çalıştırılmadan önce belirlenmiş olan bazı metriklerle ulaşıncaya kadar tekrarlanır. En küçük kareler yöntemi ile hesaplanacak hatanın kabul edilebilir düzeydeki bir seviyeye inmesi durumunda veya hatanın belli bir noktadan sonra artmaya başlaması koşullarına kadar tekrarlanır.

Öğrenme süreci tamamlandıktan sonra ağırlık değerleri belirlenmiş olan ağ, test verisi olarak ayrılan veri setinde test edilir. Böylece öğrenme seti içerisinde ulaşılan örüntülerle yapılan tahmin, ayrılan test verisi ile karşılaştırılır. Genelleme yapılmak istenirse ağ yeterli büyüklükteki veri kümesi ile eğitilmelidir (Yeh, 1998).

Yapay sinir ağları müşteri profillerinin sınıflandırılması, tıbbi teşhislerin sınıflandırılması, ses ve görüntülerin sınıflandırılması, satış tahminleri, hava tahminleri, erken uyarı sistemleri, veri birleştirme, veri kavramsallaştırma ve veri filtreleme gibi alanlarda kullanılır. Endüstriyel uygulamalar, tıp ve sağlık alanındaki uygulamalar, finans uygulamaları, askeri ve savunma sanayi uygulamaları, mühendislik uygulamaları gibi birçok alanda kullanılır.

Yapay sinir ağlarının avantajları olarak şunlardan bahsedilebilir. Veri modelleme için yüksek performansta bilgisayarlara ihtiyaç yoktur. Yapay sinir ağları hata toleranslarını karşılayabilirler. Birden çok modelin bir arada çalışması için tasarlandığından paralel çalışma yetenekleri mevcuttur. Birçok öğrenme algoritmasının temeli olarak uygulanmıştır. Veri ön işleme işlemleri öncesinde verinin ilk haliyle de anlamlı sonuçlar görülebilir.

Avantajlarının yanında bazı dezavantajları da bulunmaktadır. Yapay sinir ağları bilgi yorumlamada sınırlı nedensel çıkarım yeteneğine sahiptir. Klasik yöntemlere göre uygulanması daha zor olabilir. Klasik modellerle karşılaştırıldığında model çalışma süresi daha uzun sürebilir. Sinir ağı modelleri aşırı uyumsuz olma eğilimindedir.

2.2.5. Aşırı Öğrenme Makinesi

Bu bölümde tezde detaylı olarak incelenen ve uygulamada kullanılan aşırı öğrenme makinesi sınıflandırıcısı ile ilgili bilgilere yer verilmiş ve aşırı öğrenme makineleri literatürü incelenmiştir.

Temel Aşırı Öğrenme modeli, tek gizli katmanı olan ileri beslemeli yapay sinir ağlarını temel alır. Yapay sinir ağlarından farklı olarak geleneksel tek katmanlı ileri beslemeli yapay sinir ağı modelinde girdi ve çıktı sayıları bilinir. Geleneksel yapay sinir ağlarında modeldeki ağırlıklar

tekrarlanarak hesaplanmaktadır. Ancak Huang ve diğ. model ağırlıklarını analitik bir şekilde tek seferde hesaplayacak yeni bir yöntem önermiştir. Yeni yöntemle çalıştırılan modellerin sonuçlarının birbirinden tamamen farklı olacağı kanıtlanmıştır. Yöntem aşırı öğrenme makineleri olarak adlandırılmış ve literatüre girdiği günden itibaren detaylı olarak incelenmeye devam edilmiştir.

Aşırı öğrenme makinesi ilk olarak Huang'ın ve diğ. 2006 yılında yaptığı çalışmalarla birlikte literatüre girmiştir. Yaptığı çalışmada girdi setleri üzerinde evrensel yaklaşım ve sınırlı eğitim seti örneklerinde yaklaşım üzerinde odaklanmıştır. Çalışma sonucunda gizli düğümleri rasgele seçen ve tek gizli katmanlı ileri beslemeli sinir ağı çıktı ağırlıklarını, analitik olarak belirleyen bir teori çıkarılmıştır. Bu teori, öğrenme hızında yüksek performans yakalamıştır (Huang, 2006).

Huang ve diğ. (2007) AÖM ile ilgili çalışmalarına devam etmiş. Rasgele gizli düğümleri olan artırimsal yapıcı besleme ağılarını kullanan artışı aşırı öğrenme makinesi algoritmasını geliştirmişler. Bu algoritma birçok popüler öğrenme algoritmasından daha iyi performans gösterir. I-ELM gizli düğümleri rasgele üretir ve analitik olarak tek gizli katmanlı ileri beslemeli sinir ağılarının çıktı ağırlıklarını hesaplar, ancak I-ELM tüm çıktı ağırlıklarını yeniden hesaplamaz. Gizli düğümün rasgele eklenmesiyle dışbükeylik optimizasyon yöntemine dayalı mevcut düğümlerin çıktı ağırlıklarını yeniden hesaplayarak I-ELM'nin yakınsama oranının daha da geliştirilebileceği kanıtlanmıştır.

Sun ve diğ. (2008) moda sektöründe perakende uygulamalarında aşırı öğrenme makinesi kullanarak satış tahminleri yapmıştır. Çalışmada aşırı öğrenme makineleri, geri yayılım sinir ağılarına dayalı çeşitli satış tahmin yöntemlerinden daha iyi performans sonuçları göstermiştir.

Yuan ve diğ. (2009) veriyi birer birer, sabit ya da çeşitli yığın boyutu ile parçalayarak öğrenebilen çevrimiçi ardışık aşırı öğrenme makinesi (OS-ELM) adlı çevrimiçi sıralı öğrenme algoritmasını geliştirerek yeni bir yaklaşım önerdi. Çalışmada OS-ELM'ye dayalı çevrimiçi ardışık aşırı öğrenme makinesi (EOS-ELM) topluluğu geliştirilmiştir. Her iki algoritma 10 farklı veri seti üzerinde çalıştırılmış ve sınıflandırma performansı sonuçları karşılaştırılarak EOS-ELM'nin orijinal OS-ELM'den daha kararlı ve doğru olduğu gösterilmiştir.

Wong ve Guo (2010) moda perakende tedarik zincirindeki satış tahmini problemlerini çözmek için veri ön işleme bileşeni ve hibrit zeka tahmincisi içeren bir hibrit zeka (HI) modeli

geliştirmiştir. Optimal sinir ağı ağırlıklarını elde etmek için yeni bir öğrenme algoritması oluşturmak üzere AÖM ile yeni bir meta-buluşsal optimizasyon tekniği, harmoni arama algoritması ilk kez entegre edilmiştir ve daha iyi sinir ağı genelleme performansı elde edilmiştir. Geliştirilen hibrit zeka modeli, kısa ürün ömrü döngüsü, değişken müşteri talepleri ve çok sayıda ürün çeşitleri gibi dinamik parametreler içeren moda perakende tedarik zincirindeki orta vadeli moda satışlarının doğrusal olmayan düzensiz dağılımını etkili bir şekilde ele alabilmiştir.

Miche ve diğ. (2010), çok parçalı aşırı öğrenme makinesi (OP-ELM) modeli üzerine çalışma yapmıştır. Mevcut algoritmayı daha kuvvetli ve genel hale getirmek için geliştirmeler yaparak regresyon ve sınıflandırma problemi üzerinde uygulamıştır. Hem hesaplama zamanı hem de doğruluk değerleri açısından orijinal AÖM ve birkaç sınıflandırma algoritmasıyla karşılaştırılmıştır. Önerilen yöntemin AÖM dışında karşılaştırılan diğer algoritmalarından birkaç kat daha hızlı performans sergilediği görülmüştür.

Huang ve diğ. (2010) standart optimizasyon yöntemi açısından AÖM' yi sınıflandırma için daha ayrıntılı olarak incelemiş ve AÖM' yi spesifik bir geliştirilmiş SLFN-destek vektör ağı tipine genişletmiştir. Bu çalışmayla SVM' nin maksimum tolerans mülkiyeti ve ileri beslemeli sinir ağlarının ağırlık teorisinin asgari normunun tutarlı olduğu görülmüştür. Simülasyon sonuçları ile AÖM' nin, geleneksel SVM' den daha iyi genelleme performansı elde etme eğiliminde olduğu görülmüştür.

AÖM yönteminin yaratıcısı Huang ve diğ. (2011) tarafından gerçekleştirilen çalışmada yöntemin çeşitli varyasyonları detaylı bir şekilde incelenmiştir.

Xia ve diğ. (2012), AÖM ile elektrik fiyat tahmini gerçekleştirmişlerdir. Çalışmalarında AÖM yöntemi ile gerçekleştirilen tahmin sonuçları, yapay sinir ağları, destek vektör makineleri, ARCH, GARCH ve bulanık YSA gibi yöntemlerle gerçekleştirilen tahmin sonuçları ile karşılaştırmış ve önerdikleri yöntemin daha iyi sonuç verdiğini raporlamışlardır. 1982 ile 2012 yılları arasındaki 360 aylık veri ile çalışmalar yapılmış olup, diğer modellerden daha iyi sonuçlar elde edildiği görülmüştür.

Chi-Jie Lu (2012), bilgisayar ürünleri satış tahmini amacıyla AÖM' yi kullanmıştır. Çalışmasında toplu ampirik model ayrıştırma ve AÖM' yi birleştiren hibrit bir model önermiştir. Sabit disk, ekran kartı ve dizüstü bilgisayar ürünlerinin satış verilerinden elde edilen

deneysel sonuçlar, önerilen hibrit satış tahmin yönteminin, dört karşılaştırma modelinden daha iyi performans gösterdiği görülmüştür.

Alçın ve diğ. (2015) AÖM' nin dezavantajlarını inceleyerek, sorunların giderilmesi için seyrek tabanlı bir AÖM modeli önermişlerdir. Seyrek tabanlı AÖM modelinde, çıkış ağırlıklarının temsilini hesaplarken yinelemeli fırsatçı takip algoritması tabanlı İleri Geri Takip algoritması kullanılmıştır. Bazı sınıflandırma problemleri üzerinde yapılan çalışmalarla İGT-AÖM yönteminin üstünlük sağladığı görülmüştür. Ayrıca önerilen yöntemin başarısını ölçmek için normal AÖM, en küçük açısız regresyon, seçme operatörü, en küçük mutlak daralma ile önerilen metodun karşılaştırılması yapılmıştır.

İleri beslemeli yapay sinir ağları ile ilgili yapılan çalışmalarda öğrenme hızlarının gerekenden çok daha yavaş olduğu görülmüştür. Bunun ardında iki temel unsur tespit edilmiştir. Birincisi yavaş gradyan tabanlı öğrenme algoritmaları; sinir ağlarını eğitmek için yoğun olarak bu algoritmalar kullanılır. İkincisi bu algoritmaların doğru sonuçlar ortaya çıkarmaları için mimarideki bütün ağırlıkları yenilemeli bir şekilde değiştirmek zorunda olmalarıdır.

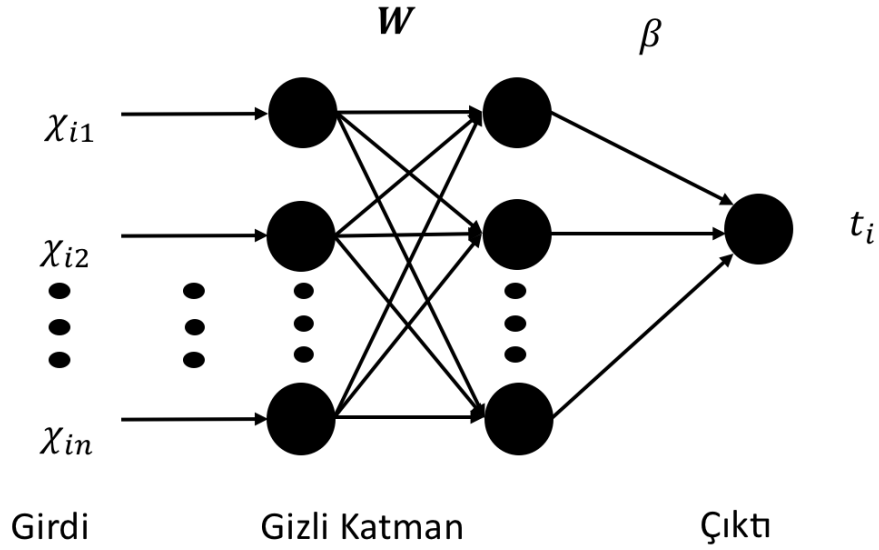
Birkaç yapay ve gerçek kıyaslama fonksiyonu yaklaşımı ve çok büyük karmaşık uygulamalar dahil olmak üzere sınıflandırma problemlerine dayanan deneysel sonuçlar yeni algoritmanın çoğu durumda iyi genelleme performansı üretebildiğini ve ileri beslemeli sinir ağları için geleneksel popüler öğrenme algoritmalarından binlerce kez daha hızlı öğrenebileceğini göstermektedir.

Geleneksel olarak, ileri beslemeli ağların tüm parametrelerine ayar yapılması gerekir ve parametrelerin farklı katmanları arasında bazı bağımlılıklar vardır. AÖM geliştirilmeden önceki yıllarda eğim iniş temelli yöntemler baz alınarak ileri beslemeli sinir ağlarının çeşitli öğrenme algoritmalarında kullanılmıştır. İniş temelli öğrenme yöntemlerinin, yanlış öğrenme adımlarından dolayı genellikle çok yavaş olduğu veya yerel minime kolayca yakınsayabileceği görülmüştür.

Geleneksel öğrenme algoritmalarından farklı olarak, AÖM sadece en küçük eğitim hatasına ve aynı zamanda en küçük ağırlık normuna ulaşma eğiliminde değildir. Daha küçük eğitim hatasına ulaşan beslemeli sinir ağları için, beslemeli sinir ağlarının genelleme performansı, ağırlıkların normu ne kadar küçükse ağların eğitimi için daha iyi genelleme performansı gösterir. Bu nedenle AÖM, ileri beslemeli sinir ağları için iyi bir genelleme performansı

gösterme eğilimindedir. Kolaylıkla uygulanabilir olduğu için, ağırlıkların en küçük normunu ve iyi genelleme performansını elde eder. Diğer popüler Tek Gizli Katmanlı İleri Beslemeli Sinir Ağları (SLFN) öğrenme algoritmalarına göre son derece hızlı çalışır.

Aşırı öğrenme modeli Şekil 2.7’de görüldüğü üzere girdi katmanı, gizli katman ve çıktı katmanından oluşur (Huang, 2006).



Şekil 2.7: Aşırı öğrenme makinesi.

Aşırı Öğrenme Makineleri tek gizli katmanlı ileri beslemeli sinir ağıdır. Giriş ağırlık matrisi rasgele seçilir. $\vec{W}$ değeri, i . giriş sinir hücresi ve i . gizli sinir hücresine bağlı ağırlık vektörüdür. β değeri, i . gizli sinir hücresine ve çıkış sinir hücrelerine bağlı olan ağırlık vektörüdür. İleri beslemeli sinir ağının çıktı ağırlık matrisi analitik çözümü β' yı belirler.

N rasgele seçilen eğitim verisi seti örnekleri (Denklem 2.8. ve Denklem 2.9); öğrenmek için k adet gizli nöronları ve bir aktivasyon fonksiyonu vektörü Denklem 2.10 olan bir tek gizli katmanlı ileri beslemeli sinir ağı (SLFN) eğitildiği varsayılırsa Denklem 2.11 modeline ulaşılır.

$$x_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in R_n \quad (2.8)$$

$$t_i = [t_{i1}, t_{i2}, \dots, t_{in}]^T \in R_m \quad (2.9)$$

$$g(x) = (g_1(x), g_2(x), \dots, g_k(x)) \quad (2.10)$$

$$\sum_{i=1}^K \beta_i g_i (w_i \cdot x_j + b_i) = \pi t_j \cdot j = 1, \dots, N, \quad (2.11)$$

Denklem 2.11’de, Denklem 2.12 gizli nöron ve giriş nöronlarını bağlayan ağırlık vektörü, Denklem 2.13 gizli nöron ile çıktı nöronlarının birleştiren ağırlık vektörünü ifade eder.

$$w_i = [w_{i1}, \dots, w_m] \quad (2.12)$$

$$\beta_i = [\beta_{i1}, \dots, \beta_m] \quad (2.13)$$

b_i i. gizli düğümün rasgele seçilen eşik değeridir.

SLFN' ler N örneği sıfır hata ile yaklaşık olarak değerlendirebilirse aşağıdaki eşitlik elde edilir.

$$\sum_{j=1}^N \|y_j - t_j\| = 0 \quad (2.14)$$

Denklem 2.14 SLFN' nin çıkış değeridir.

Tüm YSA modellerinde olduğu gibi buradaki de amaç, amaç ölçütünün en aza indirgenmesidir.

Denklem 2.11’ de $w_i \cdot x_j$ işlemi; w_i ve x_j ’ nin iç çarpımını göstermektedir. Çıkış düğümleri lineer olarak seçilmektedir. Bu denklem kompakt olarak Denklem 2.15’deki gibi yazılabilir:

$$H \cdot \beta = T \quad (2.15)$$

Burada Denklem 2.17 gizli katman çıktı matrisidir, Denklem 2.16 X_i , Denklem 2.18 ile ilgili gizli nöron çıktı ağırlıklarının matrisidir. Denklem 2.19 amaç matrisidir.

$$H = \{h_{ij}\} (i = 1, \dots, N \text{ ve } j = 1, \dots, k), \quad h_{ij} = g(w_j \cdot x_i + b_j)x_i \quad (2.16)$$

$$H(W_1 \dots W_M; b_1 \dots b_M; X_1 \dots X_N) = \begin{bmatrix} g(W_1 X_1 + b_1) & \dots & g(W_M X_1 + b_M) \\ \vdots & \ddots & \vdots \\ g(W_1 X_N + b_1) & \dots & g(W_M X_N + b_M) \end{bmatrix} \quad (2.17)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_M^T \end{bmatrix} \quad (2.18)$$

$$T = \begin{bmatrix} t_1^T \\ \vdots \\ t_M^T \end{bmatrix} \quad (2.19)$$

AÖM' de, girdi ağırlıkları ve gizli sapmalar, rasgele üretilir. Böylece, çıktı ağırlıklarının belirlenmesi (gizli katmanın çıktı katmanına bağlanması), verilen doğrusal sisteme en küçük kareler çözümü bulma kadar basittir. Doğrusal sisteme ait minimum en küçük kareler çözümü ile aşağıdaki Denklem 2.20 eşitliği elde edilir.

$$\hat{\beta} = H^{\dagger}T \quad (2.20)$$

Yukarıdaki denklem β ' nin tahmini değeri olarak kullanılır. $H^{\dagger}$, H matrisinin Moore–Penrose genelleştirilmiş tersidir.

Minimum normal doğrusal sistem çözümü benzersizdir ve tüm doğrusal sistem çözümleri arasında en küçük norma sahiptir. AÖM, Moore–Penrose ters yöntemi kullanarak öğrenme hızında belirgin bir artış sağlayan iyi bir genelleme performansı elde etme eğilimindedir (Choi ve diğ., 2006).

AÖM algoritması aşağıdaki gibi özetlenir.

Adım 1: Rasgele gizli düğüm parametreleri üreterek rasgele gizli düğümler ata.

$$(w_i, b_i), i = 1, 2, \dots, L.$$

Adım 2: Gizli katman çıkış matrisini hesapla.

$$H$$

Adım 3: Çıkış ağırlık matrisini hesapla.

$$\hat{\beta} = H^{\dagger}T$$

Aşırı Öğrenme Makineleri için farklı aktivasyon fonksiyonları kullanılabilir. Aktivasyon fonksiyonlarının çeşidi farklı sınıflandırma performansları doğurur. AÖM' de kullanılan aktivasyon fonksiyonları Tablo 2.3'de gösterilmiştir.

Tablo 2.3: AÖM' de kullanılabilen aktivasyon fonksiyonları.

Aktivasyon Fonksiyonu	Denklemleri
1 Sigmoid	$a = \frac{1}{1 + \exp(-n)}$
2 Sinüs	$a = \sin(n)$
3 Kosinüs	$a = \cos(n)$
4 Kesin Sınırlayıcı (Hard Limit)	$a = \begin{cases} 1, & \text{eğer } n \geq 0 \\ 0, & \text{diğer durumlarda} \end{cases}$
5 Üçgensel Bazlı	$a = \begin{cases} 1 - \text{abs}(n), & \text{eğer } -1 \leq n \leq 1 \\ 0, & \text{diğer durumlarda} \end{cases}$
6 Radyal Bazlı	$a = \exp(-n^2)$
7 Normalleştirilmiş Radyal Bazlı	$a = \frac{\exp(-n^2)}{\sum \exp(-n^2)}$
8 Elliot Sigmoid	$a = \frac{0.5 * n}{(1 + \text{abs}(n))} + 0.5$
9 Log-sigmoid	$a = \frac{1}{(1 + \exp(-n))}$
10 Saf Doğrusal	$a = n$
11 Soft Max	$a = \frac{\exp(n)}{\sum(\exp(n))}$
12 Ters (Inverse)	$a = \frac{1}{n}$
13 Doygun (Saturated) Doğrusal	$a = \begin{cases} 0, & \text{eğer } n \leq 0 \\ n, & \text{eğer } 0 \leq n \leq 1 \\ 1, & \text{eğer } 1 \leq n \end{cases}$
14 Pozitif Doğrusal	$a = \begin{cases} n, & \text{eğer } n \geq 0 \\ \text{eğer } n < 0 \end{cases}$

AÖM sadece geleneksel gradyen tabanlı öğrenme algoritmalarından çok daha hızlı öğrenmesiyle kalmayıp, aynı zamanda durdurma kriterleri, öğrenme oranı, öğrenme aralıkları, yerel minimumlar ve aşırı duyarlı problemler gibi gradyen tabanlı öğrenme yöntemleri tarafından karşılaşılan birçok zorluğun önüne geçer.

2.2.6. Genetik Algoritma

Çalışmada vadeli mevduat kampanyasına katılım yapabilecek müşterilerin AÖM sınıflandırıcısıyla en doğru oranda sınıflandırılabilmesi için genetik algoritma optimizasyonundan faydalanılmıştır.

Genetik algoritmalar, matematiksel hesaplama boyunca biyolojik evrim teorisini taklit eden güçlü bir optimizasyon paradigmasını temsil eder. Genetik algoritmalar, probleme çeşitli çözümleri rasgele değerlendiren ebeveyn kromozomları ile başlar. Bir kromozom içindeki bileşenler gen olarak adlandırılır. Yeni çocuk kromozomlar çaprazlama ve mutasyon ile

oluşturulur. Çeşitlilik iki veya daha fazla kromozom arasında yapılan olasılıksal gen değişimi ile gerçekleşir. Mutasyon, kromozomdaki genlerin rasgele değişmesiyle gerçekleşir. Tüm kromozomlar daha sonra uyum fonksiyonuna göre değerlendirilir ve en uygun olan kromozom gelecek nesile aktarılır (Cook, 2000). Bir probleme daha iyi çözümler üretmek için zaman içerisinde gen havuzu oluşturan algoritmalarlardır.

Genetik algoritmalar; işyeri planlaması, üretim planlama, hat dengeleme, kereste kesme optimizasyonu ve süreç optimizasyonu için başarılı bir şekilde kullanılmıştır (Cook, 1991).

Parametre optimizasyonu için genetik algoritmaların kullanıldığı birçok çalışma yer almaktadır.

Houck ve diğ. (2008), genetik algoritmayı parametre optimizasyonu için kullanmıştır. MATLAB yazılım programında yaptıkları çalışma ile program üzerinde araç kutusu oluşturmuşlardır. Böylece program içerisinde matematiksel kapasitesi yüksek, esnek, rahatça kullanılabilen modüler bir araç kutusunu oluşturarak implemente etmişlerdir.

Wu ve diğ. (2007), genetik algoritma tabanlı karar destek makineleri modeli önermişlerdir. Model sonucunda ulaşılan parametreleri diğer algoritmalar ile karşılaştırmıştır. Karar destek makineleri, sinir ağları, lojistik regresyon modelleriyle karşılaştırıldığında önerilen yöntemin geleneksel karar destek makinelerine göre daha başarılı olduğu görülmüştür.

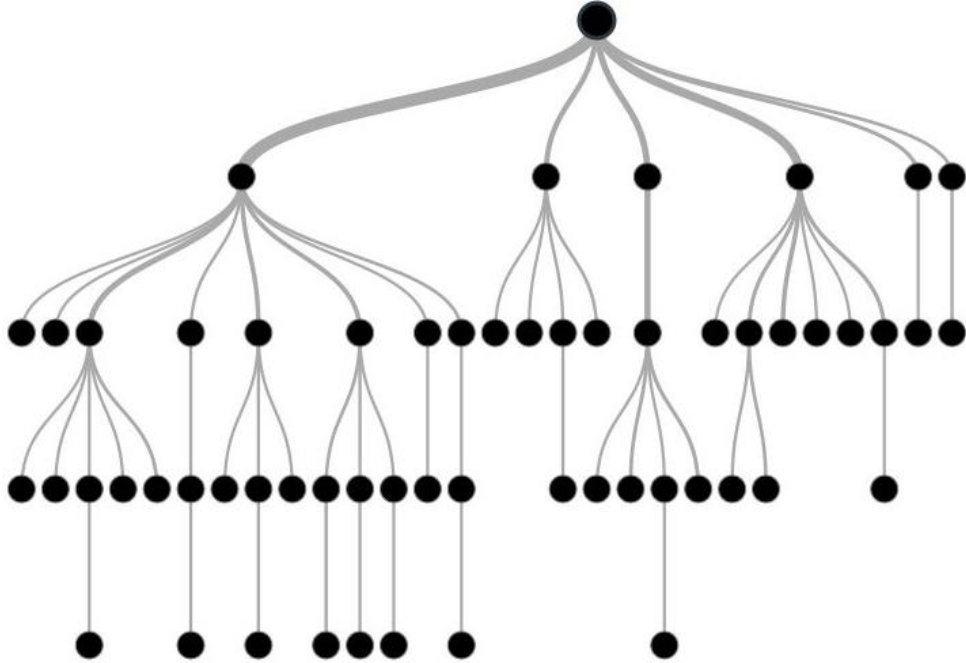
Kim ve diğ. (2007), optimizasyon problemlerinin çözümü için literatürde ilk defa *Escherichia coli* bakterisi davranışını modelleyen bir çalışma yapmıştır. Önerilen yöntemle otomatik voltaj regresörü için ayarlama yapmışlardır. Simulasyon sonuçlarıyla etkili ve kolayca genişletilebilecek bir yaklaşım olduğu görülmüştür.

Friedrichs ve Igel (2005), karar destek sistemleri modeli ile yapılan çalışmada model parametrelerinin belirlenmesinde genetik algoritma yaklaşımını kullanmıştır. Izgara araması yöntemiyle karşılaştırılarak en iyi parametre değerlerine göre ilgili sınıflandırıcı performansı karşılaştırılmıştır.

2.2.7. Karar Ağacı Sınıflandırıcısı

Tez çalışması kapsamında niteliklerin seçimi için karar ağacı sınıflandırıcısından faydalanılmıştır. Girdilerin seçimi sınıflandırma sonuçlarını etkileyen temel unsurlardandır. Çalışmada yer alan 21 niteliğin amaç niteliğine olan etkisini değerlendirmede. U yöntemden faydalanılmıştır.

Karar ağaçları hedef değişkenleri veri seti içerisinde yer alır. En baştan, aşağı doğru ilerleyen bir süreci vardır. Karar ağacı sınıflandırıcıları, ağaçlar üzerinde örneklere dayanarak oluşturulur. Ağaç yapısında yer alan düğümler ikiye ayrılır. Kök ve iç düğümler ile yaprak düğümleri. Kök ve iç düğümler özniteliklerle ilişkilendirilir, yaprak düğümleri sınıflarla ilişkilendirilir. Temel olarak, her yaprak olmayan düğümün, düğümle ilişkili özniteliğin olası her değeri için giden bir dalı vardır. Bir karar ağacını kullanarak yeni bir örneğin sınıfını belirlemek için, kök değerden başlayarak, bir yaprak düğümüne ulaşılan kadar ardışık iç düğümler kontrol edilir. Kök düğüme ve her bir iç düğüme bir test uygulanır. Testin sonucu, geçiş yapılan dalı ve bir sonraki düğümü ziyaret eder. Örnek için sınıf, son yaprak düğümünün sınıfıdır. Karar ağacının gösterimi Şekil 2.8’de verilmiştir.



Şekil 2.8: Karar ağacı gösterimi.

Bu çalışmada CART (Classification and Regression Trees) sınıflandırma algoritması kullanılmıştır. Bu algoritmada en iyi dallara ayırma kriterini belirlemek için entropiden değerinden faydalanır (Dunham, 2003). Fakat ayırma kriterini belirlemek için C4.5 ve ID3 algoritmalarından farklı bir formül kullanır ve yine ID3 algoritmasından farklı olarak dallara ayırma kriterini hesaplarken kaybolan verileri önemsemez.

Kategorik değişkenler ile yapılan karar ağacı uygulamalarında performans ölçümü için en sık olarak kullanılan yöntemler entropi, gini değeri ve sınıflandırma hatasıdır. Nümerik verilerde ise En Küçük Kareler yöntemi kullanılır.

Entropi, veri ile ilgili belirsizliği ifade eder. Veri seti içerisinde sadece bir sınıf/etiket varsa entropi değeri düşüktür. Veri entropi değerini en düşük seviyeye getirecek şekilde bölünmelidir. Bölünmeler ne kadar iyi olursa, tahminleme de o kadar iyi olur.

Denklem 2.15’de entropi hesaplaması verilmiştir.

$$H = - \sum p(x) \log p(x) \quad (2.15)$$

$p(x)$ belirli bir sınıfa ait grubun yüzdesini ve H ise entropiyi belirtmektedir. Veri, karar ağacının entropi değerini en aza indirgeyecek şekilde bölünmelidir. En iyi bölünmeyi belirlemek içinde bilgi kazancı kullanılır. Bilgi kazancı aşağıdaki Denklem 2.16 ile hesaplanır:

$$Gain(S, D) = H(S) - \sum_{V \in D} \frac{|V|}{|S|} H(V) \quad (2.16)$$

Bilgi kazancında S orijinal veri kümesidir ve D ise veri kümesinin bölünmüş bir parçasıdır. V kümeleri S kümesinin alt kümeleridir. V ’nin tümü ayrıktır ve S ’yi oluşturmaktadır. Bilgi kazancı, veri seti bölünmeden önceki orijinal hali ile hesaplanan entropi ve her bir özneliliğin entropi değeri arasındaki fark olarak tanımlanır.

Karar ağaçlarının avantajları:

- Görsel gösterimi olduğundan anlaşılması kolaydır.
- Veri ön hazırlığı diğer modellere göre daha kısa sürer, ancak kayıp verinin analizi önemlidir. Kayıp veriyi desteklemez.
- Ağaç maliyeti, ağacı eğitmek için kullanılan örneklemelerin çeşitliliğine bağlıdır.
- Hem sayısal hem de kategorik veriler işlenebilir.

- Çıktı sayısının çok olduğu problemler ele alınabilir.
- İstatistiksel testlerle model doğrulaması sağlanabilir.

Karar ağaçlarının dezavantajları:

- Veri setinde fazla sayıda etiket olması, karmaşık veri setlerinde büyük boyutta karar ağacı oluşmasına sebep olabilir. Bu durumda ağaç dallanması takip edilemeyebilir.
- Aşırı öğrenme görülebilir. Aşırı öğrenmenin engellenmesi için model parametrelerinde ayar ve veri azaltma gibi yöntemler kullanılabilir. Veri azaltma işlemi, az sayıda örnekleme içeren yaprak düğümlerin karar ağacı çiziminden atılmasını ifade etmektedir.

Bu çalışmada nitelik seçimi için karar ağacı sınıflandırıcısı kullanılmıştır. Literatürde nitelik seçiminde karar ağaçların kullanıldığı birçok çalışma vardır. Bu çalışmalardan bazıları şu şekilde özetlenebilir.

Grabczewski ve Jankowski (2005) yaptığı çalışmada karar ağacı sınıflandırmasıyla nitelik seçiminin avantajlarını ele almıştır. Ayrık değerler ile ilgili tabanlı algoritmalarla sürekli ve kategorik veriler üzerinde inceleme yapılmıştır.

Tahir ve diğ. (2006), yaptığı çalışma ile nitelik seçiminde Kümülatif Varyans ve Scree testini içeren temel bileşen analizinin temel kurallarına dayanan en iyi öz vektörleri belirlemek için bir yöntem geliştirmişlerdir.

Esmaili ve Gabor (2009) tarafından yapılan çalışmada özellik sıralaması algoritması kullanılmış ve kural geliştirmek için yukarıdan aşağı karar ağaçlarının indüksiyonu kullanılmıştır. Ağacın büyüklüğü, yaprak sayısı, hata oranı, geri çağırma ve hassasiyet gibi değerlendirme parametreleri hesaplama sonuçları önerilen yöntemle oluşturulan ortalama yaprak sayısının ve karar ağacı boyutunun diğer algoritmalarla karşılaştırıldığında daha iyi olduğunu görülmüştür. J48 veya BFTree gibi algoritmalarla karşılaştırıldığında daha küçük bir ağaç ürettiği görülmüştür.

Gayatari ve diğ. (2010) tarafından yapılan çalışmada karar verme indüksiyonuna dayanan özellik seçimi için yeni bir yöntem önerilmiştir. Karar ağacı indüksiyon prosesi ile elde edilen bu yeni veri setiyle farklı sınıflandırıcılar öğrenilmiş ve daha iyi nitelik seçimi performansları elde edilmiştir. Karar ağacı ile nitelik seçimleri Destek Vektör Makineleri ve RELIEF özellik

seçim teknikleriyle yapılan nitelik seçimleriyle karşılaştırıldığında önerilen yöntemin, diğer iki grup için daha iyi performans gösterdiği gözlenmiştir.

2.2.8. K En Yakın Komşu Algoritması

K-en yakın komşu (KNN) algoritmaları, analogik öğrenmeye dayanır. Bilinmeyen bir örnek verildiğinde, bilinmeyen örneğe en yakın olan KNN için desen benzerliği arar. Yakınlık mesafe cinsinden tanımlanır. Bilinmeyen örnek, KNN arasında en yaygın olan sınıfa atanır. Bu yaklaşımın en büyük avantajı, sınıflandırmadan önce öngörü modelinin kurulmasına gerek olmamasıdır. Dezavantajlar ise KNN'nin basit bir sınıflandırma olasılık formülü üretmemesi ve tahmin doğruluğunun komşunun uzaklığı ve kardinalliği k ölçüsünden büyük ölçüde etkilenmesidir (Yeh ve Lien, 2007).

KNN algoritmaları, sınıflandırma ve tahmin ve tahmin amaçlı olarak kullanılmaktadır. K-en yakın komşu, örnek tabanlı öğrenme metotlarından biridir. KNN uygulanmadan, eğitim veri setinin orijinal hali saklanmalıdır. Böylece sınıflandırılmamış yeni bir kayıt için bir sınıflandırma temel olarak eğitim setindeki benzer kayıtlarla karşılaştırarak keşfedilebilir.

2.2.9. Naive Bayes

Bayes ağları, yapay zeka, istatistik ve olasılığın kesişiminde yer almaktadır. Bilgi keşfi teknikleri arasında önemli bir yere sahiptir. Bayes ağlarının temel özellikleri, grafik ve olasılık teorileri kombinasyonundan ortaya çıkan genel bir model sınıfı olarak olasılıksal olmasıdır. Modelin ana avantajı, karmaşık olasılık modelleri ile bunları daha küçük bileşenlere ayırarak idare edebilmektir.

Bir olasılık modeli, düğümlerin stokastik değişkenleri sembolize ettiği ve yayların bu değişkenler arasında simgelandığı bir grafikte karakterize edilir. Grafik değişkenlerini bağlayan olasılıklı bir grafik olduğundan Bayes ağı olarak adlandırılır.

Bayes ağının iki bileşeni vardır: yönlendirilmiş bir asiklik grafik ve olasılık dağılımı. Yönlendirilmiş asiklik grafikteki düğümler, stokastik değişkenleri temsil eder ve yaylar, koşullu olasılık dağılımları ile ölçülen değişkenler arasındaki yönlendirilmiş bağımlılıkları temsil eder (Maimon ve Rokach, 2005).

2.2.10. Gradyen Artırma Modeli

Düğüm artırma, gerileme ve sınıflandırma problemleri için bir makine öğrenme tekniği olup, zayıf tahmin modelleri, genellikle karar ağaçları şeklinde bir tahmin modeli üretir (Friedman,2002).

Yükseltme, diğer hızlandırma yöntemleri gibi uygun bir maliyet fonksiyonu üzerinde bir optimizasyon algoritması olarak yorumlanabilir. Gradyen artırma, zayıf öğrencileri tek bir güçlü öğreniciye tekrarlayan şekilde birleştirir (Jain, 2015).

2.2.11. Rasgele orman

Rasgele ormanlar, hem sınıflandırma hem de regresyon problemleri için bir topluluk öğrenme yöntemidir (Breiman, 2001). Rasgele Ormanlar, özellik alanının rasgele seçilen alt alanlarında büyüyen karar ağaçlarının bir araya gelmiş halidir.

Rasgele ağaç algoritmalarının çalışma prensibi, her biri öğrenme örneğinden gelen bir önyükleme örneği ve her düğümde rasgele seçilen bir özellik alt kümesi kullanılarak oluşturulan bir dizi ikili karar ağacını birleştirmektir. Dolayısıyla, CART model oluşturma stratejisindeki durumdan farklı olarak, rasgele tek bir ağaç, öğrenme noktalarının bir alt kümesi üzerinde ve her düğümde ayrılacak özellik alt kümeleri üzerinde kurulmaktadır.

Ayrıca ormandaki ağaçlar maksimum boyuta kadar büyütülmekte ve budama basamağı atlanmaktadır. Topluluktaki ağaçların önyükleme numuneleri kullanılarak yerleştirilmesinden sonra, nihai karar, topluluk üzerinde toplanarak, yani gerileme çıktısının ortalaması alınarak veya sınıflandırma için oylama ile elde edilir.

Torbalama denen bu prosedür, modelin stabilitesini ve doğruluğunu artırır, varyansı azaltır ve aşırı öğrenmeyi önlemeye yardımcı olur (Dudek, 2014).

2.2.12. Lojistik Regresyon

Lojistik regresyon, ikili kategorik hedefler için geliştirilen özel bir regresyon şekli olan bir öngörücü modelleme tekniğidir. İki veya daha fazla bağımsız değişken ve tek bir ikili bağımlı değişken arasındaki ilişkiyi bulur. Lojistik regresyon, bağımlı değişkenin türüne, tahmin yöntemlerine ve altta yatan dağılım hakkındaki varsayımlara göre çoklu regresyondan farklıdır.

Lojistik regresyon, normal dağılım yerine ayrı değişkenlerin hata binom dağılımını izler ve ikili değişkeninin varyansı sabit değildir (Hair ve ark., 1998).

Lojistik regresyon, bir olayın meydana gelme olasılığını tahmin eder, bununla birlikte, tahmin edilen değeri 0 ve 1 aralığında tutmak için, bağımsız ve bağımlı değişkenler arasındaki ilişkiyi simgeleyen S şeklinde bir lojistik eğri kullanır.

2.2.13. Destek Vektör Makineleri

Destek vektör makineleri, temel olarak sınıflandırma ve regresyon temelli problemler için kullanılan bir makine öğrenme yöntemidir. Bir alternatif, n ölçüt vektörü ile temsil edilir. Burada n kriterlerin sayısını belirtir. Bu türden alternatiflerin bir kümesi $(n-1)$ boyutlu bir hiperdüzlem ile geometrik olarak sınıflara ayrılmaya çalışılır. Bu ayırıcı hiperdüzlem de doğrusal sınıflandırıcılar olarak adlandırılır. Farklı sınıfların alternatifleri uzayda yer aldığı anda, birbirini izleyen her bir sınıf arasında geometrik ayırma için kullanılabilecek birden fazla hiperdüzlem bulunur.

En büyük ayrımı sağlayan hiperdüzlem ile iyi bir ayırma sağlanır. Marjin, sınırların, alternatifleri temsil eden herhangi bir noktaya çarpmadan önce olabildiğince genişleyebildiği genişlikte tanımlanır. Bu hiperdüzlem, maksimum marjlı hiperdüzlem olarak adlandırılır ve bu hiperdüzlem tarafından tanımlanan doğrusal sınıflandırıcı, bir maksimum kenar boşluğu sınıflandırıcısı olarak adlandırılır. Maksimum marj hiperdüzlemini bulmak amacıyla destek vektörleri oluşturulur. N boyutlu uzaydaki herhangi bir nokta, kaynağı nokta ile birleştiren bir n boyutlu vektör (desen vektörü) olarak temsil edilebilir (Duman, 2010).

Destek vektör makinelerinin avantajları şunlardır:

- Yüksek boyutlu veri setlerinde etkilidir.
- Boyutların sayısının örnek sayısından daha büyük olduğu durumlarda da etkilidir.
- Karar fonksiyonunda eğitim skorlarının bir alt kümesini kullanır, bu nedenle mdeol hafızası etkin olarak kullanılır.
- Karar fonksiyonu için farklı çekirdek fonksiyonları belirlenebilir.

2.2.14. Lineer Diskriminant Analizi

Lineer Diskriminant Analizi (LDA) iyi kurulmuş ve basit sınıflandırma algoritmalarından biridir. Eğitim verisinden farklı sınıfları temsil eden verileri ayırmak için hiperdüzlemler kullanır. Yöntem, n boyutlu alanda bir yön w bulur.

Sınıflardan birine yeni bir örnek vermek için, örneğin w üzerine izdüşümü ve yeni örneklemden eğitim sınıflarının izdüşümüne olan uzaklığı hesaplanır. Amaç, her sınıftaki değişkenliği en aza indirirken, sınıflar arası değişkenliği en üst düzeye çıkarmaktır. Bu amaca ulaşmak için, yansıtılan araçlar arasındaki fark en üst düzeye çıkar ve her sınıftaki yansıtılan noktaların varyansı en aza indirgenir.

LDA yönteminin avantajı, sadeliği ve hesaplama verimliliğidir. W vektörü basit bir prosedür kullanılarak hızlı bir şekilde bulunabilir. Ayrıca, çok değişkenli verilerin boyutlarını azaltmak için güçlü bir yol sağlar (Dror, 2005).

LDA'nın en büyük dezavantajı, sınıfların araçları benzer olduğunda, düşük verimde çalışmasıdır. Girdi değişkenlerinin sayısı, uygulanacak yöntemin örnek sayısından daha küçük olmalıdır (Burget, 2004).

2.2.15. Model Performansının Topluluk Metotlarıyla Geliştirilmesi

İstatistik ve makine öğreniminde tek başına kullanılan öğrenme algoritmalarından elde edilebilecek tahmin sonuçlarından daha iyi bir tahmin sonucu elde etmek için çoklu öğrenme algoritmaları kullanır. Tek bir modele kıyasla daha güçlü ve genelleşebilir sonuçlar elde etmek amacıyla birden fazla baz modelin tahmin sonuçlarını birleştirir. Bu yöntemlerin başarısı iki ölçüte göre olur; temel öğrencilerin öğrenme başarısı ve birbirlerinden farklılıklarıdır.

Bir topluluk kendi başına denetimli bir öğrenme algoritmasıdır, algoritma eğitilebilir ve daha sonra tahminlerde bulunmak için kullanılabilir. Eğitimli topluluk tek bir hipotezi temsil eder. Bununla birlikte, bu hipotez zorunlu olarak kurulan modellerde daha fazla esnekliğe ulaşılabilir. Bu esneklik, teorik olarak, eğitim verilerinin tek bir modelden daha fazlasına uymasını sağlayabilir.

Topluluk öğrenmesi, metin kategorizasyonu (Dong ve Han, 2004), yüz tanıma (Lu et al, 2006) ve gen tanıma gibi makine öğrenme problemleriyle ilgili çeşitli alanlarda uygulanmıştır.

Wang ve diğ. (2002), kredi kartı sahtecilik veri seti üzerinde topluluk algoritması ile ilgili çalışma yapmışlardır. Çalışmalarında ağırlıklı grup sınıflandırıcıları kullanarak veri madenciliği konsepti dahilinde veri akışları için genel bir model önerilmiştir. C4.5, RIPPER, Naive Bayes gibi sınıflandırıcılar kullanılmıştır. Önerilen tek sınıflandırıcılı tahmin yöntemiyle avantaj sağlandığı kanıtlanmıştır.

Street ve Kim (2001), artırıcı ve torbalama topluluk algoritmalarının büyük verinin sınıflandırılmasındaki avantajlarını incelemişlerdir. Önerdikleri metotta ayrı sınıflandırıcılarla sezgisel bir yer değiştirme stratejisi kullanarak topluluk algoritması ile performans iyileştirme görülmüştür.

Bifet ve diğ. (2009), yaptıkları çalışmada deneysel veri akışı sistemi ile iki yeni torbalama metodu önermişlerdir. Bu yöntemler ADWIN Torbalama ve Uygulanabilir Boyutlu Hızlandırma Ağacı Torbalama' dır. Yöntemleri literatürde sıkça kullanılan veri setleriyle uygulamışlardır.

Topluluk modelleri üç önemli nedenden ötürü çoğunlukla klasik tek öğrenim modellerine göre tercih edilmektedir. Birincisi, klasik tek öğrenim modellerinde hangi sınıflandırıcının eğitim verisinde daha iyi performans gösterdiğine karar vermek için yeterli bilgi bulunmayabilir. İkincisi, uygulanan öğrenme algoritması en iyi bir hipoteze ulaşsa bile, en uygun hipotezlerle sonuçlanan kusurlu arama süreçleri uygulayabilmesidir. Üçüncüsü ise öğrenme stratejisinin, bir öğrencinin arama aşamasında gerçek bir hedef işlevine erişememesi nedeniyle iyi bir yaklaşıma yol açmasıdır.

Örüntü tanıma sistemlerinin geliştirilmesi boyunca, pek çok sınıflandırıcı, iki temel görüş altında, yani istatistiksel ve sözdizimsel olarak önerilmiştir. İstatistiksel örüntü tanıma, kalıpların olasılıksal kaynaklar tarafından üretildiğini varsayarak, modellerin istatistiksel karakterizasyonlarına dayanmaktadır. İstatistiksel model tanıma sistemleri için örnekler, Fisher Doğrusal Diskriminant, Naive Bayes Sınıflandırıcısı, ikinci dereceden sınıflandırıcılar, k -en yakın komşu, hızlandırma, karar ağaçları, Yapay Sinir Ağları, Bayes Ağları, Destek Vektör Makineleri, Gizli Markov Modelleri gibi doğrusal sınıflandırıcılardır (Duda, 2000).

Öte yandan, yapısal örüntü tanıma, yapısal ilişkilerden ve özelliklerin tutarlılığından kaynaklanmakta, özellikler, istatistiksel sınıflandırmada kullanılan basit sayısal özellik

vektörlerinden daha karmaşık ilişkiler arasındaki ilişkiyi hesaba katabilen desen yapıları sunmaktadır (Schalkoff, 1992).

Sözdizimsel örüntü tanımlamasındaki sorunlar, resmi bir dilin ya da ilişkilerin grafiklerinin dizelerinde temsil edilir.

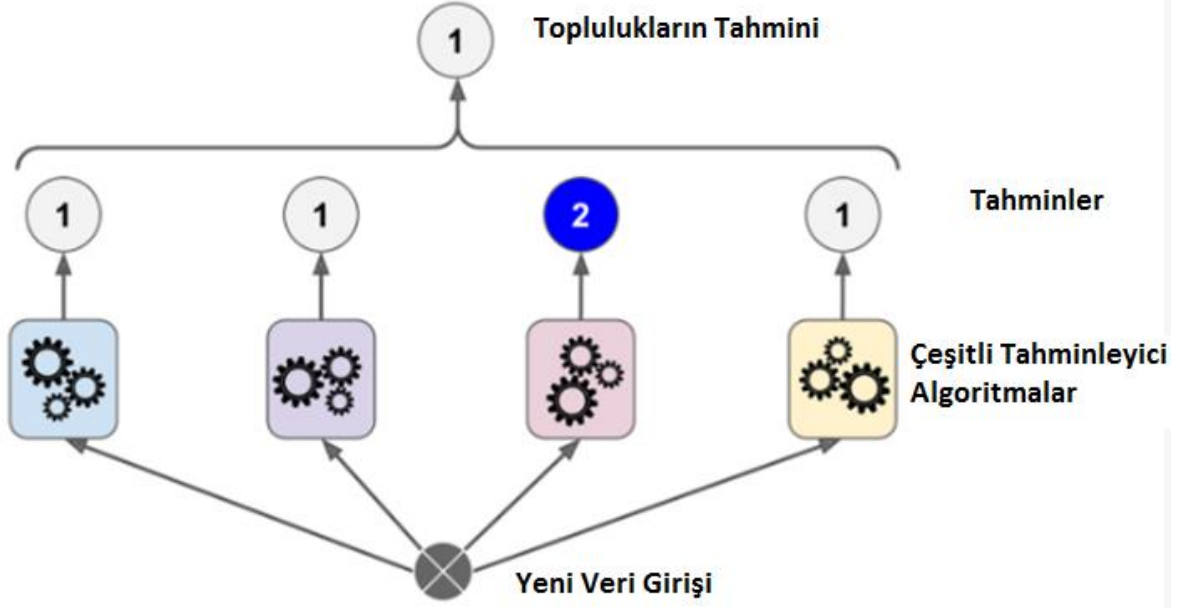
Ampirik olarak, topluluklar, modeller arasında önemli bir çeşitlilik olduğunda daha iyi sonuçlar verme eğilimindedir (Kuncheva ve diğ., 2003). Birçok topluluk yöntemi birleştirildikleri modeller arasında çeşitliliği önermektedir (Sollich ve diğ., 1996)

Topluluk metotları üçe ayrılır. Bunlar;

- Oylama (Voting)
- Torbalama (Bagging)
- Yükseltme (Boosting)

2.2.15.1. Oylama

Daha iyi bir sınıflandırıcı oluşturmanın basit bir yolu, bu sınıflandırıcılara tahminlerini sormak ve en çok oy alan sınıfı bulmaktır. Tahminleri çoklu makine öğrenimi algoritmalarında birleştirmenin en basit yollarından biridir. Eğitim veri kümesinden iki veya daha fazla bağımsız model oluşturmaya çalışır. Alt modellerin tahminleri ağırlıklandırılabilir, ancak sınıflar için ağırlıkların elle veya hatta sezgisel olarak belirlemek zordur (Brownlee, 2016). Şekil 2.9'da oylama yöntemi gösterilmiştir.

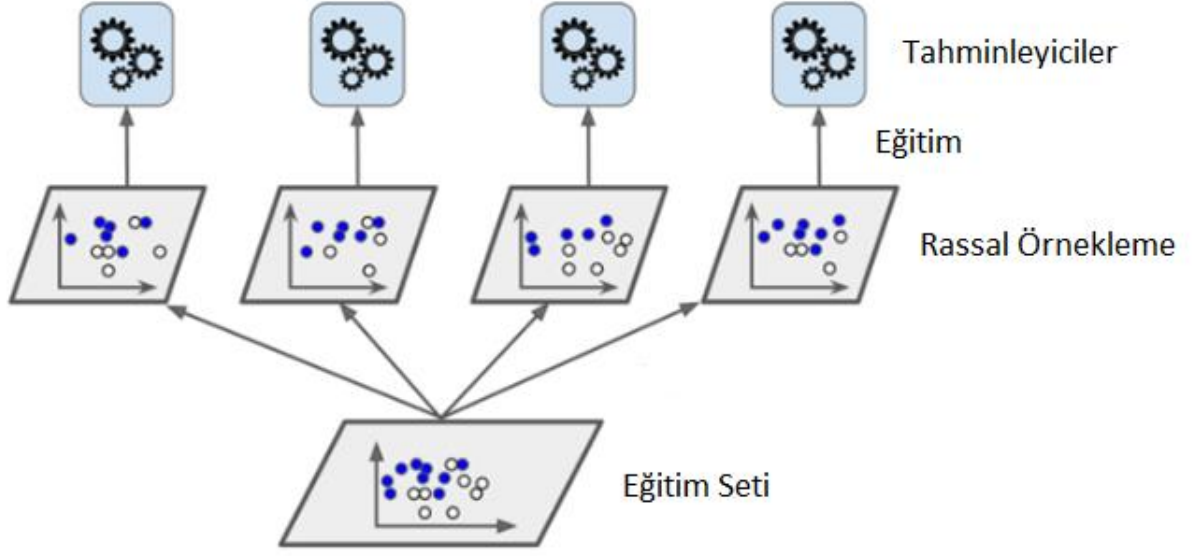


Şekil 2.9: Oylama yöntemi.

2.2.15.2. Torbalama

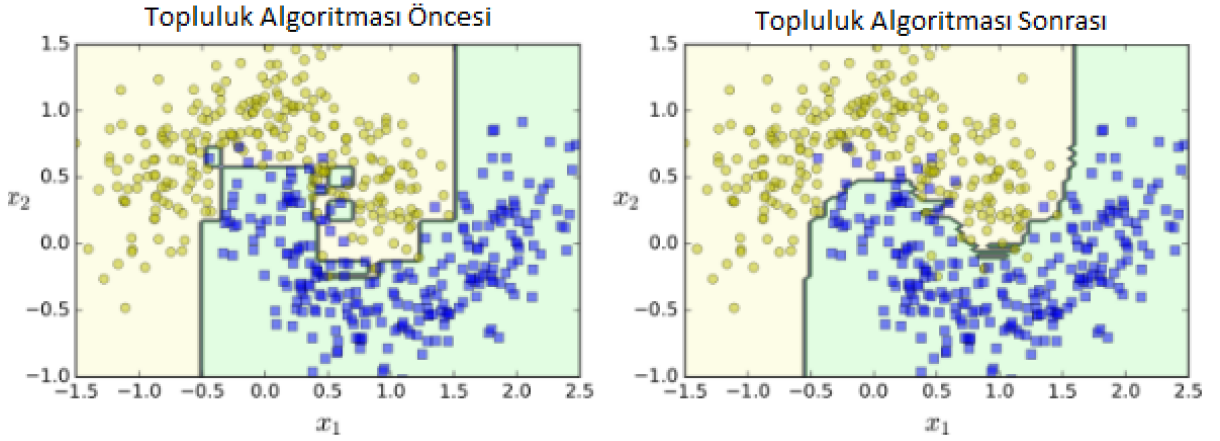
Aynı algoritmayı verinin farklı altkümeleri üzerinde çalıştırmaktır. Sabit bir kuralı kullanarak çıktıların bir araya getirilmesiyle toplanmış çoklu örneklerini üreten bir grup öğrenme yaklaşımıdır.

Bir komite içindeki farklı modeller arasındaki değişkenliği sınanarak en iyi sonucu aranır. Model varyansını teşvik etmek için, torbalama, eğitim setinin rasgele çizilmiş bir alt kümesini kullanarak topluluktaki her modeli eğitir. Altkümeleri oluştururken, örnekleme işlemi yerine koyma ile yapılırsa, bu yöntem torbalama olarak, yerine koyma ile yapılmıyor ise, bu yöntem yapııştırma olarak adlandırılır. Tüm tahmin ediciler eğitildikten sonra, tahminleri birleştirmek için, sınıflandırma probleminde sert oylama sınıflandırıcı (ya da yumuşak oylama sınıflandırıcı), regresyon probleminde ise tahminlerin ortalaması kullanılabilir. Torbalama algoritmalarında genel olarak rasgele ağaç algoritması, ekstra ağaçlar, torbalanmış karar ağaçları algoritmaları kullanılır. Şekil 2.10'da torbalama yöntemi gösterilmiştir.



Şekil 2.10: Torbalama yöntemi.

Topluluk algoritmasının uygulanması öncesi ve sonrasını gösteren örnek model görüntüsü Şekil 2.11’ de gösterilmiştir.



Şekil 2.11: Topluluk algoritması uygulanmadan önce ve sonra model gösterimi.

2.2.15.3. Artırma

Toralama yönteminin farklı bir versiyonudur. Fark; öğrenme sonuçlarının sonraki öğrenici için kullanılıyor olmasıdır. Önceki modellerin yanlış sınıflandırıldığı eğitim örneklerini vurgulamak için her yeni model örneğini eğiterek bir topluluk oluşturmayı içerir. Artan topluluk algoritmaları, sıradaki modellerin hatalarını düzeltmeye çalışan bir dizi model oluşturur. Diğer yöntemlere göre daha yaygındır, hızlı çalışır az bellek kullanır. Eğitim için ayrılan veri setinden

bir temel öğrenici için rasgele seçim yapılır. Öğrenme gerçekleşir, model test edilir. Sonuçlardan yanlış sınıflandırılan örnekler belirlenir. Bunlar bir sonraki öğrenici için örnek seçiminde önceliklendirilir. Her seferinde bu bilgi güncellenir. Torbalama yönteminde her bir örneğin seçilme şansı ettir.

Bugüne kadar, artırma algoritmaları içerisinde en yaygın uygulaması Adaboost'dur, ancak daha iyi sonuçlar elde etmek için bazı yeni algoritmalar bildirilmiştir (Brownlee, 2016).



3. MALZEME VE YÖNTEM

3.1. PROBLEMİN TANIMLANMASI

Bankacılık sektörü, tele pazarlamaya yönelik hedeflemeyi optimize etmek, karlılığı artırmak ve maliyetleri düşürmek konusunda sürekli baskı altındadır. Bankalar için pazarlama kampanyalarının önemi gün geçtikçe artmaktadır. Çalışma kapsamında uygulanacak pazarlama kampanyası ile arama yapılan müşteriye vadeli mevduat hesabı açarak bankadaki likiditeyi arttırmak, bankaya çapraz satış fırsatları yaratmak ve müşteri ilişkilerini geliştirmek hedeflenmiştir. Doğrudan yapılan pazarlama aramaları müşteri memnuniyetsizliği oluşturabilir. Arama yapılacak müşterilerin doğru olarak sınıflandırılması, doğru müşteri grubuna ulaşarak müşteri memnuniyetsizliğini azaltacak ve arama maliyetlerini düşürecektir. Ayrıca kaynakların verimli kullanımından doğan ek bütçe, çalışanların eğitimi ya da teknoloji yatırımı gibi banka süreçlerini iyileştirecek alanlarda ek yatırım yapma imkanı sağlanacaktır.

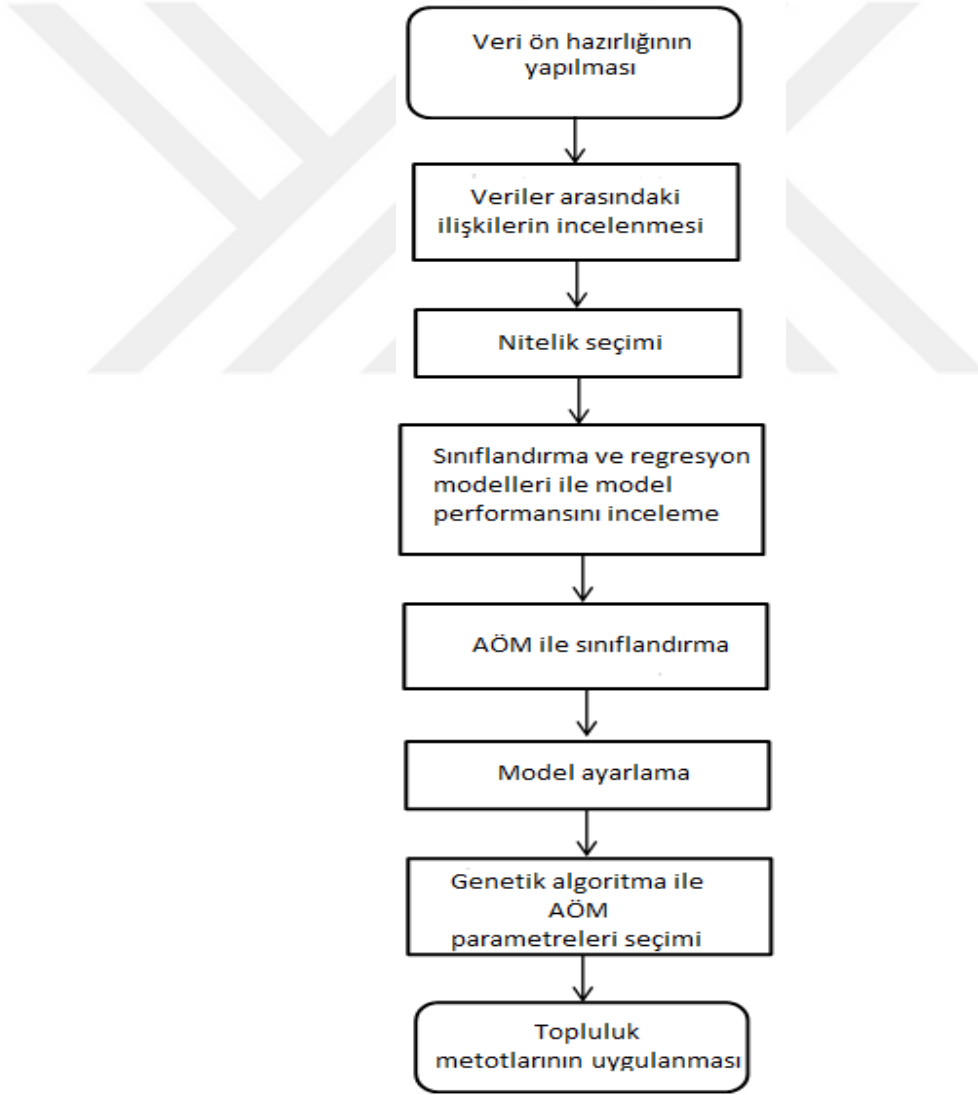
Veri setinde genel olarak farklı mevsim, iş gücü, müşteri profilleri vs. gibi özelliklere göre müşterilerin kampanyaya katılımı yani vadeli mevduat ürünü açtırması veya kampanyaya katılmama durumları yer almaktadır.

Bu çalışmada bankanın geçmiş pazarlama verileri ile yapılan analizlerde veriyi sınıflandırma için AÖM sınıflandırıcısı önerilmiş, vadeli mevduat açılışı kampanyasına katılabilecek müşteriler ile yapılan telefon görüşmeleri sınıflandırılmıştır. Özellik seçimi öncesinde ve sonrasında ki özelliklerle diğer sınıflandırma algoritmalarının performansları karşılaştırılmıştır. Önerilen yöntemin parametreleri genetik algoritma ile optimize edilerek sınıflandırma için en doğru tahminleme yapılmıştır. En doğru tahminleme için topluluk algoritmaları kullanılmıştır.

Veri analizi sürecinde sırasıyla aşağıdaki işlemler yapılmıştır (Şekil 3.1):

- Analiz öncesi veri ön hazırlığı,
- Veriler arasındaki ilişkilerin incelenmesi,
- Nitelik seçimi,
- Nitelik seçimi öncesinde farklı sınıflandırma algoritmalarıyla varsayılan parametre değerleri ile analizi,

- Nitelik seçimi sonrasında farklı sınıflandırma algoritmalarıyla varsayılan parametre değerleri ile analizi,
- Nitelik seçimi öncesi ve sonrasında sınıflandırma performanslarının karşılaştırılması,
- Nitelik seçimi sonrasında Aşırı Öğrenme Makinesi (AÖM) ile sınıflandırma sonucunda ulaşılan model değerlendirme ölçütlerinin incelenmesi,
- Model ayarlama teknikleriyle model performansı için en uygun parametrelerin bulunması,
- Genetik algoritma ile en iyi parametrelerin tespiti,
- En iyi parametrelerle çalıştırılan modelin topluluk algoritmaları ile iyileştirilmesi.



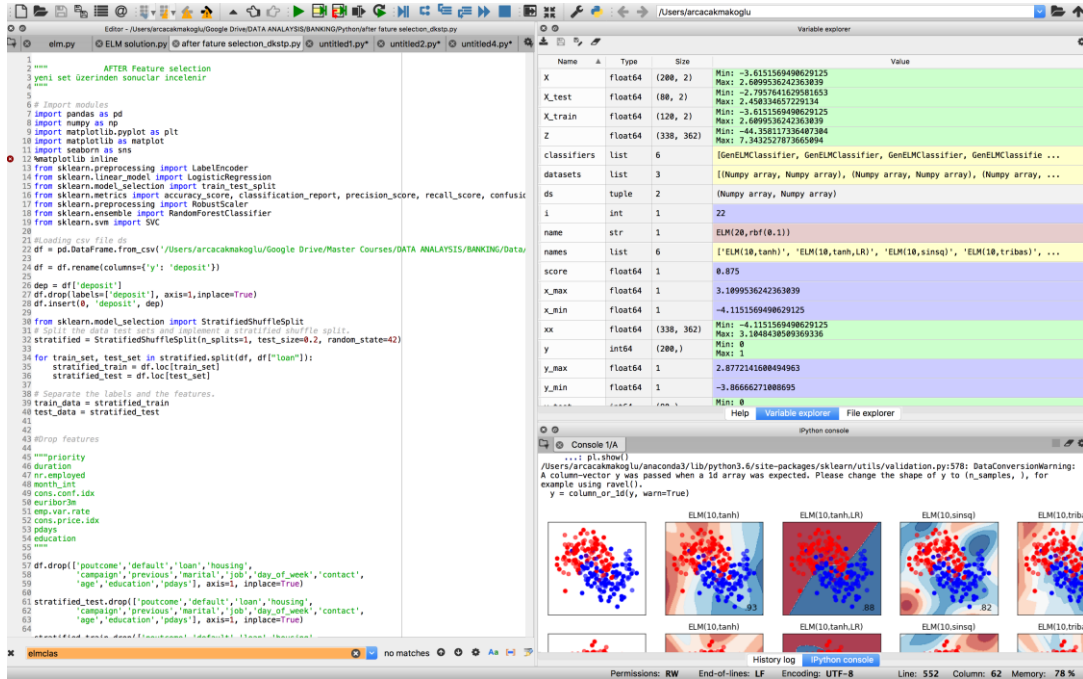
Şekil 3.1: Veri analizi süreci akış şeması.

3.2. KULLANILAN YAZILIM GELİŞTİRME ORTAMI

Python, yorumlanmış, etkileşimli, nesne yönelimli bir programlama dilidir. Python’ da modüller, istisnalar, dinamik yazım, üst düzey dinamik veri türleri ve veri sınıfları bulunmaktadır. Birçok sisteme entegre edilebilir ve içerisinde birçok kütüphane bulundurulur. Bazı Windows uygulamalarında, C veya C++ ile genişletilebilir. Programlanabilir uygulamalar için ön yüz ihtiyacını karşılama işlemleri için kullanılabilir. Söz dizimi yazımı ile hızlı yazılım üretmeyi destekleyen bir kod yapısı vardır.

Unix, Linux, Symbian, Mac, Windows, Amiga platformlarını destekler. Python ile sistem yazılımı, kullanıcı arabirimi yazılımı, ağ yazılımı, uygulama ve veri tabanı yazılımı gibi birçok alanda yazılım geliştirilebilir. C ya da C++ gibi dillere göre büyük yazılım projelerinde hızlı prototip ihtiyacı olduğunda avantaj sağlar.

Guido van Rossum 1990 yılında Amsterdam’da bu uygulamayı geliştirmeye başlamıştır. Python açık kaynaklı bir yazılım dilidir. Python Yazılım Vakfı tarafından toplanan gönüllülerin gönüllüler ile geliştirilmeye devam edilmektedir. Python 1.0 sürümü Ocak 1994’te çıkmıştır. En güncel versiyonu 3.6.5’ tir. Şekil 3.2’de programa ait bir ekran görüntüsü verilmiştir.



Şekil 3.2: Programa ait ekran görüntüsü (bankacılık verisi ile).

3.3. VERİNİN ANLAŞILMASI

Veri seti, Kaliforniya Irvine Üniversitesi (UCI) Makine Öğrenimi Veri Kütüphanesi web sitesinden alınmıştır. UCI makine öğrenme uygulamaları, veri analizi üzerine çalışan akademisyenler ve profesyoneller tarafından kullanılan bir veri tabanı kütüphanesidir. UCI web sitesindeki “Bank Marketing Data Set” dosyaları içerisinde bulunan bank-additional-full.csv dosyası ile çalışma yapılmıştır (Moro, 2014) .

Uygulamada kullanılan veriler, bir Portekiz Bankası’nın doğrudan pazarlama kampanyası için toplanmıştır. Veri seti pazarlama kampanyaları için yapılan telefon görüşmeleri ile ilgili detayları içermektedir. Veri detayında müşteriler ile ilgili demografik bilgiler (yaş, meslek, cinsiyet ...), bankacılık bilgileri (ihtiyaç kredisi, ev kredisi, ...), finans metrikleri (eurobor, aylık fiyat endeksi,...) ve önceki kampanyalara dair bilgiler yer almaktadır. Toplamda 41.188 adet arama ile ilgili veri kaydedilmiştir. Bu veriler Mayıs 2008’den Kasım 2010’a kadar toplanmıştır. Veri setinde 20 farklı nitelik ve bir hedef verisi bulunmaktadır. Hedef verisi müşterilerin kampanyaya katılıp katılmadığı bilgisidir. Kampanya kapsamında aynı müşteriye birden çok arama yapılmıştır. Müşterilere iyi faiz getirisi olan uzun vadeli mevduat hesabı önerilmiştir. Veri içerisinde yer alan nitelikler ve niteliklerin özellikleri Tablo 3.1’de belirtilmiştir. Şekil 3.3.’ te veri setinin programdaki görüntüsüne yer verilmiştir.

Tablo 3.1: Ön işleme öncesi veri.

Nitelikler	Veri Tipi	Açıklama
Yaş	int64	20,21,...
Meslek	object	Yönetici, bilinmiyor, işsiz, yönetim, müstahdem, girişimci, öğrenci, mavi yaka, serbest meslek sahibi, emekli, teknisyen, hizmet sektöründe çalışan.
Medeni Durum	object	Boşanmış, evli, bekar, bilinmiyor.
Eğitim Durumu	object	İlkokul mezunu, ortaokul mezunu, lise mezunu, okumamış, profesyonel kurs almış, üniversite mezunu, bilinmeyen.
Mevcut Kredi Bilgisi	object	Evet, Hayır, Bilinmiyor
Ev Kredisi	object	Evet, Hayır, Bilinmiyor
İhtiyaç Kredisi	object	Evet, Hayır, Bilinmiyor
İletişim Yöntemi	object	Cep telefonu, Sabit Telefon
Ay	object	Ocak, Şubat, ..., Aralık
Gün	object	Pazartesi, Salı, Çarşamba, Perşembe, Cuma

Tablo 3.1 (devam): Ön işleme öncesi veri.

Nitelikler	Veri Tipi	Açıklama
Görüşme Süresi	int64	1,2, ... saniye
Kampanya Sayısı	int64	Son arama dahil olmak üzere kampanya amacıyla yapılan arama sayısı
Arama Sıklığı	int64	Son iki arama arasında geçen gün sayısı: Daha önce arama yapılmamışsa 999.
Toplam Arama Sayısı	int64	Bugüne kadar yapılan arama sayısı
Önceki Kampanya Sonucu	object	Önceki aramanın kampanya sonucu: Başarısız, Başarılı, Bilinmiyor.
İstihdam Oranı	float64	3 aylık oran
Tüketici Fiyat Endeksi	float64	Aylık
Tüketici Güven Endeksi	float64	Aylık
Eurobor Endeksi	float64	3 aylık EUROBOR endeksi Avrupa bankalar arası faiz oranı, Avrupa Birliği bankalar arası piyasasında bankalar arasında borç verme için kullanılan temel faiz oranı ve aynı zamanda diğer krediler üzerinde faiz oranının belirlenmesi için referans olarak kullanılmıştır.
İşsizlik Oranı	float64	3 aylık
Kampanyaya Katılım	object	Evet, Hayır

Spyder (Python 3.6)						df - DataFrame		
Index	age	job	marital	education	default	housing	loan	contact
0	56	housemaid	married	basic.4y	no	no	no	telephone
1	57	services	married	high.school	unknown	no	no	telephone
2	37	services	married	high.school	no	yes	no	telephone
3	40	admin.	married	basic.6y	no	no	no	telephone
4	56	services	married	high.school	no	no	yes	telephone
5	45	services	married	basic.9y	unknown	no	no	telephone

Şekil 3.3: Veri setinin programda gösterimi.

month	day_of_week	duration	campaign	pdays	previous	poutcome
may	mon	261	1	999	0	nonexistent
may	mon	149	1	999	0	nonexistent
may	mon	226	1	999	0	nonexistent
may	mon	151	1	999	0	nonexistent
may	mon	307	1	999	0	nonexistent
may	mon	198	1	999	0	nonexistent

Şekil 3.3 (devam): Veri setinin programda gösterimi.

emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
1.100000000	93.994000000	-36.4000000...	4.857000000	5191.000000...
1.100000000	93.994000000	-36.4000000...	4.857000000	5191.000000...
1.100000000	93.994000000	-36.4000000...	4.857000000	5191.000000...
1.100000000	93.994000000	-36.4000000...	4.857000000	5191.000000...
1.100000000	93.994000000	-36.4000000...	4.857000000	5191.000000...
1.100000000	93.994000000	-36.4000000...	4.857000000	5191.000000...

Şekil 3.3 (devam): Veri setinin programda gösterimi.

3.4. VERİ İŞLEME

Veri setinin orijinal hali analiz öncesinde kontrol edilmiştir. Veri tiplerine göre gerekli ön işlemlerden geçirilmiş ve analiz için hazır hale getirilmiştir. Veri seti içerisinde kayıp değerler olmadığından veri türetme ihtiyacı olmamıştır. Veri setin ön işlemleri için sırasıyla aşağıdaki işlemler yapılmıştır.

- Öncelikle karar niteliği belirlenmiştir. Karar sütunu veri setinden ayrılmıştır. Kampanya sonucunu ifade eden karar niteliği, müşterinin vadeli mevduata üye olma durumunu ifade eder. Programda kullanılmak üzere sütun deposit olarak adlandırılır.
- Veri seti içerisinde evet/hayır (boolean) formatında olan veriler 0-1 formatında (binary) değişkenlere dönüştürülür. Mevcut kredi (default), ev kredisi (housing), ihtiyaç kredisi (loan) ve vadeli mevduata üyelik durumu (deposit) nitelikleri için bu işlem yapılır.
- Ay (month) ve gün (day) kolonları numerik veriye (Int_32) dönüştürülür.
- Konuşma süresi (duration) daha doğru karşılaştırmalar yapılabilmesi için dakikaya çevrilir.

Ön işlemler tamamlandıktan sonraki veri tipi özellikleri Tablo 3.2’de gösterilmiştir.

Veri ön işleme sonrasında veri seti için özetler, Tablo 3.3’de gösterilmiştir. Özet tabloda numerik veriler için sırasıyla ilgili niteliğin sayısı, ortalaması, standart sapması, minimum ve maksimum değerleri ile 1. ve 3. çeyrek değerlerine gelen değer ve medyan gösterilmiştir.

Tablo 3.2: Ön işlem sonrası veri seti.

Nitelikler	Veri Tipi	Açıklama
Yaş	int64	20,21,...
Meslek	object	Yönetici, bilinmiyor, işsiz, yönetim, müstahdem, girişimci, öğrenci, mavi yaka, serbest meslek sahibi, emekli, teknisyen, hizmet sektöründe çalışan.
Medeni Durum	object	Boşanmış, evli, bekar, bilinmiyor.
Eğitim Durumu	object	İlkokul mezunu, ortaokul mezunu, lise mezunu, okumamış, profesyonel kurs almış, üniversite mezunu, bilinmeyen.
Mevcut Kredi Bilgisi_int	int64	Evet, Hayır, Bilinmiyor
Ev Kredisi_int	int64	Evet, Hayır, Bilinmiyor
İhtiyaç Kredisi_int	int64	Evet, Hayır, Bilinmiyor
İletişim Yöntemi	object	Cep telefonu, Sabit Telefon
Ay_int	int64	Ocak, Şubat, ..., Aralık
Gün_int	int64	Pazartesi, Salı, Çarşamba, Perşembe, Cuma
Görüşme Süresi	int64	1,2, ... dakika
Kampanya Sayısı	int64	Son arama dahil olmak üzere kampanya amacıyla yapılan arama sayısı
Arama Sıklığı	int64	Son iki arama arasında geçen gün sayısı: Daha önce arama yapılmamışsa 999.
Toplam Arama Sayısı	int64	Bugüne kadar yapılan arama sayısı
Önceki Kampanya Sonucu	object	Önceki aramanın kampanya sonucu: Başarısız, Başarılı, Bilinmiyor.
İstihdam Oranı	float64	3 aylık oran
Tüketici Fiyat Endeksi	float64	Aylık
Tüketici Güven Endeksi	float64	Aylık
Eurobor Endeksi	float64	3 aylık EUROBOR endeksi Avrupa bankalar arası faiz oranı, Avrupa Birliği bankalar arası piyasasında bankalar arasında borç verme için kullanılan temel faiz oranı ve aynı zamanda diğer krediler üzerinde faiz oranının belirlenmesi için referans olarak kullanılmıştır.
İşsizlik Oranı	float64	3 aylık
Kampanyaya Katılım	object	Evet, Hayır

Tablo 3.3: Veri seti özeti (Nitelik seçimi öncesi).

Nitelikler	Önceki Kampanya Sonucu	Arama Sıklığı	Kampanya Sayısı	Konuşma Süresi (dk)	Yaş	Kampanyaya Katılım	Gün	Ay	Mevcut Kredi	İhtiyaç Kredisi	Ev Kredisi	Eurobor Endeksi	İstihdam Oranı	Tüketici Güven Endeksi	İşsizlik Oranı
Ortalama	0,172	962,475	2,567	4,304	40,02	0,112	2,98	6,607	0,208	0,175	0,547	3,621	-40,502	93,575	0,081
Standart Sapma	0,494	186,91	2,77	4,321	10,42	0,316	1,41	2,04	0,406	0,38	0,497	1,734	4,628	0,578	1,57
Minimum değer	0	0	1	0	17	0	1	3	0	0	0	0,634	-50,8	92,201	-3,4
1.Çeyrekte ki Veri	0	999	1	1,7	32	0	2	5	0	0	0	1,344	-42,7	93,075	-1,8
Medyan	0	999	2	3	38	0	3	6	0	0	1	4,857	-41,8	93,749	1,1
3.Çeyrekte ki Veri	0	999	3	5,32	47	0	4	8	0	0	1	4,961	-36,4	93,994	1,4
Maksimum Değer	7	999	56	81,97	98	1	5	12	1	1	1	5,045	-26,9	94,767	1,4

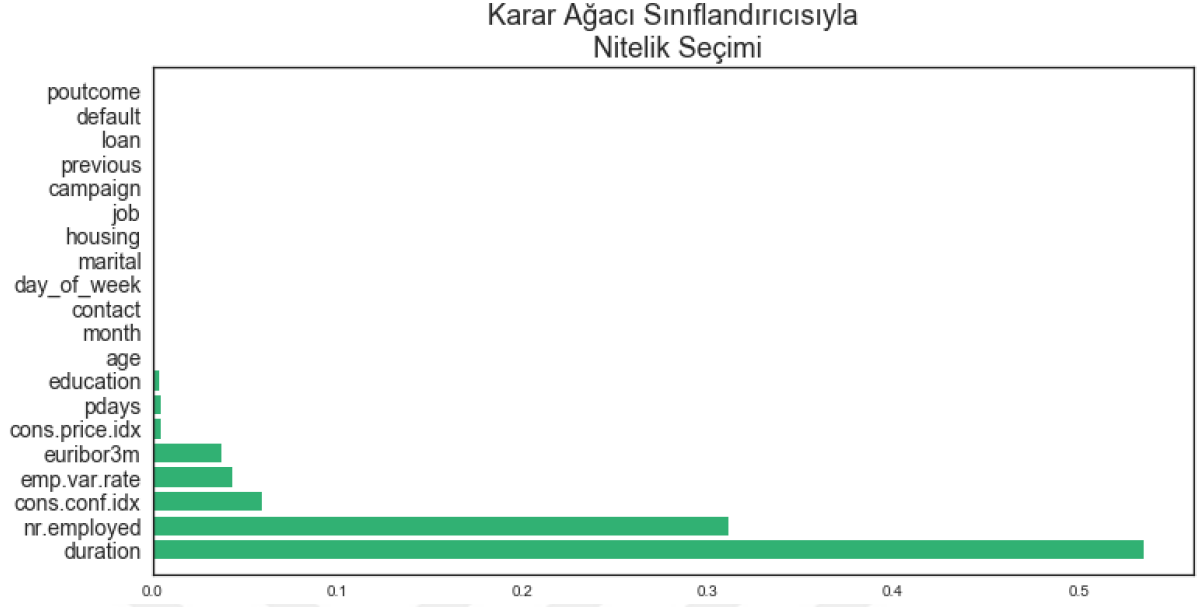
3.4.1. Nitelik Seçimi

Nitelik seçimi, mümkün olduğunca küçük bir özellik alt kümesini belirleyerek modelin aşırı öğrenmesini engellemek ve daha hızlı, hesaplama maliyeti açısından daha uygun modeller sunmayı sağlamaktadır. Bu çalışmada nitelik seçimi öncesinde ve nitelik seçimi sonrasında sınıflandırma algoritmaları çalıştırılarak değişim durumları incelenmiştir.

Nitelik seçimi için Python scikitlearn kütüphanesinde yer alan DecisionTreeClassifier kullanılmıştır. Kararı etkileyebilecek niteliklerin ilişki oranına göre niteliklerin önem derecesi sıralanmıştır. Python’ da ilgili karar ağacı algoritması CART’ dir (Classification and Regression Tree). Şekil 3.4’ te ve Tablo 3.4’te niteliklerin ilişki oranları gösterilmiştir.

Sonuçlara göre, müşterilerin kampanyaya üye olması için en önemli nitelikler sırasıyla konuşma süresi, çalışan sayısı, tüketici güven endeksi, istihdam oranı, eurobor endeksi, tüketici fiyat endeksi, önceki kampanya sonrasında arama yapılana kadarki süre, eğitim derecesi, yaş, ay, arama kanalı, gün bilgisi, medeni durum, ev kredisi olup olmadığı ve meslektir.

Nitelik seçimi sonuçlarına göre kampanyaya katılımı etkilemeyen ihtiyaç kredisi (loan), kredi (default) nitelikleri, önceki kampanya başarısı (poutcome) ve aranan müşteri dahil daha önce yapılan arama sayısı (previous) önem derecesi sıfır çıktığından veri setinden çıkarılmıştır. Bu çalışmada varsayılan parametreleriyle çalıştırılan modeller nitelik seçimi öncesindeki veri seti ve nitelik seçimi sonrasındaki veri setiyle çalıştırılmıştır. Nitelik seçimi sonrasında daha doğru sınıflandırma sonuçlarına ulaşıldığından, AÖM ile yapılan çalışma nitelik seçimi sonrasında ulaşılan veri seti ile yapılmıştır.



Şekil 3.4: Nitelik seçimi.

Tablo 3.2: Nitelik ilişkileri.

Nitelik Numarası	Önem Derecesi (*1.00)	Nitelik Numarası	Önem Derecesi (*1.00)
duration	0,535468	contact	0,000143
nr.employed	0,311355	day_of_week	0,000086
cons.conf.idx	0,058994	marital	0,000019
emp.var.rate	0,043389	housing	0,000019
euribor3m	0,037382	job	0,000015
cons.price.idx	0,004288	campaign	0,000000
pdays	0,004242	previous	0,000000
education	0,004009	loan	0,000000
age	0,000385	default	0,000000
month	0,000205	poutcome	0,000000

3.5. MODELİN KURULMASI

Bu çalışmada birden çok sınıflandırma algoritması, kampanyaya katılım sonucunu sınıflandırmak için kullanılmıştır. Ayrıca nitelik seçimi ve modeli geliştirme evrelerinde birçok algoritmadan faydalanılmıştır. Model nitelik seçimi öncesi ve sonrası olmak üzere iki veri seti üzerinde çalıştırılmıştır. Kullanılan algoritmalar ve Python kütüphaneleri Tablo 3.5'te gösterilmiştir.

Tablo 3.5: Python’ da kullanılan algoritmalar.

Kullanılan Fonksiyonlarının	Algoritmalar /	Kullanılan Kütüphane / Paket	Kullanım Amacı
StratifiesShuffleSplit		sklearn.model_selection	Veri Ayırma
CategoricalEncoder		sklearn.base	Veri Ön Hazırlığı
NumericalEncoder		sklearn.preprocessing	Veri Ön Hazırlığı
LabelEncoder		sklearn.preprocessing	Veri Ön Hazırlığı
StandartScaler		sklearn.preprocessing	Veri Ön Hazırlığı
Gradient Boosting Classifier		sklearn.ensemble	Sınıflandırma Algoritması
Neural Classifier		sklearn.neuralnetwork	Sınıflandırma Algoritması
Logistic Regression		sklearn.linear_model	Sınıflandırma Algoritması
Random Forest		sklearn.ensemble	Sınıflandırma Algoritması
Support Vector Machine		sklearn.svc	Sınıflandırma Algoritması
K Nearest Neighbors		sklearn.neighbors	Sınıflandırma Algoritması
G Extreme Learning Machine		sklearn.extreme_learning_machines.elm	Sınıflandırma Algoritması
Extreme Learning Machine		sklearn.extreme_learning_machines.elm	Sınıflandırma Algoritması
Decision Tree (CART)		sklearn.tree	Sınıflandırma Algoritması / Nitelik Seçimi
Linear Discriminant Analysis		sklearn.discriminant_analysis	Sınıflandırma Algoritması
Naives Bayes Classifier		sklearn.naive_bayes	Sınıflandırma Algoritması
BaggingClassifier		Sklearn.ensembles	Model Geliştirme /
VotingClassifier			Sınıflandırma Algoritmaları
AdaboostClassifier			

3.6. MODEL PERFORMANS DEĞERLENDİRME YÖNTEMLERİ

Modeller, nitelik önceliklendirmesi öncesinde ve sonrasında belirtilen sınıflandırıcılar için varsayılan parametreler ile çalıştırılmış, model performans ölçümleri yapılmıştır. Tezin esas konusunu içeren AÖM sınıflandırıcısı ile model ayarlama işlemleri yapıp farklı parametrelere göre çalıştırılmış ve performans ölçümleri yapılmıştır. Bulgular bölümünde her bir modelin performans ölçümleri açıklanmıştır. Analizlerde;

- Eğitim ve test verileri çapraz geçерleme yöntemi ile 10 parçaya bölünerek ayrılmıştır.
- Karmaşıklık matrisi, duyarlılık derecesi, F1- skor değeri, doğruluk oranı ve destek değeriğine göre performans ölçümleri karşılaştırılmıştır.
- Izgara arama ve genetik algoritma ile parametre optimizasyonu model ayarlama yöntemleri kullanılarak en iyi sınıflandırmanın yapılabilmesi için kullanılması gereken AÖM sınıflandırıcısı parametreleri belirlenmiştir.
- Topluluk öğrenmesi algoritması uygulanıp model geliştirilmiştir. Topluluk algoritması olarak oylama, torbalama ve artırıcı algoritmalar kullanılmıştır.

3.7. MODELİN UYGULANMASI VE YORUMLANMASI

Veri seti üzerinde farklı sınıflandırma algoritmaları ile analizler yapılmış, AÖM sınıflandırıcısı parametreleri genetik algoritma yöntemiyle optimize edilirken iyi parametrelerin seçimi sağlanmıştır. Ulaşılan model üzerinde topluluk algoritmaları çalıştırılarak en iyi sınıflandırma doğruluğuna ulaşılması hedeflenmiştir. Sınıflandırma sonuçları ve bu sonuçlara göre yapılan yorumlar dördüncü bölümde yer almaktadır.

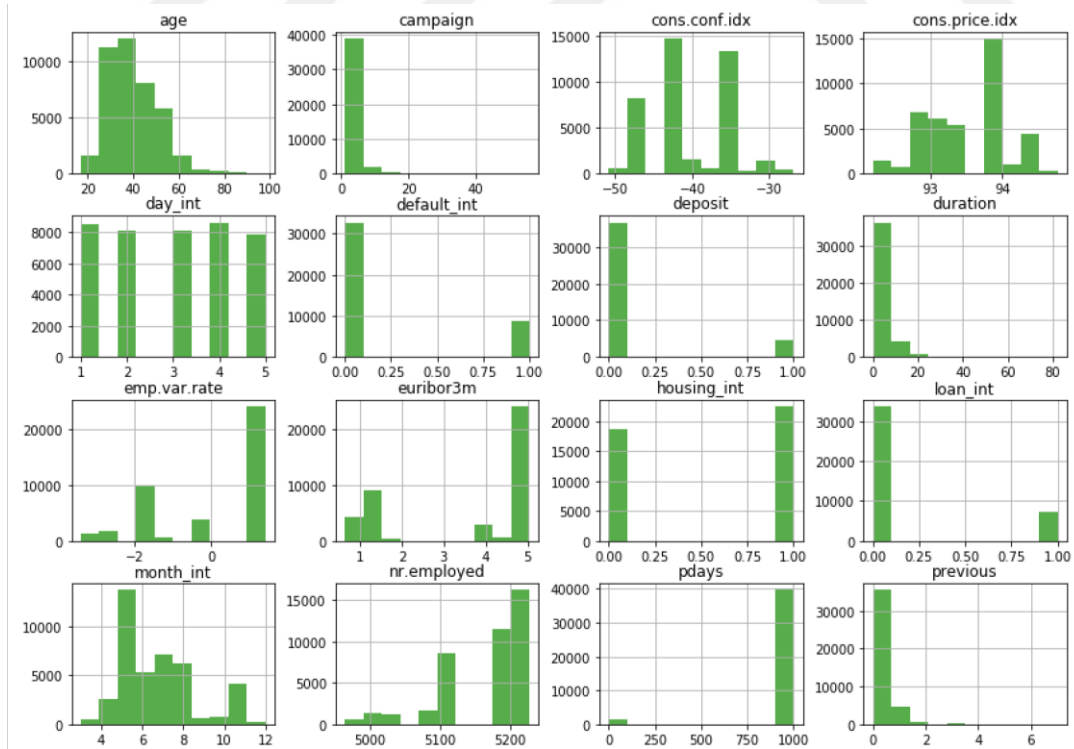


4. BULGULAR

Bu bölümde veri setiyle ilgili temel istatistiksel bilgiler verilmiş ardından veri madenciliği analizlerinin sonuçları incelenmiştir. Veri içerisinde yer alan bazı nitelikler ile kampanyaya katılım oranları karşılaştırılarak yorumlar yapılmıştır. Veriler arasındaki ilişkiler yorumlandıktan sonra çeşitli sınıflandırıcılarla analiz yapılmıştır. Sonrasında Aşırı Öğrenme Makinesi (AÖM) sınıflandırıcısıyla analizler yapılmıştır.

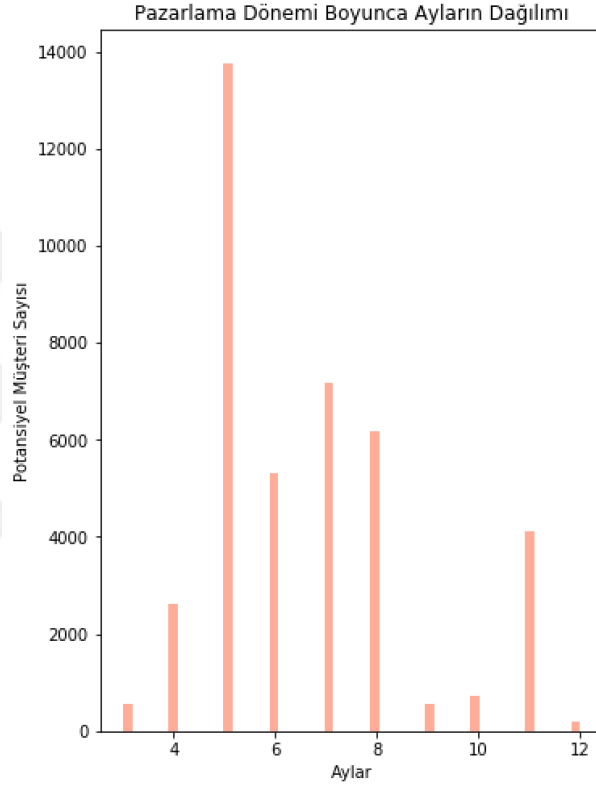
4.1. VERİLER ARASINDAKİ İLİŞKİLERİN İNCELENMESİ

Nümerik olan her bir niteliğin tüm veri içerisindeki dağılımlarının histogram tablosunda gösterimi Şekil 4.1’ dendir. Burada sırasıyla yaş aralıklarının, önceki kampanya arama sayılarının, tüketici güven endeksi, tüketici fiyat endeksi, aramaların yapıldığı günlerin, mevcut kredilerin, daha önce yapılan arama sayısı, konuşma süresinin, istihdam oranının, eurobor endeksinin, kredilerin, ayların ve daha önce yapılan arama sonrası geçen zamanın dağılımları görülmektedir.

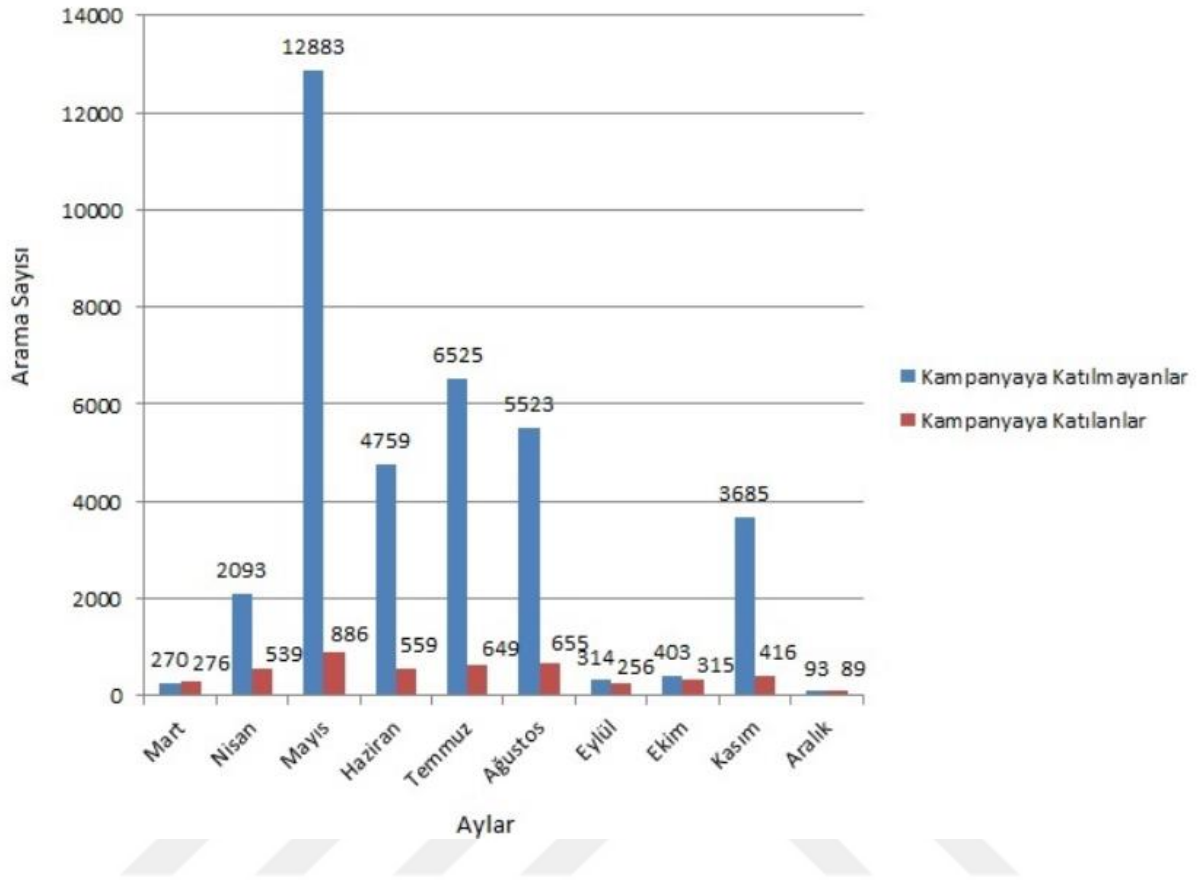


Şekil 4.1: Veri dağılımlarının incelenmesi.

Şekil 4.2’de görüldüğü üzere pazarlama etkinliği en yüksek olan ay Mayıs ayıdır. Ocak, Eylül, Ekim ve Aralık aylarında ise arama sayıları daha düşüktür. Şekil 4.3’de kampanyaya katılım oranlarının aylar üzerindeki etkisi incelenmiştir. Kampanyaya katılan ve katılmayan müşterilerin oranlarının en yüksek olduğu ay Mayıs aylarıdır. Buna karşın Ocak, Şubat, Mart, Nisan, Eylül, Ekim ve Aralık aylarında ki kampanyaya katılım oranı, katılmama oranlarından daha yüksektir. Bu ayların kampanya açısından daha verimli aylar olduğu görülmektedir.

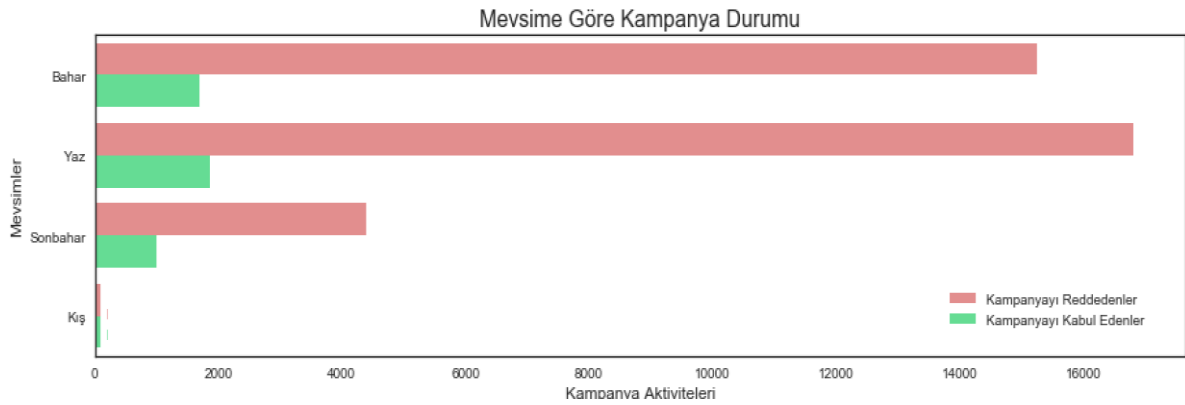


Şekil 4.2: Kampanyaya katılımın aylara göre dağılımı.



Şekil 4.3: Aylara göre kampanya başarısı.

Potansiyel müşteriler vadeli mevduatları için sonbahar ve kış mevsimlerini tercih etmektedir (Şekil 4.4). Bir sonraki pazarlama kampanyası faaliyetlerinde bu mevsimlere göre yoğunlaşılabilir. Ayrıca günlerin kampanya üzerindeki etkisi incelendiğinde belirgin bir farklılık olmamakla birlikte kampanya için en iyi günün Perşembe günü olduğu görülmektedir.



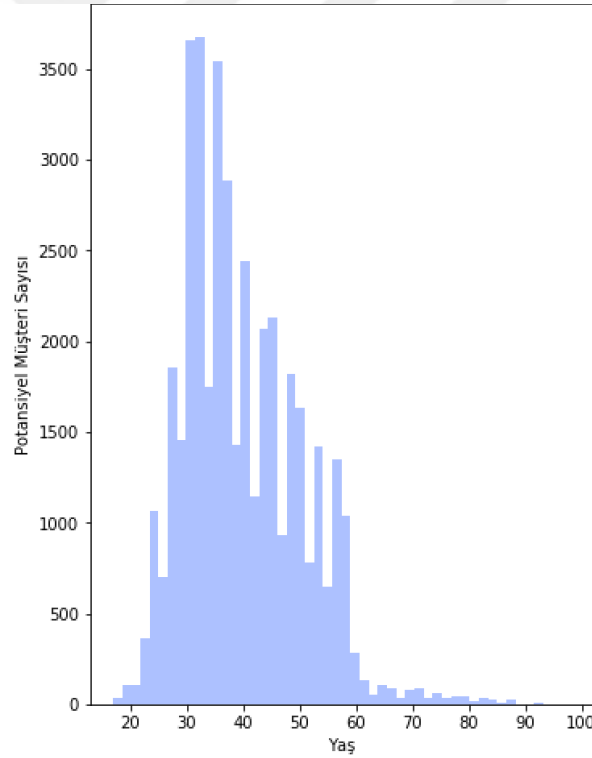
Şekil 4.4: Mevsimlere göre kampanya başarısı.

Veri seti içerisindeki müşterilerin yaş dağılımları Tablo 4.1’de gösterilmiştir. Veri setinde bulunan müşteriler genel olarak 30-45 yaş aralığındadır. Müşteri yaş aralıklarına göre kampanyaya katılım incelendiğinde en ideal yaşların, altmışlı yaşlar olduğu görülmüştür. En düşük katılım oranı kırk yaşlarındaki müşterilerde görülmüştür. Kampanyaya katılım durumları karşılaştırıldığında altmışlı yaşlardan sonra sırasıyla yirmili yaşlar, ellili yaşlar ve otuzlu yaşlar gelmektedir.

Tablo 4.1: Yaş kategorisine göre kampanya katılım oranları.

Yaş Kategorisi	20	30	40	50	60
Kampanyaya Katılmayanlar	%83,73	%89,87	%92,07	%89,84	%60,43
Kampanyaya Katılanlar	%16,26	%10,12	%7,92	%10,15	%39,56

Yaşlara göre müşteri sayısı dağılımı Şekil 4.5’de gösterilmiştir.

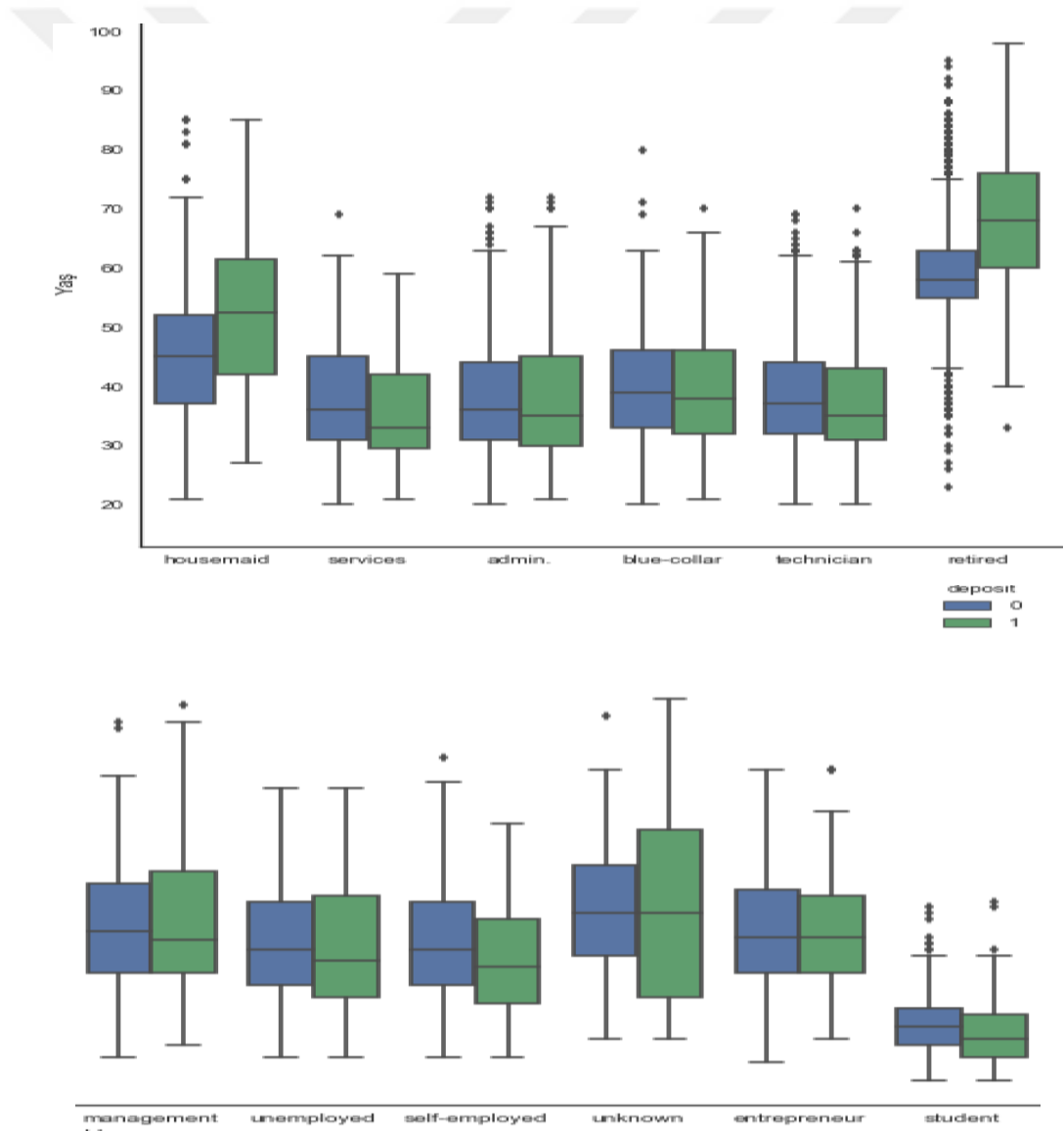


Şekil 4.5: Yaşlara göre müşteri sayısı dağılımı.

Arama yapılan kişilerin meslek grupları ve yaşları arasındaki ilişki Şekil 4.6’de kutu grafiği üzerinde gösterilmiştir. Kutu grafiğinde kampanyaya katılan ve katılmayan müşteriler, mesleklerine göre ayrılmış ve bu mesleklerin hangi yaş aralığında dağıldığı gösterilmektedir. Ev kadınları ve emekliler dışındaki meslek gruplarında kampanyaya katılan ve katılmayan

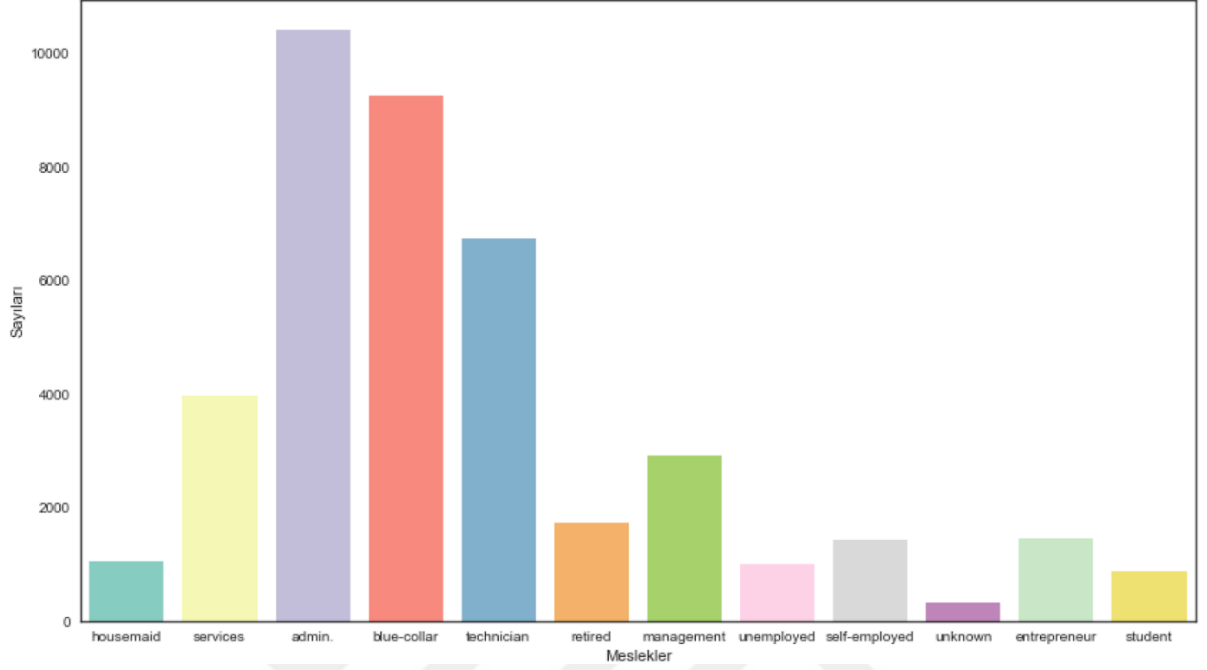
müşterilerin çoğu aynı yaş ortalamalarıdır. Örneğin, kampanyaya katılan ev kadınlarının ortalama yaşı 55 iken, kampanyaya katılmayan ev kadınları ortalama 45' tir. Kutu grafiği ile mesleklere göre hangi yaş aralıklarına odaklanılması gerektiği görülebilmektedir.

Öte yandan meslek gruplarına göre kampanya arama sayıları incelendiğinde en yüksek arama yönetim grubuna yapılmıştır. Bu gruptakilerin yaklaşık yüzde on ikisi kampanyaya katılım sağlamıştır. Meslek grupları kampanya arama sayıları göz önüne alınmadan incelendiğinde kampanyaya olumlu cevap verenler arasında en yüksek oran öğrencilerdir. Emekli bireyler, faiz ödemeleri yoluyla bir miktar nakit kazanmak için daha fazla vadeli mevduat bulundurma eğilimindedir.



Şekil 4.6: Kampanyaya katılım durumunun yaş ve mesleklerle ilişkisi.

Şekil 4.7’ de potansiyel müşterilerin meslek dağılımları verilmiştir.



Şekil 4.7: Potansiyel müşterilerin meslek dağılımları.

Tablo 4.2’ de meslek gruplarına göre kampanyaya katılım yüzdeleri görülmektedir. Yönetimde çalışanlar, mavi yaka ve teknisyenler vadeli mevduat kampanyasında en çok arama yapılan kesim olmuştur. Öğrenciler, girişimciler ve işsizler için daha az arama yapılmıştır. Öğrencilerin %31,42’si vadeli mevduat için onay vermiş, kampanyayı kabul etmiştir. Emekli olanların %25,23’ü vadeli mevduata onay vermiş, işsizlerin %14’ü kampanyaya olumlu bakmıştır.

Tablo 4.2: Meslek gruplarının kampanyaya katılım oranları.

Meslekler	Katılmayanların Oranı %	Katılanların Oranı %
Yönetim	87,02	12,97
Mavi Yaka	93,10	6,89
Girişimci	91,48	8,51
Ev Kadını	90,0	10,0
Yönetici	88,78	11,21
Emekli	74,76	25,23
Serbest Meslek	89,51	10,48
Hizmet	91,86	8,13
Öğrenci	68,57	31,42
Teknisyen	89,17	10,82
Çalışmayanlar	85,79	14,20
Bilinmiyor	88,78	11,21

Müşterilerin medeni durumlarına göre kampanyaya katılım durumları incelendiğinde en yüksek katılım bekar müşterilerde görülmüştür. Bekar müşterilerin %14'ü kampanyaya katılmıştır.

Ev kredisi olan müşterilerin kampanyaya katılım oranları incelendiğinde yüzde 11,58'i kampanyaya olumlu tepki vermiş, ancak ev kredisi olmayanların yüzde 10,88'i kampanyaya katılmıştır. Ev kredisinin niteliğinin kampanya açısından değerlendirilebilir bir nitelik olmadığı görülür (Tablo 4.3).

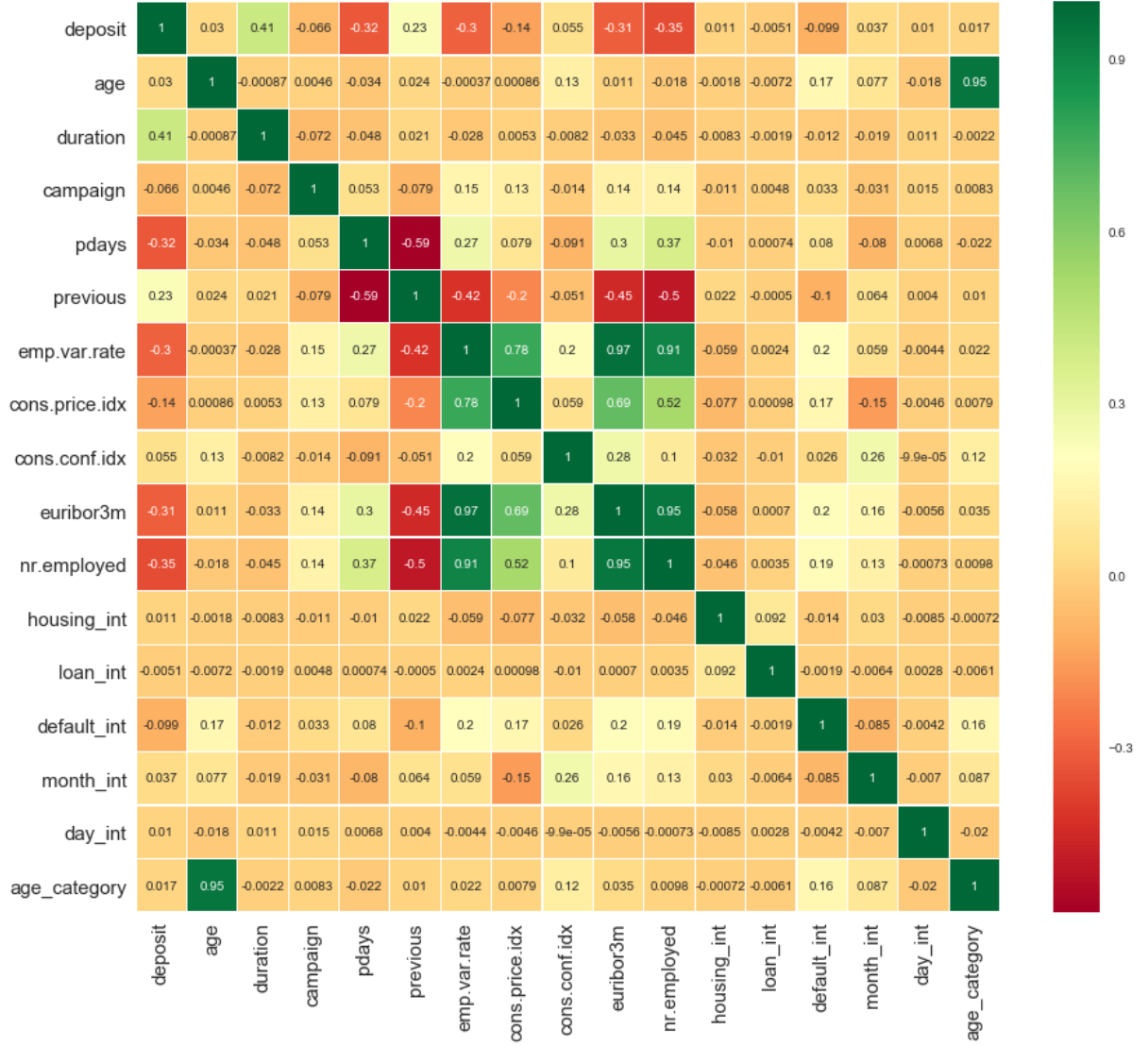
Tablo 4.3: Ev kredisi olan müşterilerin kampanyaya katılım oranları.

Ev Kredisi	Kampanyaya Katılmayanların Oranı (%)	Kampanyaya Katılanlar Oranı (%)
Yok	89,12	10,88
Var	10,87	11,58

Nitelikler arasındaki ilişkilerin birbirlerine olan etkisini incelemek için korelasyon matrisi oluşturulmuştur (Şekil 4.8). Korelasyon matrisinde bir kare ne kadar 1 değerine yakın olursa, sütun ile kolondaki nitelik arasında o kadar pozitif ilişki olduğu görülür. İlgili kolon ve sütundaki nitelikler birbirine bağlı olarak kararı etkilerler. Ters şekilde kare ne kadar -1'e yakın olursa, kolon ve sütundaki değer arasında o kadar negatif ilişki vardır. 0 değerine yakın olan kolon ve sütunlardaki nitelikler birbirleri ile daha az ilişkilidirler.

İşsizlik oranı ve Eurobor oranı ile istihdam ve tüketici fiyat endeksi arasında pozitif yönde ilişki vardır. Ayrıca işsizlik oranı ve Eurobor oranı arasında da pozitif yönde ilişki vardır. Daha önce yapılan aramaların sayısı ile en son arama yapılan zaman arasında negatif ilişki vardır. İşsizlik oranı, istihdam ve eurobor oranı ile daha önce yapılan arama sayıları arasında negatif ilişki vardır.

Kampanyaya katılım açısından niteliklerle olan ilişkiler incelendiğinde konuşma süresi ile pozitif yönde ilişkili olduğu görülmektedir. Öte yandan önceki arama yapılan zaman, istihdam oranı, eurobor oranı ve işsizlik oranı ile negatif ilişki olduğu görülmektedir.



Şekil 4.8: Korelasyon matrisi.

4.2. UYGULAMA

Çalışmada önceki bölümde yer verilen veri seti üzerinde Aşırı Öğrenme Makinesi sınıflandırıcısı kullanılarak, tele pazarlama kampanyasına katılacak müşterileri en iyi doğruluk oranı ile sınıflandırmak amaçlanmıştır. Veri seti gerekli ön işlemlerden geçirildikten sonra farklı sınıflandırma ve regresyon modelleri kullanılarak nitelik seçimi yapılmadan ve nitelik seçimi yapıldıktan sonra ki sınıflandırma sonuçları karşılaştırılmıştır. Modelin aşırı uyum ve düşük uyum durumlarını yönetebilmek için çapraz geçerleme kullanılmıştır. Çapraz geçerlemede k değeri 10 olarak seçilmiştir.

Nitelik seçimi öncesinde ve sonrasında model değerlendirme ölçütleri karşılaştırılmıştır.

Çalışmanın sonunda nitelik seçimi yapılarak Aşırı Öğrenme Makinesi sınıflandırma algoritması parametreleri ile farklı testler yapıp en iyi sınıflandırma sonuçlarına ulaşmak için hangi parametrelerin seçileceği incelenmiştir. Son olarak en iyi parametre seçimi için Genetik algoritma ile parametre optimizasyonu yapılmıştır.

Ayrıca elde edilen en iyi parametrelerle topluluk algoritmaları çalıştırılarak ulaşılan daha iyi sınıflandırma sonuçları incelenmiştir.

4.2.1. Parametre Değerleriyle Sınıflandırıcıların Değerlendirilmesi

AÖM sınıflandırıcısıyla model ayarlama öncesinde farklı sınıflandırma algoritmalarının performans sonuçları incelenmiştir. Veri seti nitelik seçimi öncesindeki haliyle Tablo 4.4'te yer alan sınıflandırma ve regresyon yöntemleriyle çalıştırılmıştır. Parametreler bölümünde sınıflandırıcı veya regresörlerin genel parametre değerleri belirtilmiştir. Parametrelerin genel parametre değerleri ile çalışmalar yapılmıştır.

Tablo 4.4: Sınıflandırıcı ve regresör parametreleri.

Sınıflandırma Algoritması	Parametreler
Gradyen Artırıcı Sınıflandırıcı	loss='deviance', learning_rate=0.1, n_estimators=100, subsample=1.0, criterion='friedman_mse', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_depth=3, min_impurity_decrease=0.0, min_impurity_split=None, init=None, random_state=None, max_features=None, verbose=0, max_leaf_nodes=None, warm_start=False, presort='auto'

Tablo 4.4 (devam): Sınıflandırıcı ve regresör parametreleri.

Sinir Ağları Sınıflandırıcı	hidden_layer_sizes=(100,),
	activation='relu',
	solver='adam',
	alpha=0.0001,
	batch_size='auto',
	learning_rate='constant',
	learning_rate_init=0.001,
	power_t=0.5,
	max_iter=200,
	shuffle=True,
	random_state=None,
	tol=0.0001,
	verbose=False,
	warm_start=False,
	momentum=0.9,
Lojistik Regresyon	nesterovs_momentum=True,
	early_stopping=False,
	validation_fraction=0.1,
	beta_1=0.9,
	beta_2=0.999,
	epsilon=1e-08
	penalty='l2',
	dual=False,
	tol=0.0001,
	C=1.0,
	fit_intercept=True,
	intercept,
	scaling=1,
	class_weight=None,
	random_state=None,
	solver='liblinear',
	max_iter=100,
	multi_class='ovr',
	verbose=0,
	warm_start=False,
	n_jobs=1

Tablo 4.4 (devam): Sınıflandırıcı ve regresör parametreleri.

K En Yakın Komşu	n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=1
Naives Bayes	priors=None
Karar Ağacı (CART)	criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, presort=False
Rasgele Orman Sınıflandırıcı	n_estimators=10, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None, verbose=0, warm_start=False, class_weight=None

Tablo 4.4 (devam): Sınıflandırıcı ve regresör parametreleri.

Lineer Diskriminant Analizi	solver='svd', shrinkage=None, priors=None, n_components=None, store_covariance=False, tol=0.0001
	C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', random_state=None

Tablo 4.5' te nitelik seçimi öncesi varsayılan genel parametre değerleriyle eğitim verisi üzerinde yapılan sınıflandırma sonuçları gösterilmektedir. Sınıflandırma performanslarına göre sonuçlar karşılaştırıldığında değerlerin 0,91627 ile 0,88807 doğruluk oranları arasında değiştiği görülmektedir. Pozitif ön görüş, duyarlılık oranları ve F1 skor değerleri 0,92 ile 0,89 arasında değişmektedir. Sonuçlar incelendiğinde en yüksek doğruluk oranı 0,91627 ile gradyen artırıcı sınıflandırıcı olduğu görülmüştür.

Tablo 4.5: Sınıflandırma performans sonuçları.

Sınıflandırma Algoritması	Doğruluk Oranları	Pozitif Öngörü Değeri	Duyarlılık Oranı	F1 Değeri	Skor
Gradyen Artırıcı Sınıflandırıcı	0,91627	0,91	0,92	0,91	
Sinir Ağları Sınıflandırıcı	0,91184	0,90	0,91	0,91	
Lojistik Regresyon	0,91120	0,90	0,91	0,91	
Lineer Diskriminant Analizi	0,90889	0,90	0,91	0,90	
Rasgele Orman Sınıflandırıcı	0,90871	0,90	0,91	0,90	
Destek Vektör Makinesi	0,90737	0,89	0,91	0,89	
K En Yakın Komşu	0,89982	0,89	0,90	0,89	
Naives Bayes	0,89982	0,89	0,90	0,89	
Aşırı Öğrenme Makinesi	0,89712	0,88	0,90	0,87	
Karar Ağacı (CART)	0,88807	0,89	0,89	0,89	

Her bir algoritmanın karmaşıklık matrisi sonucu Tablo 4.6’da gösterilmiştir.

Tablo 4.6: Karmaşıklık matrisi sonuçları.

Sınıflandırma Algoritması	Doğru Negatif Sayısı	Yanlış Pozitif Sayısı	Yanlış Negatif Sayısı	Doğru Pozitif Sayısı
Gradyen Artırıcı Sınıflandırıcı	28250	1000	1759	1941
Sinir Ağları Sınıflandırıcı	28238	1012	1893	1807
Lojistik Regresyon	28456	794	2132	1568
Lineer Diskriminant Analizi	28084	1166	1836	1864
Rasgele Orman Sınıflandırıcı	28426	824	2184	1516
Destek Vektör Makinesi	28689	561	2491	1209
K En Yakın Komşu	28127	1123	2178	1522
Naives Bayes	28127	1123	2178	1522
Aşırı Öğrenme Makinesi	28835	415	2975	725
Karar Ağacı (CART)	27365	1885	1803	1897

Tablo 4.7’de varsayılan parametrelerle sınıflandırma yapılması için geçen eğitim süresi gösterilmiştir. Verinin büyüklüğüne ve veri setinin boyutuna göre bu süre değişmektedir. Destek vektör makinesi eğitim süresinin en yüksek olduğu görülmüştür.

Tablo 4.7: Sınıflandırma algoritmalarına göre eğitim süreleri.

Sınıflandırma algoritması	Eğitim Süresi (saniye)
Naives Bayes	0,0275
Aşırı Öğrenme Makinesi	0,1172
Lineer Diskriminant Analizi	0,2089
Karar Ağacı (CART))	0,2501
Lojistik Regresyon	0,2856
K En Yakın Komşu	0,3855
Rasgele Orman Sınıflandırıcı	0,4894
Sinir Ağları Sınıflandırıcı	2,9811
Gradyen Artırıcı Sınıflandırıcı	7,1556
Destek Vektör Makinesi	22,4603

Eğitim süreleri karşılaştırıldığında destek vektör makineleri en uzun sürede eğiten sınıflandırıcı olmuştur. Diğer algoritmaların eğitim süresi sırasıyla Gradyen Artırıcı Sınıflandırıcı, Sinir Ağları Sınıflandırıcı, Rasgele Orman Sınıflandırıcısı, K En Yakın Komşu, Lojistik Regresyon, Karar Ağacı Sınıflandırıcısı, Lineer Diskriminant Analizi, Aşırı Öğrenme Makinesi, Naives Bayes sınıflandırıcıları olduğu görülmüştür. Model performansını etkileyen en önemli etkenlerden biri nitelik seçimidir. Karar ağacı sınıflandırması kullanılarak nitelik seçimi yapılmıştır.

4.2.2. Nitelik Seçimi Sonrasında Varsayılan Parametre Değerleriyle Sınıflandırıcıların Değerlendirilmesi

Malzeme ve Yöntem bölümünde yapılan karar ağacı sınıflandırıcısıyla nitelik seçimi sonuçlarına göre ulaşılan ihtiyaç kredisi (loan), kredi (default) nitelikleri, önceki kampanya başarısı (poutcome) ve ilgili müşteri dahil daha önce yapılan arama sayısı (previous) veri setinden çıkarılmıştır. Nitelik seçimi sonrasında varsayılan genel parametre değerleriyle eğitim verisi üzerinde yapılan sınıflandırma tekrarlanmıştır. Model sonuçları Tablo 4.8’ de gösterilmektedir.

Tablo 4.8: Sınıflandırma performans sonuçları.

Sınıflandırma Algoritması	Doğruluk Oranları	Pozitif Öngörü Değeri	Duyarlılık Oranı	F1 Skor Değeri
Gradyen Artırıcı Sınıflandırıcı	0,91669	0,91	0,92	0,91
Sinir Ağları Sınıflandırıcı	0,91147	0,90	0,91	0,91
Lojistik Regresyon	0,90989	0,90	0,91	0,90
Lineer Diskriminant Analizi	0,90886	0,90	0,91	0,90
Rasgele Orman Sınıflandırıcı	0,90640	0,90	0,91	0,90
Destek Vektör Makinesi	0,90586	0,89	0,91	0,89
K En Yakın Komşu	0,89933	0,89	0,90	0,89
Naives Bayes	0,89933	0,89	0,90	0,89
Aşırı Öğrenme Makinesi	0,89903	0,88	0,90	0,88
Karar Ağacı (CART)	0,88622	0,89	0,89	0,89

Kampanya katılımına etkisi olmadığı tespit edilen nitelikler çıkarıldıktan sonra test verisi ile yapılan sınıflandırma sonuçlarında daha yüksek doğruluk oranları görülmüştür. Bazı algoritmalarda ise düştüğü gözlemlenmiştir. Tablo 4.9’ da algoritmalar arasındaki değişim oranı gösterilmiştir. Gradyen Artırıcı Sınıflandırıcı sınıflandırma oranı %0.042 ve Aşırı Öğrenme Makinesi sınıflandırma oranı %0.191 artmıştır. Sinir Ağları Sınıflandırıcı sınıflandırma oranı %0.037, Lojistik Regresyon sınıflandırma oranı %0.131, Lineer Diskriminant Analizi sınıflandırma oranı %0.003, Rasgele Orman Sınıflandırıcısı sınıflandırma oranı %0.231, Destek Vektör Makinesi sınıflandırma oranı %0.151, K En Yakın Komşu sınıflandırma oranı %0.049, Naives Bayes sınıflandırma oranı %0.049 ve Karar Ağacı sınıflandırma oranı %0.185 kadar düşmüştür.

Tablo 4.9: Doğruluk oranları değişimi.

Sınıflandırma Algoritması	Değişim Oranları %
Gradyen Artırıcı Sınıflandırıcı	0.042
Sinir Ağları Sınıflandırıcı	-0.037
Lojistik Regresyon	-0.131
Lineer Diskriminant Analizi	-0.003
Rasgele Orman Sınıflandırıcı	-0.231
Destek Vektör Makinesi	-0.151
K En Yakın Komşu	-0.049
Naives Bayes	-0.049
Aşırı Öğrenme Makinesi	0.191
Karar Ağacı (CART)	-0.185

4.2.3. Aşırı Öğrenme Makinesi Sınıflandırma Algoritması ile Uygulama

Bu bölümde tezin araştırma konusu olan Aşırı Öğrenme algoritması ile bankacılık pazarlama veri setinin sınıflandırma çalışması yapılmıştır. Sınıflandırıcının parametreleri incelenmiş her bir parametre değişikliğinin model performans ölçütlerine olan etkisi incelenmiştir. En iyi tahmin sonucuna ulaşabilmek için farklı iterasyonlar ile model çalıştırılmıştır.

Analiz için, Python yazılım programında Aşırı Öğrenme Makinesi sınıflandırıcısı için elm kütüphanesinde yer alan ELMClassifier kullanılmıştır (Lambert, 2013).

Sınıflandırıcının Python' da kullanılabilen parametreleri ve açıklamaları varsayılan parametre değerleri ve kullanılabilen değerleri ile birlikte aşağıda yer almaktadır.

- `n_hidden` (int) - (varsayılan = 20) SimpleRandomLayer'da oluşturulacak gizli nöron sayısı.
- `alpha` (float) - (varsayılan = 0,5) Uzaklık ve nokta ürün giriş aktivasyonları için karıştırma katsayısı:

$$activation = alpha * mlp_activation + (1-alpha) * rbf_width * rbf_activation$$

- `activation_func` (string) – (varsayılan = tanh) giriş aktivasyonunu etkinleştirmek için kullanılan aktivasyon. Gizli birim aktivasyonlarını hesaplamak için kullanılacaktır.

Kullanılabilecek fonksiyonlar: *tanh, sine, tribas, sigmoid, hardlim, softlim, gaussian, multiquadric, inv_multiquadric, reclinear*

- `activation_args` (string) - Aktivasyon fonksiyonu için ek fonksiyonlar kullanılır.
- `regressor` (string) - Gizli birim aktivasyonlarından çıktılara ve sonraki tahminlere gerilemeyi gerçekleştirmek için kullanılır. Kullanılmazsa rasgele lineer en küçük kareler yöntemiyle yerleştirilir.
- `random_state` (int) - Sınıflandırıcı uyum zamanı için gizli birim ağırlıklarında kullanılan rasgele pseudo sayısı üretimini kontrol eden zaman değeri. Değer girilmezse rasgele sayı üretilir.

ELM Classifier kullanılarak Aşırı Öğrenme Makinesi parametreleri üzerinde model ayarlama ile altı parametrenin model performansı üzerindeki etkisi detaylı olarak incelenmiştir.

Parametreler üzerinde yapılan değişikliklerle, belli bir oranda sınıflandırma sonuçları yükseltilebilir. Model ayarlama farklı teknikler kullanılmaktadır. Bu çalışmada model ayarlama tekniklerinden biri ızgara arama yöntemi ve genetik algoritma ile parametre optimizasyonu yöntemleri incelenmiştir. Parametre aşırı uyuma ulaşması modelin aşırı eğitilmesi anlamına gelebilir. Belirli bir orandan sonra parametreler üzerinde değişiklikler yapmak yerine topluluk algoritmaları çalıştırarak modeli eğitmek daha doğru bir yaklaşımdır.

Son olarak model ayarlama teknikleriyle ulaşılan en iyi parametrelere göre topluluk metotlarıyla model geliştirilmiştir.

İlk olarak ızgara arama yöntemiyle gizli nöron sayısı değişikliğinin sınıflandırma sonuçlarına olan etkisi incelenmiştir. Tablo 4.10'da varsayılan gizli nöron sayısı olan 20 değerinden başlayarak 150, 200, 600, 2000 gizli nöron için test sonuçları karşılaştırılmıştır. Doğruluk oranlarına göre en yüksek değer seçildiğinde en uygun gizli nöron sayısının 150 adet olduğu görülmüştür. Model test doğruluk oranı 0,89903' ten 0,90750 oranına yükselmiştir.

Tablo 4.10: Gizli nöron sayısı değişikliklerinin sınıflandırmaya etkisi.

Gizli Nöron Sayısı	Alpha Değeri	Aktivasyon Fonksiyonu	Doğruluk Oranı	Pozitif Öngörü Değeri	Duyarlılık Oranı	F1 Skor Değeri
20	0,5	tanh	0,90155	0,88149	0,90155	0,87922
150	0,5	tanh	0,9075	0,89294	0,9075	0,89475
200	0,5	tanh	0,90289	0,88578	0,90289	0,88836
600	0,5	tanh	0,89973	0,88247	0,89973	0,88653
2000	0,5	tanh	0,88966	0,86682	0,88966	0,8735

150 gizli nöron seçilerek algoritma çalıştırılmaya devam edilmiştir. Sonraki parametre olan alpha değerinin varsayılan değeri 0,5' tir. Alpha değerini 0 ile 1 arasında farklı değerler vererek çalıştırıldığında en iyi sonuca 0,65 değerinde ulaşılmıştır. Doğruluk oranı 0,90896' ya ulaşmıştır. Alpha değeri değişikliklerinin sınıflandırmaya etkisi Tablo 4.11'de gösterilmiştir.

Tablo 4.11: Alpha değeri değişikliklerinin sınıflandırmaya etkisi.

Gizli Nöron Sayısı	Alpha Değeri	Aktivasyon Fonksiyonu	Doğruluk Oranı	Pozitif Öngörü Değeri	Duyarlılık Oranı	F1 Skor Değeri
150	0,01	tanh	0,9075	0,89216	0,9075	0,888
150	0,10	tanh	0,90325	0,88469	0,9033	0,88222
150	0,25	tanh	0,90058	0,87982	0,9006	0,87953
150	0,5	tanh	0,9075	0,89294	0,9075	0,89475
150	0,65	tanh	0,90896	0,89529	0,909	0,89713
150	0,75	tanh	0,90872	0,89544	0,9087	0,89778
150	0,99	tanh	0,90774	0,89335	0,9077	0,89519

Aktivasyon fonksiyonu için varsayılan fonksiyon *tanh*’ dır. ELMClassifier’ da kullanılabilen dokuz farklı fonksiyon incelendiğinde doğruluk oranlarına göre en iyi sonuç *inv_multiquadric* fonksiyonun vermiştir. Bu fonksiyon ile 0,91078 doğruluk oranına ulaşılmıştır. *inv_multiquadric* fonksiyondan sonra en başarılı sınıflandırıcı fonksiyonları sırasıyla *sigmoid*, *tanh*, *multiquadric*, *hardlim*, *softlim*, *gaussian*, *tribas*, *sine* (Tablo 4.12).

Tablo 4.12: Aktivasyon fonksiyonu değişikliklerinin sınıflandırmaya etkisi.

Gizli Nöron Sayısı	Alpha Değeri	Aktivasyon Fonksiyonu	Doğruluk Oranı	Pozitif Öngörü Değeri	Duyarlılık Oranı	F1 Skor Değeri
150	0,65	tanh	0,90896	0,89529	0,90896	0,89713
150	0,65	sine	0,88844	0,87529	0,88844	0,88053
150	0,65	tribas	0,89791	0,8776	0,89791	0,88133
150	0,65	sigmoid	0,90908	0,89559	0,90908	0,89754
150	0,65	hardlim	0,90714	0,89194	0,90714	0,89311
150	0,65	softlim	0,90714	0,89194	0,90714	0,89311
150	0,65	gaussian	0,90143	0,88338	0,90143	0,88622
150	0,65	multiquadric	0,90738	0,89268	0,90738	0,89441
150	0,65	inv_multiquadric	0,91078	0,89797	0,91078	0,89961

Tablo 4.13’de diğer parametrelerin etkilerini incelemek için aktivasyon fonksiyonu *inv_multiquadric* seçilerek alt aktivasyon fonksiyonu değişikliğinin sonuçları incelenmiştir. En iyi sonuç *softlim* alt fonksiyonunda görülmüştür. Parametre değişikliği ile 0,91114 doğruluk değerine ulaşılmıştır.

Tablo 4.13: Alt aktivasyon fonksiyonu değişikliklerinin sınıflandırmaya etkisi.

Gizli Nöron Sayısı	Alpha Değeri	Aktivasyon Fonksiyonu	Alt Aktivasyon Fonksiyonu	Doğruluk Oranı	Pozitif Öngörü Değeri	Duyarlılık Oranı	F1 Skor Değeri
150	0,65	inv_multiqu	adric	0,91029	0,89714	0,91029	0,89868
150	0,65	inv_multiqu	adric	0,90641	0,89199	0,90641	0,89461
150	0,65	inv_multiqu	adric	0,90859	0,89428	0,90859	0,89546
150	0,65	inv_multiqu	adric	0,90702	0,89245	0,90702	0,89461
150	0,65	inv_multiqu	adric	0,90617	0,89162	0,90617	0,89426
150	0,65	inv_multiqu	adric	0,91114	0,8983	0,91114	0,8996
150	0,65	inv_multiqu	adric	0,90908	0,89534	0,90908	0,89699
150	0,65	inv_multiqu	adric	0,90774	0,89339	0,90774	0,89527
150	0,65	inv_multiqu	adric	0,90835	0,89433	0,90835	0,89616

Regressör değeri üzerinde yapılan değişikliklerle lojistik regresör kullanıldığında model doğruluk ölçütlerinde bir iyileşme görülmemiştir. Tablo 4.14’de sonuçlar gösterilmektedir.

Tablo 4.14: Regresör değişikliğinin sınıflandırmaya etkisi.

Gizli Nöron Sayısı	Alpha Değeri	Aktivasyon Fonksiyonu	Alt Aktivasyon Fonksiyonu	Regresör	Doğruluk Oranı	Pozitif Öngörü Değeri	Duyarlılık Oranı	F1 Skor Değeri
150	0,65	inv_multiqu	adric	softlim	0,91114	0,8983	0,91114	0,8996
150	0,65	inv_multiqu	adric	softlim	0,90823	0,898	0,90823	0,9015

150 gizli nöron sayısı, *inv_multiquadric* aktivasyon fonksiyonu ve *softlim* alt aktivasyon fonksiyon ile en iyi doğruluk sonuçlarına ulaşılmıştır.

4.2.3.1. Genetik algoritma ile AÖM parametreleri optimizasyonu

Izgara arama yöntemi ile AÖM sınıflandırıcısı parametreleri üzerinde yapılan çalışmalarla ulaşılan değerlerin optimize edilerek daha iyi doğruluk oranlarına ulaşmak için genetik algoritma ile parametre optimizasyonu yapılmıştır.

Genetik algoritması uygulamasının amaç fonksiyonu; AÖM sınıflandırıcısı için en uygun parametrelerinin seçilerek maksimum sınıflandırma performansına ulaşmaktır. Fonksiyonun karar değişkenleri AÖM sınıflandırıcısı parametreleridir. Uygulamada kullanılan parametre değerleri aşağıda belirtilmiştir. Genetik algoritma 30 iterasyonla gen üretecek şekilde çalıştırılmıştır. Test ve eğitim verisi k değeri 10 ile çapraz geçerleme ile ayrılmıştır. Genetik Algoritma popülasyon boyutu 50 seçilmiştir. İki kromozom arasındaki gen değişim olasılığı yarı yarıya, kromozomda gen mutasyon olasılığı 0,1 seçilerek uygulanmıştır. Gen tipi olarak ailelerin iki geni arasında rastgele gen seçimi yapılmıştır.

AÖM parametreleri:

- Aktivasyon fonksiyonları: *tanh*, *sine*, *tribas*, *sigmoid*, *hardlim*, *softlim*, *gaussian*, *multiquadric*, *inv_multiquadric*
- Alpha değeri: 0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 2, 10
- Alt aktivasyon fonksiyonları: *tanh*, *sine*, *tribas*, *sigmoid*, *hardlim*, *softlim*, *gaussian*, *multiquadric*, *inv_multiquadric*
- Gizli nöron sayısı: 10, 200, 500, 700, 1000
- Rasgele sayı: 21398023

Genetik algoritma parametre optimizasyonu ile en iyi AÖM parametreleri aşağıdaki şekilde çıkmıştır. Doğruluk oranı 0,91114'ten 0,91156' ya yükselmiştir.

- Aktivasyon fonksiyonları: *multiquadric*
- Alpha değeri: 0.05,
- Alt aktivasyon fonksiyonları: *sigmoid*
- Gizli nöron sayısı: 700

4.2.3.2. Torbalama yöntemiyle modelin geliştirilmesi

Son olarak en iyi parametre değerleri ile topluluk metotları (Ensemble) uygulanmıştır (Tablo 4.15). Böylece model geliştirilerek daha doğru tahminleme yapılmaya çalışılmıştır.

- Torbalama yönteminde alt sınıflandırma algoritması aşırı öğrenme makinesi kullanılmıştır. Aynı sınıflandırıcıyı birden çok örneklem için çalıştırarak en iyi

sonuca ulaşma eğiliminde işlemler yapılır. Topluluk algoritmaları çalıştırıldığında doğruluk oranı 0,91156'da 0,90774' e düşmüştür.

- Oylama yöntemi en iyi sınıflandırıcılar olan Gradien Boost, Lojistik Regresyon, Neural ve AÖM üzerinde uygulanmıştır. Oylama sonucunda 0,91156'dan, 0,91612 doğruluk oranına ulaşılmıştır.

Tablo 4.15: Topluluk algoritmaları sonuçları.

Gizli Nöron Sayısı	Alpha Değeri	Aktivasyon Fonksiyonu	Alt Aktivasyon Fonksiyonu	Topluluk Algoritması Yöntemi	Doğruluk Oranı
700	0,05	multiquadric	sigmoid	-	0,91156
700	0,05	multiquadric	sigmoid	Torbalama	0,90774
700	0,05	multiquadric	sigmoid	Oylama (Gradien Boost, Neural, Logistic Regression, ELM)	0,91612

Veri seti üzerinde yapılan model ayarlama çalışmaları sonucunda en başarılı parametre değerlerine göre seçilen değerler, modelin yüksek performansa ulaşmasını sağlamıştır. AÖM ile geliştirilen model son olarak topluluk algoritmaları yöntemlerinden torbalama yöntemi ile farklı test kümelerinde birden çok AÖM algoritması eğitimi yapılmıştır. Test sonuçlarına göre müşterilerin kampanyaya katılımı 0,91612 oranında doğru olarak sınıflandırılmıştır.

5. TARTIŞMA VE SONUÇ

Bankacılık sektörü, tele pazarlamaya yönelik hedeflemeyi optimize etmek, karlılığı artırmak ve maliyetleri düşürmek konularında sürekli baskı altındadır. Bankalar için pazarlama kampanyalarının önemi gün geçtikçe artmaktadır. Çalışma kapsamında uygulanan pazarlama kampanyası ile arama yapılan müşteriye vadeli mevduat hesabı açarak bankadaki likiditeyi arttırmak, bankaya çapraz satış fırsatları yaratmak ve müşteri ilişkilerini geliştirmek hedeflenmiştir. Arama yapılacak müşterilerin doğru olarak sınıflandırılması, kaynağın kampanyaya olumlu cevap verecek müşterilere ulaşmak için kullanılmasını sağlayarak daha çok satış yapma fırsatı yaratacaktır. Ayrıca müşteri memnuniyetsizliğini azaltarak ve arama maliyetlerini düşürecektir. Kaynağın verimli kullanımından doğan ek bütçe çalışanların eğitimi ya da teknoloji yatırımı gibi banka süreçlerini iyileştirecek alanlarda ek yatırım yapma imkanı sağlayacaktır.

Bu çalışmada, doğrudan pazarlama ve tele pazarlama hakkında genel bilgiler verilmiştir. Pazarlama ile veri madenciliği arasındaki ilişki üzerinde durulmuş, veri madenciliği teknikleri ve süreci özetlenmiştir. Sınıflandırma algoritmalarından biri olan Aşırı Öğrenme Makinesi ve AÖM'nin temeli olan Yapay Sinir Ağları ile ilgili detaylı inceleme yapılmıştır. Çalışma sürecinde kullanılan nitelik önceliklendirme teknikleri ve veri ayırma teknikleri üzerinde durulmuştur.

Aynı veri seti kullanılarak yapılan çalışmalarda ulaşılan sonuçlar şu şekildedir: Zakaryazad ve Duman'ın (2016) yaptığı çalışmada farklı versiyondaki yapay sinir ağları ile 0,8262 ile 0,8830 değerleri arasında doğruluk oranlarına ulaşılmıştı. Apampa' nın (2016) çalışmasında ise 0,742 ile 0,766 arasında model doğruluğuna ulaşılmıştır. Ruangthong ve Jaiyen (2015) SMOTE algoritmaları ile yaptığı çalışmada 0,8051 ve 0,9081 değerleri arasında sonuçlara ulaşmıştır. Frempong ve Jayabalan' ın (2017) sınıflandırma karşılaştırma çalışmasında rasgele orman algoritması ile 0,868 doğruluk oranı görülmüştür. Rey ve diğ. (2017) geliştirdikleri yeni modelle 0,91 doğruluk oranına ulaşılmıştır. Moro ve diğ. (2014), orijinal veri setiyle yaptığı çalışmada AUC değeri ve ALIFT değeri ile model performansı ölçmüştür. Bu değerler 0,80 ve 0,67 çıkmıştır. Veri setiyle yapılan çalışmalarda ulaşılan sınıflandırma doğruluk oranları Tablo 5.1' de özetlenmiştir.

Tablo 5.1: Kullanılan veri seti ile yapılan diğer çalışmaların sonuçları.

Veri Setiyle Yapılan Çalışmalar	Doğruluk Oranları	
	En Düşük	En Yüksek
Zakaryazad ve Duman (2016)	0,826	0,883
Apampa (2016)	0,742	0,766
Ruangthong ve Jaiyen (2015)	0,805	0,908
Frempong ve Jayabalan 2017	0,868	
Rey ve diğ. (2017)	0,91	
Moro ve diğ. (2014)	0,67	0,8

Yapılan çalışma ile veri seti üzerinde AÖM sınıflandırıcısı ile yapılan model ayarlama çalışmaları sonucunda yüksek model performansına ulaşılması sağlanmıştır. Çalışmada sırasıyla özellik seçimi öncesinde AÖM ile sınıflandırma yapılmıştır. 0,89712 sınıflandırma oranına ulaşılmıştır. Veri setinde kampanyaya katılım sonucu ile ilgisiz olan özelliklerin çıkarılarak daha iyi sınıflandırma performansı elde etmek için tekrar AÖM ile sınıflandırma yapılmıştır. 0,89903 sınıflandırma oranına yükselmiştir. Izgara arama yöntemiyle en iyi AÖM parametreleri aranmış ve sınıflandırma oranı 0,91114' e ulaşılmıştır. AÖM parametreleri genetik algoritma ile optimize edilerek daha yüksek doğruluk oranlarına ulaşılmıştır. 0,91156 oranına ulaşılmıştır. AÖM ile geliştirilen model, son olarak topluluk algoritmaları yöntemlerinden torbalama yöntemi ile farklı test kümelerinde çalıştırılmış ve birden çok AÖM algoritması ile modelin eğitimi yapılmıştır. Test sonuçlarına göre müşterilerin kampanyaya katılımı 0,91612 oranında doğru olarak sınıflandırılmıştır. Tablo 5.2' de yapılan çalışmalarda ulaşılan sınıflandırma doğruluğu oranları gösterilmiştir.

Tablo 5.2: Uygulamada ulaşılan sonuçların doğruluk oranları.

Uygulamada Ulaşılan Sonuçlar	Doğruluk Oranı
Tüm Özelliklerle AÖM	0,89712
Seçilen Özelliklerle AÖM	0,89903
Izgara Arama ile AÖM	0,91114
Genetik Algoritma Parametre Optimizasyonu ile AÖM	0,91156
Topluluk Algoritmaları ile AÖM	0,91612

Yapılan çalışma ile veri setinin %91 oranında başarılı olarak sınıflandırılması, literatürde aynı veri setiyle yapılan çalışmalar arasında yüksek başarı sağlayan gruba girmesini sağlamıştır.

Daha önce aynı veri seti ile yapılan çalışmalar ile karşılaştırıldığında; sınıflandırma algoritması için Aşırı Öğrenme Makinesi' nin kullanılması, parametre seçimi için genetik algoritma optimizasyonu yapılması ve modelin torbalama algoritmaları ile geliştirilmesi bu çalışmayı diğerlerinden ayırmaktadır. Çalışmanın Aşırı Öğrenme Makinesi sınıflandırıcısı kullanmak isteyen ve model ayarlama üzerinde çalışmalar yapmak isteyen işletmeler ve akademik çalışmalara örnek teşkil etmesi umulmaktadır.

Bankalar için pazarlama kararları ve iş kararları, müşteri ile olan ilişkilerini korumak için her zamankinden daha önemlidir. Bankacılık sektörünün farklı çalışma alanlarında biriken büyük miktarda ve çeşitlilikteki büyük verinin saklanması, verinin işlenmesi ve veri üzerinden değerli bilgi elde etme için geleneksel yöntemler yetersiz kalmaktadır. Büyük veri teknolojileri ile daha hızlı, daha şeffaf, daha etkin çözümlerle bankaların farklı süreçlerinde iyileştirme noktaları, fırsatlar, verimli iş modelleri, sunulan hizmetlerin iyileştirilmesi ile müşteri memnuniyeti, yüksek kârlılık ve rekabet avantajı sağlanmaktadır. Tez çalışmasında detaylı olarak incelenen sınıflandırma modellerinin yüksek doğrulukta tahmin sonucunu ortaya koyması, hemen hemen her sektörde olduğu gibi bankacılık sektöründe de büyük kazanımlar elde etmeyi sağlamaktadır.

KAYNAKLAR

- Ahi, L., 2015, *Veri madenciliği yöntemleri ile ana harcama gruplarının paylarının tahmini*, Yüksek Lisans Tezi, Hacettepe Üniversitesi.
- Akpınar, H., 2000, Veri tabanlarında bilgi keşfi ve veri madenciliği, *İstanbul üniversitesi işletme fakültesi dergisi*, 29, 1-22.
- Alçın, Ö.F., Şengür, A. ve İnce, M.C., 2015, İleri-geri takip algoritması tabanlı seyrek aşırı öğrenme makinesi, *Gazi üniversitesi mühendislik-mimarlık fakültesi dergisi*, 30 (1), 111-117.
- Alpaydin, E., 2004, *Introduction to machine learning*. The MIT Press Cambridge, Massachusetts London, England.
- Apampa, O., 2016, Evaluation of classification and ensemble algorithms for bank customer marketing response prediction, *Journal of international technology and, information management*, 25 (4), 84-100.
- Argüden, Y. ve Erşahin, B., 2008, *Veri madenciliği: veriden bilgiye, masraftan değere*, ARGE Danışmanlık Yayınları, İstanbul.
- Asare-Frempong, J. ve Jayabalan, M., 2017, Predicting customer response to bank direct telemarketing campaign, *International conference on engineering technology and technopreneurship (ICE2T)*, 18-20 Eylül 2017, Kuala Lumpur.
- Balaban, M.E. ve Kartal, E., 2015, *Veri madenciliği ve makine öğrenmesi*, 1. baskı, Çağlayan Kitapevi, İstanbul, ISBN: 978-975-436-089-9.
- Bambang, B., Widodo, R.J., Sitalaksana, I.Z., Singgih, M., 2002, Indonesia stock market prediction (smgr/ggrm), using time series neural networks, *Proceedings of the sixth AEESEAP triennial conference*, 23-25 Ağustos 2002, Endonezya.
- Baxter, C.W., 2001, *Modelling heuristics from literature*, CIV E 729 Course Notes, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada.
- Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R. ve Gavaldà, R., 2009, New ensemble methods for evolving data streams, *KDD'09 proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, 28 Haziran-1 Temmuz 2009, Paris, 139-148.
- Birant, D., Kut, A., Ventura, M., Altınok, H., Altınok, B., Altınok, E. ve Ihlamur, M., 2010, İş zekası çözümleri için çok boyutlu birliktelik kuralları analizi, *XII. Akademik bilişim konferansı bildirileri*, 10-12 Şubat 2010, Muğla Üniversitesi, 215-222.
- Blum, A., ve Langley, P., 1997, Selection of relevant features and examples in machine learning, *Artificial intelligence*, 97, 245-271.
- Breiman, L., 2001, Random forests, *Machine learning*, 45 (1), 5-32.

- Brownlee, J., 2016, *Machine learning mastery with python*.
- Burget, L., 2004, *Complementarity of speech recognition systems and system combination. Computer graphics and multimedia*, Ph.D. Dissertation, Brno University of Technology.
- Catalao, J.P.S., Mariano, S.J.P.S., Mendes, V.M.F. ve Ferrira, L.A.F.M., 2007, Short-term electricity prices forecasting in a competitive market: A neural network approach, *Electric power system research*, 77, 1297-1304.
- Cook, D.F., Ragsdale C.T. ve Major R.L., 2000, Combining a neural network with a genetic algorithm for process parameter optimization, *Engineering Application of Artificial Intelligence*, 13, 391-396.
- Choi, T.M., Li, D. ve Yan, H., 2006, Quick response policy with Bayesian information updates, *European journal of operational research*, 170, 788–808.
- Çil, F., 2010, *Banka yatırım fonu müşteri hareketlerinin belirlenmesine yönelik bir veri madenciliği uygulaması*, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü.
- Das, S., 2001, Filters, wrappers and a boosting-based hybrid for feature selection. *Proceedings of the Eighteenth International Conference on Machine Learning*, In: Brodley, C.E. ve Danyluk, A.P. (ed.), 28 Haziran-1 Temmuz 2001, San Francisco, Morgan Kaufmann Publishers, 74–81.
- Doğan Ş. ve Türkoğlu İ., 2007, Hypothroidi and hyperthyroidi detection from thyroid hormone parameters by using decision trees, *Doğu anadolu bölgesi araştırmaları dergisi*, 5 (2), 163-169.
- Dong, Y. ve Han, K., 2004, A comparison of several ensemble methods for text categorization, *Proceedings of the 2004 IEEE international conference on services computing*, 15-18 Eylül 2004, Şangay.
- Dror, G., 2005. *Analysis of gene expression data*, Tel Aviv University School of Computer Science, <http://www.cs.tau.ac.il/~rshamir/ge/05/scribes/lec06.pdf>, [Ziyaret tarihi: 12 Nisan 2018].
- Duda, R.O., Hart, P.E. ve Stork, D.G., 2000, *Pattern classification*, 2nd ed. Wiley-Interscience, New York, ISBN: 978-0-471-05669-0.
- Dudek, G., 2014, Short-Term Load Forecasting using Random Forests, *Intelligent Systems'2014: Proceedings of the 7th IEEE International Conference Intelligent Systems IS'2014*, 24-26 September 2014, Polonya, 821-828.
- Duman, A., 2010, *Multiple criteria sorting methods based on support vector machines*, Yüksek Lisans Tezi, Ortadoğu Teknik Üniversitesi.
- Dunham, M.H. 2003, *Data mining introductory and advanced topics*, Pearson Education Inc., New Jersey.

- Emgin, Ö. ve Süngü, A., 2004, *Pazarlama kavramı içerisinde tutundurma fonksiyonu*, Mevzuat dergisi, 7 (78), <https://www.mevzuatdergisi.com/2004/06a/02.htm>, [Ziyaret tarihi:25 Nisan 2018].
- Eser, Z. ve Ilgaz Sümer, S., 2006, Pazarlama karması mal ve hizmetler için tüketiciler tarafından nasıl değerlendirilir? Betimleyici bir araştırma, *Gazi üniversitesi iktisadi ve idari bilimler fakültesi dergisi*, 8 (3), 111-132.
- Esmaeili, M. ve Gabor, F., 2011, Feature Selection as an Improving Step for Decision Tree Construction, *2009 International conference on machine learning and computing IPCSIT vol.3*, Singapur, 35-39.
- Fausett, L., 1994, *Fundamentals of neural networks: architectures, algorithms and applications*, Englewood Cliffs: Prentice-Hall.
- Fayyad, U., Piatetsky-Shapiro, G. Ve Smyth, P., 1996, The KKD process for extracting useful knowledge from volumes of data, *Communications of the ACM*, 39 (11), 27-34.
- Fayyad, U.M., 1996, Data mining and knowledge discovery: Making sense out of data, *IEEE Expert: Intelligent Systems and Their Applications*, 11 (5), 20-25.
- Friedman, J.H., 2002, Stochastic gradient boosting. *Computational statistics and data analysis*, 38 (4), 367-378.
- Friedrich, F. ve Igel, C., 2005, Evolutionary tuning of multiple SVM parameters, *Neurocomputing*, 64, 107-117.
- Gallant, S.I., 1993, *Neural network learning and expert systems*, MIT Press, Cambridge.
- Garofalakis, M., Hyun, D., Rastogi, R., ve Shim, K., 2003, Building decision trees with constraints, *Data mining and knowledge discovery*, 7 (2), 187 – 214. Me
- Gayatri, N., Nickolas, S. ve Reddy, A.V., 2010, Feature selection using decision tree induction in class level metrics dataset for software defect predictions, *Proceedings of the world congress on engineering and computer science vol I WCECS*, 20-22 Ekim 2010, San Francisco.
- Geçer, T., 2016, Ürün pazarlamasında doğrudan pazarlama gelişme potansiyeli ve tutundurma karması içerisindeki önemi: Eskişehir ilindeki işletmelerde doğrudan pazarlama uygulamalarına ilişkin çalışma, *Siirt üniversitesi iktisadi ve idari bilimler fakültesi iktisadi yenilik dergisi*, 3 (2), 46-71.
- Gelibolu, L. ve Özsoy, T. 2013, Çağrı merkezlerinin satış amaçlı kullanılması: doğrudan pazarlamanın bir unsuru olarak telepazarlama, *Çukurova üniversitesi sosyal bilimler enstitüsü dergisi*, 22 (1), 481-500.
- Grabczewski, K. ve Jankowski, N., 2005, Feature selection with decision tree criterion, *Hybrid Intelligent Systems, HIS '05. Fifth International Conference*, 6-9 Şubat 2005, Rio de Janeiro.

- Ding, X. ve Zhou, H., 2010, Optimization method based extreme learning machine for classification, *Neurocomputing*, 74, 155–163.
- Hair, J.F., Anderson, R.E., Tatham, R.L. ve Black, W.C., 1998, *Multivariate Data Analysis*, 5th ed., Prentice Hall, New Jersey.
- Han, J., Kamber, M. ve Pei, J., 2012, *Data mining: concepts and techniques*, 3rd. ed., Morgan Kaufman Publishers, ABD, ISBN: 978-0-12-381479-1.
- Hanbay, D., Turkoglu, I. ve Demir, Y., 2010, Modeling switched circuits based on wavelet decomposition and neural networks, *Journal of the franklin institute*, 347 (3), 607-617.
- Hand, D., Manilla, H., ve Smyth, P., 2001, *Principles of data mining*, MIT Press, England, ISBN: 0-262-08290-X.
- Huang, G.B., Chen, L. ve Siew, C.K., 2006, Universal approximation using incremental constructive feedforward networks with random hidden nodes, *IEEE transactions on neural networks*, 17 (4), 879-892.
- Huang, G.B., Wang, D.H. ve Lan, Y., 2011, Extreme learning machines: a survey, *International journal of machine learning and cybernetics*, 2, 107–122.
- Huang, G.B., Zhu, Q.Y. ve Siew, C.K., 2006, Extreme learning machine: Theory and applications, *Neurocomputing*, 70, 489-501.
- Huang, G.B. ve Chena, L., 2007, Convex incremental extreme learning machine, *Neurocomputing*, 70, 3056–3062.
- Huang, G., Ding, X. ve Zhou, Z., 2010, Optimization method based extreme learning machine for classification, *Neurocomputing*, 74, 155-163.
- İslamoğlu, A.H., 2002, *Pazarlama İlkeleri*, Beta Basım Yayım, İstanbul, Türkiye.
- Jain, A., Menon, M.N. ve Chandra, S., Sales forecasting for retail chains, <https://pdfs.semanticscholar.org/76a2/44f4da1d29170a9f91d381a5e12dc7ad2c0f.pdf>, [Ziyaret tarihi: 5 Nisan 2018].
- Kakıllı, S. ve Parlar, T. 2017, Using data mining techniques for detecting the important features of the bank direct marketing data, *International journal of economics and financial issues*, 7 (2), 692-696.m
- Kalıkovi A., 2006, *Veri madenciliği ve bir e-ticaret uygulaması*, Yüksek Lisans Tezi, Gazi Üniversitesi.
- Kearns, M. ve Valiant, L.G., 1988, *Learning boolean formula and finite automata*, Harvard University, Center for Research in Computing Technology, Aiken Computation Laboratory putation Laboratory, 1988.
- Kobs, J., 1991, *Profitable direct marketing*, Chicago: NTC Business Books.

- Kotler P. ve Keller K.L., 2012, *Marketing management*, 14th ed., Pearson, Upper Saddle River, N.J.
- Kotler, P. ve Armstrong, G., 2018, *Principles of marketing*, Seventeenth Edition, Pearson, United Kingdom, ISBN 13: 978-1-292-22017-8.
- Kotler, P., Wong, V., Saunders, J., ve Armstrong, G., 2005, *Principles of marketing*, Fourth European Edition, Pearson Prentice Hall, England, ISBN-13: 978-0-273-68456-5.
- Kulluk, S., 2009, *Karınca koloni optimizasyonu ile yapay sinir ağlarından kural çıkarımı*, Doktora Tezi, Erciyes Üniversitesi Fen Bilimleri Enstitüsü.
- Kuncheva, L., ve Whitaker, C., 2003, Measures of diversity in classifier ensembles, *Machine learning*, 51, 181-207.
- Lambert, D., Yazılım geliştirme projelerini içeren web sitesi, <https://github.com/dclambert/Python-ELM>, [Ziyaret tarihi: 3 Şubat 2018].
- Lan, Y., Soh, Y.C. ve Huang, G.B., 2009, Ensemble of online sequential extreme learning machine, *Neurocomputing*, 72, 3391–3395.
- Li, G., Wen, C., Li, Z. G., Zhang, A., Yang, F. ve Mao, K., 2013, Model-based online learning with kernels, *Neural networks and learning systems*, 24 (3), 356-369. metinde bulamadım
- Lu. J., Plataniotis, K.N., Venetsanopoulos, A.N. ve Li, S.Z., 2006, Ensemble-based discriminant learning with boosting for face recognition, *IEEE transactions on neural networks*, 17 (1):166-78.
- Lu, C. ve Shao Y., 2012, Forecasting computer products sales by integrating ensemble empirical mode decomposition and extreme learning machine, *Mathematical Problems in Engineering*, 15.
- Maimon O. ve Rokach L. 2005, *Data mining and knowledge discovery handbook*, Springer, New York.
- Maimon, O. ve Rokach, L., 2010, *Introduction to knowledge discovery and data mining*, Data mining and knowledge discovery handbook, 2nd ed., Springer US, ABD, ISBN: 978-0-387-09822-7, 1–15.
- Miche, Y., Sorjamaa, A., Bas, P., Simula, O., Jutten, C. ve Lendasse, A., 2010, OP-ELM: optimally pruned extreme learning machine, *IEEE transactions on neural networks*, 21 (1).
- Monino, J-L. ve Sedkaoui, S., 2016, *Big data, open data and data development*. Wiley-ISTE, ISBN: 978-1-848-21880-2.
- Moro, S., Cortez, P. ve Rita, P., 2014, A data-driven approach to predict the success of bank telemarketing, *Decision support systems* 62, 22-31, <https://archive.ics.uci.edu/ml/datasets/bank+marketing>, [Ziyaret tarihi: 18 Şubat 2018].

- Mucuk, İ., 1997, *Pazarlama ilkeleri (örnek olaylar)*, 8. Baskı, Türkmen Kitapevi, İstanbul.
- Nehdi, M., El Chabib, H. ve El Naggar, H., 2001, Predicting performance of self-compacting concrete mixtures using artificial neural networks”, *ACI materials journal*, 98 (5), 394-401.
- Öndoğan, E., 2010, Restoran pazarlamasında kullanılan temel pazarlama karma elemanları, *Ege stratejik araştırmalar dergisi*, 1 (1), 1-25.
- Özcan, C., 2014, *Veri madenciliği güvenlik uygulama alanları ve veri madenciliği sahtekarlık analizi*, Yüksek Lisans Tezi, İstanbul Bilgi Üniversitesi Sosyal Bilimler Enstitüsü.
- Özçalıcı, M., 2017, Eşik değerinin hisse senedi fiyat tahmin performansına etkisinin incelenmesi, *Hacettepe Üniversitesi iktisadi ve idari bilimler fakültesi dergisi*, 35 (1), 67-88. metinde yok
- Özdemir, Ş., 2017, Karar ağaçları, Rapidminer ile Uygulamalı Veri Madenciliği, In: Çelik, U., Akçetin, E. ve Gök, M. (ed.), Pusula Yayıncılık.
- Özkan, Y., 2008, *Veri madenciliği yöntemleri*, Papatya Yayıncılık, İstanbul.
- Özmen, M., 2013, *Pazarlamada temel kavramlar*, Pazarlama İlkeleri, In: Torlak, Ö. ve Özmen, M. (ed.), Bölüm 1, T.C. Anadolu Üniversitesi Yayını No: 3012, Eskişehir.
- Öztemel, E., 2012, *Yapay sinir ağları*, 3. Basım, Papatya Yayıncılık Eğitim, İstanbul.
- Prytherch, R., 2005, *Harrod's librarians' glossary and reference book: a dictionary of over 10,200 terms*. 10. ed. Ashgate, England.
- Rumelhart, D.E., Hinton, G.E., ve Williams, R.J., 1986, *Learning internal representations by error propagation*, Parallel Distributed Processing: Explorations in the Microstructures of Cognition, In: Rumelhart D.E., McClelland, J.L. (eds.), MIT Press, Cambridge, 318-362.
- Ruangthong, P. ve Jaiyen, S., 2015, Bank direct marketing analysis of asymmetric information based on machine learning, *Computer science and software engineering*, 22-24 Haziran 2015 Thailand, 93-96.
- Sankur, B., 2004, *İngilizce -Türkçe ansiklopedik bilişim sözlüğü*. Pusula, İstanbul.
- Sarıyer, N., 2007, Banka çağrı merkezi pazarının bölümlendirilmesi-Yozgat il merkezi'nde bir uygulama, *ZKÜ sosyal bilimler dergisi*, 3 (6), 149-162.
- Savaş, S., Topaloğlu, N. ve Yılmaz, M., 2012, Veri madenciliği ve Türkiye'deki uygulama örnekleri, *İstanbul ticaret üniversitesi fen bilimleri dergisi*, 11 (21), 1-23. metinde yok
- Schalkoff, R., 1992, *Pattern recognition-statistical, structural and neural approaches*. John Wiley and Sons, USA.
- Sollich, P. ve Krogh, A., 1996, Learning with ensembles: How overfitting can be useful, *Advances in neural information processing systems*, 8, 190-196.

- Street, W.N. ve Kim, Y., 2001, A streaming ensemble algorithm (sea) for large-scale classification,
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.482.3991&rep=rep1&type=pdf>, [Ziyaret tarihi: 22 Nisan 2018].
- Sun, Z.L., Choi, T.M., Au, K.F. ve Yu, Y., 2008, Sales forecasting using extreme learning machine with applications in fashion retailing, *Decision support systems*, 46 (1), 411-419.
- Tahir, N.M., Hussain, A., Samad, S.A., Ishak, K.A. ve Halim, R.A., 2006, Feature selection for classification using decision tree, *4th Student conference on research and development (SCORed 2006)*, 27-28 Temmuz 2006, Selangor.
- Tu, J.V., 1996, Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, *Journal of clinical epidemiology*, 49 (11), 1225-1231. metinde yok
- Uğur, A., ve Kınacı, A.C., 2006, Yapay zeka teknikleri ve yapay sinir ağları kullanılarak web sayfalarının sınıflandırılması, *XI. Türkiye'de İnternet" Konferansı (inet-tr'06)*, 21-23 Aralık 2006 Ankara, TOBB Ekonomi ve Teknoloji Üniversitesi, 1-4. metinde yok
- Varinli, İ. Ve Kahraman, Ç., 2008, *Güncel pazarlama yaklaşımlarından seçmeler*. Detay Yayıncılık. Ankara.
- Villuendas-Rey, Y., Rey-Benguría, C.F., Ferreira-Santiago, A., Camacho-Nieto, O., Yáñez-Márquez, C., 2017, The Naïve Associative Classifier (NAC): A novel, simple, transparent, and accurate classification model evaluated on financial data, *Neurocomputing*, 265, 105–115.
- Wang, H., Fan, W., Yu, P.S. ve Han, J., 2002, Mining concept-drifting data streams using ensemble classifiers, *KDD'03 proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining*, 24-27 Ağustos 2003, Washington, D.C., 226-235.
- Wong, W.K. ve Guo, Z.X., 2010, A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm, *International journal of production economics*, 128, 614-624.
- Wu, C., Tzeng G., Goo Y. ve Fang W., 2007, A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy, *Expert systems with applications*, 32, 397-408.
- Xia, C., Dong, Z.Y., Meng, K., Xu, Y., Wong, K.P. ve Ngan, H.W., 2012, Electricity price forecasting with extreme learning machine and bootstrapping, *IEEE transactions on power systems*, 27 (4), 1-8.
- Yalabık, İ., 2007, *A pattern classification approach boosted with genetic algorithms*, Yüksek Lisans Tezi, Ortadoğu Teknik Üniversitesi. Metnde yok
- Ye, N., 2003, *The handbook of data mining*, 1st ed., Lawrance Erlbaum.

- Yeh, I., Lien, C., 2007, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, *Expert systems with applications*, 36 (2), 2473-2480.
- Yeh, I.C., 1998, Modeling of strength of high-performance concrete using artificial neural networks, *Cement and concrete research*, 28 (12), 1797- 1808.
- Yuan, L., Soh, Y. ve Huang, G., 2009, Ensemble of online sequential extreme learning machine, *Neurocomputing*, 72, 3391-3395.
- Yurdakul, M. ve Kiracı, H., 2008, Sanal pazarlama karması bileşimi, *Süleyman demirel üniversitesi iktisadi ve idari bilimler fakültesi dergisi*, 13 (2), 165-185.
- Zakaryazad, A. ve Duman, E., 2016, A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing, *Neurocomputing*, 175, 121–131.
- Zengin, A.M., 2010, *Doğrudan pazarlama aracı olarak cep telefonu kullanımı ve tüketici tutumları*, Yüksek Lisans Tezi, Selçuk Üniversitesi.

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Arca ÇAKMAKOĞLU
Doğum Yeri	İstanbul
Doğum Tarihi	05.05.1990
Uyruğu	☒ T.C. ☐ Diğer:
Telefon	05384256646
E-Posta Adresi	arcacakmakoglu@gmail.com
Web Adresi	

Eğitim Bilgileri	
Lisans	
Üniversite	Yıldız Teknik Üniversitesi
Fakülte	Makina
Bölümü	Endüstri Mühendisliği
Mezuniyet Yılı	18.06.2013

Yüksek Lisans	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Endüstri Mühendisliği Anabilim Dalı
Programı	Endüstri Mühendisliği Programı