

T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ



YÜKSEK LİSANS TEZİ

FİLM ENDÜSTRİSİNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE YENİ ÜRÜNLERİN GELİR TAHMİNİ

Özge Hüsniye NAMLI

Endüstri Mühendisliği Anabilim Dalı

Endüstri Mühendisliği Programı

DANIŞMAN
Dr. Öğr. Üyesi Tuncay ÖZCAN

Haziran, 2018

İSTANBUL

Bu çalışma,25.06.2018 tarihinde aşağıdaki jüri tarafındanEndüstri Mühendisliği Anabilim Dalı

,Endüstri Mühendisliği Programında Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi

Dr. Öğr. ÜyesiTuncay ÖZCAN(Danışman)
İstanbul Üniversitesi
Mühendislik Fakültesi

Prof. Dr.Şakir ESNAF
İstanbul Üniversitesi
Mühendislik Fakültesi

Prof. Dr.Alp BARAY
İstanbul Üniversitesi
Mühendislik Fakültesi

Doç. Dr.Tarık KÜÇÜKDENİZ
İstanbul Üniversitesi
Mühendislik Fakültesi

Dr. Öğr. ÜyesiZeynep GERGİN
İstanbul Kültür Üniversitesi
Mühendislik 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 abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

ÖNSÖZ

Tez çalışması süresince her türlü konuda bilgi ve tecrübesi ile bana yol gösteren, yardımını ve desteğini esirgemeyen çok değerli danışmanım Dr. Öğr. Üyesi Tuncay ÖZCAN'a sonsuz teşekkürlerimi sunarım.

Hayatımın her anında olduğu gibi, tez çalışmam boyunca da bana destek olan, sonsuz güvenlerini daima arkamda hissettiğim sevgili aileme çok teşekkür ederim.

Haziran 2018 Özge Hüsniye NAMLI

İÇİNDEKİLER

Sayfa No

ÖNSÖZ ..... **iv**
İÇİNDEKİLER ..... **v**
ŞEKİL LİSTESİ ..... **vii**
TABLO LİSTESİ ..... **viii**
SİMGE VE KISALTMA LİSTESİ ..... **ix**
ÖZET ..... **xi**
SUMMARY ..... **xiii**
1. GİRİŞ **1**
2. GENEL KISIMLAR **4**
2.1. VERİ MADENCİLİĞİ 4
2.1.1. Veri Madenciliği Tanımı 4
2.1.2. Veri Madenciliği Süreci 5
2.1.3. Veri Madenciliğinin İşlevleri 6
2.1.3.1. Kümeleme Analizi 6
2.1.3.2. Birliktelik Kuralları ve İlişki Analizi 8
2.1.3.3. Sınıflandırma 9
2.2. MAKİNE ÖĞRENMESİ 11
2.2.1. Makine Öğrenmesi Tanımı 11
2.2.2. Bileşik Makine Öğrenmesi Tanımı 12
2.3. SINIFLANDIRMA 13
2.3.1. Bireysel Sınıflandırma Algoritmaları 13
2.3.1.1. Yapay Sinir Ağları 13
2.3.1.2. Destek Vektör Makinaları 15
2.3.2. Bileşik Makine Öğrenmesi Algoritmaları 19
2.3.2.1. Torbalama Yöntemi 19
2.4. MODEL GEÇERLEME YÖNTEMLERİ 20
2.4.1. Belirli Oranlarda Eğitim ve Test Verisi Olarak Bölme 20
2.4.2. Çapraz Geçerleme 21
2.5. MODEL PERFORMANS DEĞERLENDİRME ÖLÇÜLERİ 21
2.5.1. Karışıklık Matrisi 21

2.5.2. Doğruluk Oranı 22
2.5.3. Kesinlik 22
2.5.4. Anma 23
2.5.5. F-Ölçütü 23
2.5.6. Kappa İstatistiği 23
2.6. LİTERATÜR İNCELEMESİ 24
3. MALZEME VE YÖNTEM 41
3.1. PROBLEMİN TANIMI 41
3.2. VERİ SETİNİN OLUŞTURULMASI 42
3.3. MAKİNE ÖĞRENMESİ ALGORİTMALARININ UYGULANMASI 46
3.4. PARAMETRE ENİYİLEMESİ İLE ÖNERİLEN YÖNTEMLERİN PERFORMANSLARININ İYİLEŞTİRİLMESİ 48
4. BULGULAR 52
5. TARTIŞMA VE SONUÇ 56
KAYNAKLAR 59
EKLER 69
EK 1. OMDB API'si ile IMDB'Nin İnternet Sitesinden İstenen Verinin Çekilmesi için C# Kodu 69
EK 2. Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Yapay Sinir Ağları Modeli Kodları 71
EK 3.Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Destek Vektör Makinaları Modeli Kodları 72
EK 4.Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Torbalama Modeli Kodları 73
EK 5. Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Yapay Sinir Ağları Modeli En İyi Parametre Arama Kodları 74
EK 6. Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Destek Vektör Makinaları Modeli En İyi Parametre Arama Kodları 75
EK 7. Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Torbalama Modeli En İyi Parametre Arama Kodları 76
ÖZGEÇMİŞ 77

ŞEKİL LİSTESİ

Sayfa No

Şekil 2.1: Bilgi keşfi süreci. 5

Şekil 2.2: Sınıflandırma algoritmalarına genel bakış. 10

Şekil 2.3: Yapay Sinir Ağlarının yapısı. 14

Şekil 2.4: Doğrusal ayrılabilen veriler için DVM. 16

Şekil 3.1: Çalışmanın akış şeması. 42

Şekil 3.2: Veri setinin özet görünümü. 46

TABLO LİSTESİ

Sayfa No

Tablo 2.1: DVM'deki çekirdek fonksiyonların performansları. 18

Tablo 2.2: 5-kat çapraz geçerlemenin örnek gösterimi. 21

Tablo 2.3: Karışıklık matrisi. 22

Tablo 2.4: Kappa istatistiğinin yorumu. 24

Tablo 2.5: Gişe hasılatı tahmini problemine ilişkin literatür özeti. 38

Tablo 3.1: MPAA film derecelendirme sistemi. 43

Tablo 3.2: Filmlerin aylara göre mevsimsellik değerleri. 43

Tablo 3.3: Veri setindeki değişkenler. 45

Tablo 3.4: YSA algoritmasının varsayılan parametre değerleri. 47

Tablo 3.5: DVM algoritmasının varsayılan parametre değerleri. 47

Tablo 3.6: Torbalama algoritmasının varsayılan parametre değerleri. 47

Tablo 3.7: Çalışmada kullanılan algoritmaların parametrelerinin değer aralıkları. 51

Tablo 4.1: Sınıflandırma modellerinin performans analizi (çapraz geçerleme 10-kat ile). 52

Tablo 4.2: Sınıflandırma modellerinin performans analizi (çapraz geçerleme 5-kat ile). 53

Tablo 4.3: Sınıflandırma modellerinin performans analizi (%80 eğitim-%20 test). 54

Tablo 4.4: Sınıflandırma modellerinin performans analizi (%66,6 eğitim-%33,3 test). 55

SİMGE VE KISALTMA LİSTESİ

Kısaltmalar **Açıklama**

AAP : Ağızdan Ağıza Pazarlama
AHST : Akıllı Hızlı Satış Tahmini
AÖM : Aşırı Öğrenme Makinaları
BAU : Bağımsız Alt Uzay
BBK : Bayes Bilgi Kriteri
ÇBK : Çarpıklık Bilgi Kriteri
ÇDR : Çoklu Doğrusal Regresyon
ÇDURU : Çok Değişkenli Uyarlamalı Regresyon Uzanımları
ÇKFBA : Çok Katmanlı Fonksiyonel Bağlantı Ağları
ÇKA : Çok Katmanlı Algılayıcı
ÇLR : Çoklu Lojistik Regresyon
ÇRA : Çoklu Regresyon Analizi
DA : Diskriminant Analizi
DR : Doğrusal Regresyon
DTÇ : Durum Tabanlı Çıkarsama
DVM : Destek Vektör Makineleri
DVR : Destek Vektör Regresyonu
DYSA : Dinamik Yapay Sinir Ağları
ESA : Evrimci Sinir Ağları
GA : Genetik Algoritma
GAÖM : Gri Aşırı Öğrenme Makinesi
GHKK : Göreli Hata Karelerinin Karekökü (RRSE)
GİA : Gri İlişkisel Analiz
GM : Gri Model
GMH : Göreli Mutlak Hata (RAE)
HA : Hızlandırılmış Ağaç
HO : Hareketli Ortalama
k-EYK : k-En Yakın Komşu
KÇ : Kelime Çantası
KÖH : Kendini Örgütleyen Haritalar

M5A : M5 Model ağaçları
MYHM : Mutlak Yüzde Hataların Medyanı
MZ : Melez Zekâ
MZMCS : Markov Zinciri Monte Carlo Simülasyonu
ODTO : Ortalama Doğru Tahmin Etme Oranı
OHKK : Ortalama Hata Karelerinin Karekökü (RMSE)
OKH : Ortalama Hata Kareleri
OMH : Ortalama Mutlak Hata
OMS : Ortalama Mutlak Sapma
OMÖH : Ortalama Mutlak Ölçekli Hata
OMYH : Ortalama Mutlak Yüzde Hata
OYH : Ortalama Yüzde Hata
OYHKK : Ortalama Yüzde Hata Karelerinin Karekökü
PSO : Parçacık Sürüsü Optimizasyonu
RAU : Rastgele Alt Uzay
RP : Rasgele Projeksiyonlar
RO : Rastgele Orman
SAS : Sosyal Ağ Servisi
SRA : Sınıflandırma ve Regresyon Ağacı
ÜD : Üstel Düzeltme
YSA : Yapay Sinir Ağları

ÖZET

YÜKSEK LİSANS TEZİ

FİLM ENDÜSTRİSİNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE YENİ ÜRÜNLERİN GELİR TAHMİNİ

Özge Hüsniye NAMLI

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Endüstri Mühendisliği Anabilim Dalı

Danışman : Dr. Öğr. Üyesi Tuncay ÖZCAN

Film endüstrisi istikrarlı olarak büyümekte olan dinamik bir sektördür. Her yıl üretilen film sayısının sürekli olarak artış gösterdiği gerçeği göz önüne alındığında film sektörüne harcanan bütçenin ne kadar büyük miktarlarda olduğu görülmektedir. Sürekli gelişen teknoloji, değişen koşullar her sektörü etkilediği gibi film sektörünü de etkilemekte ve kaçınılmaz bir rekabet ortamı yaratmaktadır. Bu rekabet ortamı içerisinde işletmelerin doğru yatırım kararları vermeleri önemli bir problem haline gelmektedir. Bu noktada filmlerin gişe hasılatının ne olacağını önceden bilmek ve buna göre stratejik kararlar almak büyük önem arz etmektedir.

Bu çalışmada, filmlerin vizyona girmeden önce gişe hasılatının tahmin edilmesi için makine öğrenmesi esaslı sınıflandırma metodolojileri kullanılmıştır. Sınıflandırma modellerinin oluşturulmasında; makine öğrenmesi esaslı yöntemlerden Yapay Sinir Ağları ve Destek

Vektör Makinaları, bileşik makine öğrenmesi esaslı yöntemlerden ise Torbalama algoritması gişe hasılatının tahmin edilmesi probleminin çözümünde uygulanmıştır.

Çalışmada, hem bireysel sınıflandırma algoritmaları ile bileşik makine öğrenmesi algoritmasının sınıflandırma probleminin çözümündeki başarıları karşılaştırılmış hem de algoritmaların parametre değerlerinin eniyilenmesinin modellerin performansına etkisi araştırılmıştır. Sunulan üç modelin gişe hasılatının tahmini için birbirleriyle karşılaştırılabilir ve oldukça güçlü sonuçlar verdiği görülmüştür. Parametre eniyileme ise modellerin performansını önemli ölçüde iyileştirmiştir.

Haziran 2018, 92 sayfa.

Anahtar kelimeler: Gişe hasılatı, yapay sinir ağları, destek vektör makineleri, torbalama, parametre eniyileme.

SUMMARY

M.Sc. THESIS

REVENUE FORECASTING FOR NEW PRODUCTS WITH MACHINE LEARNING ALGORITHMS IN MOVIE INDUSTRY

Özge Hüsniye NAMLI

İstanbul University

Institute of Graduate Studies in Science and Engineering

Department of Industrial Engineering

Supervisor : Assist. Prof. Dr. Tuncay ÖZCAN

The movie industry is a dynamic industry that is growing steadily. Considering the fact that the number of films produced each year is increasing continuously, it is seen how large the amount of the total budget spent on the movie industry is. Constantly developing technology and changing conditions affect the movie industry just as it affects all other sectors and they create an inevitable competitive environment. Within this competitive environment, making investment decisions becomes a serious problem. At this point it is important to know in advance what will be the box office revenue of films and to make strategic decisions accordingly.

In this study, machine learning based classification methodologies were used to forecast box office revenue of movies before its theatrical release. In constructing prediction models,

Artificial Neural Networks and Support Vector Machines based-machine learning methods and Bagging algorithm based-ensemble machine learning methods were used to solve the problem of forecasting box-office revenue.

In this study, both the performance of the individual classification algorithms and ensemble machine learning algorithm were compared in solving the classification problem and the effect of optimizing the parameter values of the algorithms was investigated on the performance of the models. Proposed models were comparable to each other for forecasting box office revenue and showed very strong results. It was found that parameter optimization always has an increasing effect on the performance of the models.

June 2018,92pages.

Keywords: Box-office revenue, artificial neural networks, support vector machines, bagging, parameters optimization.

1. GİRİŞ

Temeli 1800'li yılların sonuna dayanan sinema filmleri, yıllar içerisinde gelişimini tamamlayarak; günümüzde yönetmenler, yapımcılar, senaristler, oyuncular, teknik ekipler, dağıtımcı firmalar, salonlar ve izleyicilerden oluşan geniş bir sektör haline gelmiştir.

Birçok alanı olduğu gibi gelişen teknoloji ve değişen dünya film endüstrisini de etkilemiştir. Üretimi sırasında pek çok yenilikçi işlem ve yaratıcı çalışma gerektiren filmler yüksek maliyetler oluşturmaktadır. Yıldız kadrosu, özel çekim efektleri, büyük dekorlar, pahalı kostümler gibi öğeler filmlerin maliyetini daha da yükseltmektedir. Maliyetlere bağlı olarak filmler için ayrılan bütçeler de çok yüksek rakamlara ulaşmaktadır. Bu doğrultuda, bir filme yatırım yapma ve gerekli bütçeyi ayırma kararı ciddi ve bir o kadar da stratejik bir karar haline gelmiştir.

Sinema filmi endüstrisindeki yatırımcıların, karlarını arttırmak için dağıtım aşamasında doğru kararlar vermeleri gerekmektedir (Sawhney & Eliashberg, 1996; Hur ve diğ., 2016). Karar verme sürecinde, sadece daha önce piyasaya sürülen filmler için değil, aynı zamanda vizyona girmeyi bekleyen filmler için de gişe hasılatını tahmin etmek büyük önem taşımaktadır (Jun ve diğ., 2011; Hur ve diğ., 2016).

Mevcut karar destek sistemlerinin yetersiz kaldığı ve yüksek güvenilirlikle sonuçlar veremediği tahmin problemlerinde, makine öğrenmesi algoritmaları kullanarak üretim öncesi güçlü tahminler yapmak dikkat çeken bir konu haline gelmiştir.

Bu çalışmada; yatırım kararlarına temel oluşturma amacıyla, vizyona girmeden önce filmlerin gelir tahmininin sınıflandırma algoritmaları ile yapılması amaçlanmıştır. Düşük hata oranları ve yüksek doğruluk oranı ile güçlü bir tahmin yapmak için makine öğrenmesi teknikleri kullanılarak tahmin modelleri oluşturulmuştur. Tahmin modeli oluşturmada, bireysel makine öğrenmesi metotlarından olan Yapay Sinir Ağları (YSA) ve Destek Vektör Makineleri (DVM) algoritmaları ile bileşik makine öğrenmesi algoritmalarından biri olan Torbalama (Bagging) algoritması kullanılmıştır. Oluşturulan tahmin modellerinin doğruluğunu iyileştirmek için modellerin parametreleri eniyilenmeye çalışılmıştır. Çalışmada, hem parametre eniyilemenin modelin tahmin başarısına etkisini gözlemlemek hem de YSA, DVM

ve Torbalama algoritmasının başarısını karşılaştırmak hedeflenmiştir. Tahmin modellerinin karşılaştırılmasında; Doğruluk oranı, Kesinlik, Anma, F-Ölçütü ve Kappa istatistiği gibi performans ölçütleri kullanılmıştır. Algoritmaların uygulanmasında; 10 kat çapraz geçerleme, 5 kat çapraz geçerleme, %80 eğitim-%20 test ve %66,6 eğitim-%33,3 test yöntemleri ile eğitim ve test verileri oluşturulmuştur.

Çalışmanın bundan sonraki bölümlerinin gelişimi aşağıdaki gibi özetlenebilir:

Genel kısımların yer aldığı ikinci bölümde; ilk olarak veri madenciliği kavramı, veri madenciliği süreci ve veri madenciliği fonksiyonlarından özetle bahsedilmiş ve çalışmanın ana konusunu oluşturan sınıflandırma fonksiyonu ayrıntıları ile ele alınmıştır. Daha sonra, gişe hasılatı tahmin problemi için tez kapsamında uygulanacak bireysel sınıflandırma algoritmalarından YSA ve DVM, bileşik makine öğrenmesi algoritmalarından Torbalama algoritmasının metodolojisi detaylı olarak anlatılmıştır. Model geçerleme yöntemleri olarak çalışmada kullanılan belirli oranlarda bölme (Hold-out) ve k-kat çapraz geçerleme metotlarından kısaca bahsedilmiş ve modellerin performanslarını değerlendirmek için kullanılacak performans değerlendirme ölçülerine yer verilmiştir. Son olarak, gişe hasılatı tahmin problemine ve tez çalışmasında kullanılan sınıflandırma ve tahmin algoritmalarının yer aldığı çalışmalara ait literatür incelemesi sunulmuştur.

Malzeme ve Yöntemin yer aldığı üçüncü bölümde, ilk olarak gişe hasılatı tahmin problemi anlatılmıştır. Daha sonra, problem için tahmin modellerini uygulamak üzere veri setinin nasıl oluşturulduğundan bahsedilmiş ve filmlerin vizyon öncesi gişe hasılatının tahmini için yapay zeka metotlarından olan YSA ve DVM, ve topluluk makine öğrenmesi algoritmalarından biri olan torbalama algoritması ile oluşturulan tahmin modelleri 10 kat çapraz geçerleme, 5 kat çapraz geçerleme, %80'e %20 ayırma ve %66,6'ya %33,3 ayırma metotları ile test ve eğitim seti olarak ayrılan veri üzerine uygulanmıştır. Son olarak, uygulanan bu tahmin modellerinin başarısını arttırmak amacıyla Izgara Arama (Grid Search) algoritması kullanılarak tahmin modellerinin parametre değerleri eniyilenmeye çalışılmıştır.

Bulguların yer aldığı dördüncü bölümde, uygulanan tahmin modellerinin performans değerleri sunulmuştur. Parametre eniyilemenin modelin tahmin başarısına etkisi incelenmiş ve geleneksel yapay zeka algoritmaları ile topluluk makine öğrenmesi algoritmasının tahmin başarısı gişe hasılatının erken tahmini problemi için karşılaştırılmıştır. Tahmin sonuçlarının

karşılaştırılmasında; Doğruluk oranı, Kesinlik, Anma, F-Ölçütü ve Kappa istatistiği gibi performans ölçütleri kullanılmıştır.

Çalışmanın son kısmı olan, tartışma ve sonucun yer aldığı beşinci bölümde ise; çalışmanın sonuçları özetlenmiş, bu çalışmanın literatüre katkılarından bahsedilmiş ve ele alınan problemle ilgili olarak geleceğe yönelik yapılabilecek farklı çalışmalar ile ilgili önerilere ve değerlendirmelere yer verilmiştir.

2. GENEL KISIMLAR

Bu bölümde, ilk olarak veri madenciliği kavramı ve süreci ile veri madenciliği fonksiyonlarından kısaca bahsedilmiş ve çalışmanın temelini oluşturan sınıflandırma fonksiyonu detaylı olarak incelenmiştir. Daha sonra, gişe hasılatı tahmin problemi için uygulanan YSA, DVM ve Torbalama algoritmaları hesaplama adımları ile birlikte detaylı olarak analiz edilmiştir. Sınıflandırma problemlerinin çözümü için kullanılan model geçerleme yöntemlerine ve modellerin performanslarını değerlendirmek için kullanılacak performans değerlendirme ölçülerine yer verilmiştir. Son olarak, gişe hasılatı tahmin problemine ve tez çalışmasında kullanılan sınıflandırma ve tahmin algoritmalarının yer aldığı çalışmalara ait literatür incelemesi sunulmuştur.

2.1.VERİ MADENCİLİĞİ

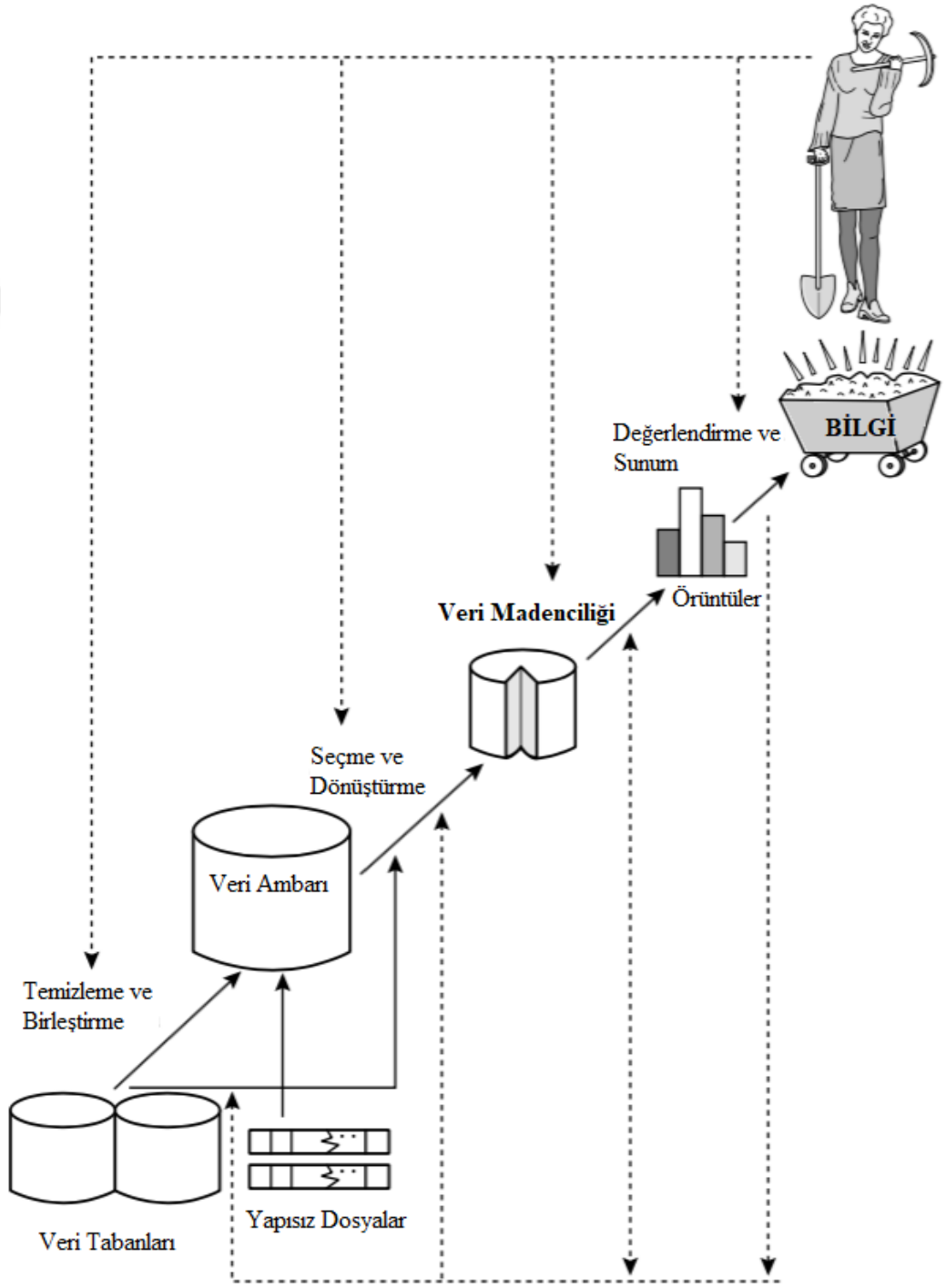
Veri madenciliği, büyük veri setlerinde ilginç, beklenmedik veya değerli yapıların keşfini amaçlamaktadır (Hand, 2007). Bu noktada veri madenciliği birçok disiplinin kavşak noktasında yer almaktadır.

2.1.1. Veri Madenciliği Tanımı

Veri madenciliği, büyük veri tabanlarında bulunan ve daha önce bilinmeyen yararlı bilgiyi elde etmek için kullanılan güçlü bir bilgi teknolojisi aracıdır. Veri madenciliği, ham veriler arasındaki ilişkiyi ve veri içerisindeki örüntüleri işleme sürecini otomatik hale getirir ve karar destek sisteminde kullanılabilecek veya karar vericiler tarafından değerlendirilebilecek sonuçlar sunmaktadır (Fernandez, 2010). Veri madenciliğinde; doğru bir analiz veya desen çıkarımı için verilerin temiz olması gerekmektedir. Temizleme ve önişleme sürecinde veri setini kusurlu, gürültülü veya eksik verilerden arındırmak için birkaç adımı içermektedir. Bu noktada, verilerden doğru ve güvenilir bilgi elde etmek için uygulanması gereken adımlar; özellik seçimi, özellik çıkarımı ve kullanılan yöntemlerin gerekliliklerine göre verilerin hazırlanması olarak sıralanabilir (Piramuthu, 2004).

2.1.2. Veri Madenciliği Süreci

Veriden, bilgi elde edilmesi için gereken işlem adımlarını içeren bilgi keşfi süreci Şekil 2.1'de gösterilmektedir (Han ve Kamber, 2006).



Şekil 2.1: Bilgi keşfi süreci.

Veri madenciliği aşağıdaki temel adımları içermektedir (Gupta ve diğ., 2011):

- Problem tanımı: İlk adım hedefleri tanımlamaktır. Tanımlanan hedefe dayanarak, ilgili davranış modelini oluşturmada veriler için uygun araçlar seçilir.
- Veri araştırması: Verilerin kalitesi doğru bir model için uygun değilse, o zaman gelecekteki veri toplama ve depolama stratejileri ile ilgili öneriler yapılabilir. Analiz için, tüm verilerin tutarlı bir şekilde ele alınıp birleştirilmesi gerekir.
- Veri hazırlama: Bu adımın amacı, verileri temizlemek ve dönüştürmektir. Böylece eksik ve geçersiz değerlerin işlenmesi ve bilinen tüm geçerli değerlerin daha sağlam analizler için tutarlı hale getirilmesi sağlanmış olur.
- Modelleme: Verilere ve istenen sonuçlara dayanarak, analiz için bir veri madenciliği algoritması veya algoritmalarının kombinasyonu seçilir. Bu algoritmalar; istatistikler, komşuluklar ve kümeleme gibi klasik teknikleri, ayrıca karar ağaçları, ağlar ve kural tabanlı algoritmalar gibi yeni nesil teknikleri içerir. Özel algoritmalar, hedefe ve analiz edilecek verilerin kalitesine göre seçilir.
- Değerlendirme: Veri madenciliği algoritmalarının sonuçlarına göre, analizden elde edilen önemli sonuçları belirlemek ve değerlendirilmek üzere bir dizi öneri oluşturmak için analiz yapılır.

2.1.3. Veri Madenciliğinin İşlevleri

Veri madenciliğinin sınıflandırma, kümeleme ve birliktelik kuralları ve ilişki analizi olmak üzere üç temel işlevi bulunmaktadır. Borzemski (2007) bu üç temel işlevi kısaca şöyle açıklamıştır. Sınıflandırmada, önceden tanımlanmış kurallara göre önceden tanımlanmış kategorilere veri atanır. Kümelemede, bazı yakınlık ölçümleri kullanılarak verilerin özelliklerinin analizi yoluyla veri kümeleri tanımlanır, ancak önceden belirlenmiş kümeler yoktur. Kümeler, verilerde tanımlanan modellerle oluşturulur. Birliktelik kuralları ve ilişki analizinde ise gelecekteki eğilimlerin tahmin edilmesine olanak tanımak için kuralları belirleyen tanımlayıcı modeller oluşturulur.

2.1.3.1. Kümeleme Analizi

Kümeleme; aynı aileden geldiği sınıflandırma ve regresyondan, tahmin edilecek önceden belirlenmiş bir hedef özelliğinin olmamasıyla farklılaşan bir tümevarımsal öğrenme aracıdır.

Önceden tanımlanmış sınıflardan ziyade, verilerde tanımlanan benzerlik modellerine dayanan kendi kendine keşfedilmiş sınıflandırma olarak düşünülebilir (Cichosz, 2015).

Kümeleme, benzer örnek veya nesnelerin gruplandırılması olduğundan, iki nesnenin benzer olup olmadığını belirleyen bir çeşit ölçü gereklidir. Benzerlik ilişkisini tahmin etmek için kullanılan iki temel ölçü türü vardır. Bunlar; mesafe ölçüleri ve benzerlik ölçüleridir (Maimon & Rokach, 2010).

Kümeleme algoritmalarının temelinde kullanılan veri ve uzaklık matrisleri aşağıdaki gibidir (Han & Kamber, 2006):

Veri matrisi:

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{ip} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Uzaklık matrisi:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & 0 & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Nesneler arasındaki uzaklığı bulmak için farklı uzaklık ölçüleri kullanılabilir. Uzaklık ölçütü yerine benzerlik ölçütü kullanılarak da kümeleme gerçekleştirilebilir.

Literatüre baktığımızda birçok kümeleme algoritması geliştirildiğini görüyoruz. Bunlar; bölümlemeli yöntemler (k-ortalamalar, k-medoids), aşamalı yöntemler, yoğunluk bazlı yöntemler, ızgara bazlı yöntemler, model bazlı yöntemler (KÖH), yüksek boyutlu veriler için yöntemler ve kısıtlamaya dayalı yöntemler olarak sınıflandırılabilir. Bazı algoritmalar birden fazla kategoriye ait olabilir (Han & Kamber, 2006).

Kümeleme algoritmalarının uygulama alanları olarak; pazarlama, ekonomi bilimi, biyoloji, kütüphaneler, sigorta, şehir planlama, deprem çalışmaları, web ve belge kümeleme alanları sayılabilir (Madhulatha, 2012).

2.1.3.2. Birliktelik Kuralları ve İlişki Analizi

Birliktelik kuralları, verilerdeki birliktelik ilişkilerini bulmayı amaçlayan bir dizi tekniktir. Bu algoritmaların en yaygın uygulaması, perakende işlem verilerinin analizidir. Bu analiz; X ürününü satın alan bir müşterinin, aynı zamanda Y ürünü satın alması olasılığını bulmayı amaçlamaktadır. Birliktelik kuralları algoritmaları, sıralı desenleri bulmak için de formüle edilebilir. İlişkilendirme analizi için genellikle veri hacimleri çok büyük olan işlem verilerine ihtiyaç duyulur (García ve diğ., 2015).

İlişkisel veri tabanlarındaki birliktelik kuralları, bazı niteliklerin değerlerini, aynı değişken grubundaki diğer bazı niteliklerin değerlerinin varlığı ile ilişkilendirir. $[A = a] \Rightarrow [B = b]$ kuralı, A niteliği bir değişken grubunda bir değer aldığında, B'nin aynı değişken grubunda b değerini aldığını söyler. İlişkilendirme kurallarının doğruluğu ve önemi, genellikle sırasıyla güven ve destek olarak adlandırılan iki olasılık ölçüsü ile tahmin edilir (Delgado ve diğ., 2001).

Genel olarak, bir ilişkilendirme kuralları madenciliği algoritması aşağıdaki adımları içerir (Kotsiantis & Kanellopoulos, 2006):

- Aday k-öğe setlerinin kümesi, önceki iterasyonda üretilen yaygın (k -1) – öğe setlerinin 1-uzantıları tarafından üretilir.
- Aday k-öğe setleri için destek, veri tabanı üzerinde bir geçiş ile oluşturulur.
- Minimum desteğe sahip olmayan öğeler atılır ve kalan öğeler yaygın k-öğe setleri olarak adlandırılır.

Bu işlem, yaygın öğe setleri bulunmayana kadar tekrarlanır. Daha sonra, yaygın öğe setlerinin ikili bölünmesi ile elde edilen aday birliktelik kurallarının minimum güven koşulunu sağlama durumu denetlenerek birliktelik kuralları elde edilir.

Birliktelik kuralı madenciliğinde kullanılan temel algoritmalar Apriori, FP-büyüme gibi algoritmalar veri tabanının taranma sayısını azaltarak birliktelik kurallarını makul sürede elde edilmesini amaçlamaktadır.

2.1.3.3.Sınıflandırma

Sınıflandırma probleminde, kategoriler veya sınıflar olarak adlandırılan, gruplara ayrılan örnek veri kümesinin yapısını öğrenmek amaçlanmaktadır. Bu kategorilerin öğrenilmesi bir model ile sağlanır. Veri örneğinin grup tanımlayıcılarını (sınıf etiketlerini) tahmin etmek için kullanılan bu modelde, sınıflandırma probleminin girdilerinden biri, zaten farklı sınıflara ayrılmış örnek veri kümesidir. Sınıfların grup tanımlayıcıları, sınıf etiketleri ve öğrenilen model, eğitim modeli olarak adlandırılır. Sınıflandırılması gereken daha önce görülmemiş veri noktaları da, toplu olarak test veri seti olarak adlandırılır (Aggarwal, 2015). Sınıflandırma modelleri, sınıf üyeliği olasılığını, tahmin edici değişkenlerin bir fonksiyonu olarak tahmin etmek için nominal cevap kullanırlar. Genel olarak, sınıflandırma modelleri, nominal cevap ile öngörücü değişkenler arasındaki ilişkinin şeklinin belirlemesine ve kategorik yanıtla anlamlı bir şekilde ilişkili olan öngörücü değişkenlerin daha fazla araştırılmasına olanak sağlarlar (Fernandez, 2010).

Sınıflandırma, veri madenciliği de dahil olmak üzere çeşitli alanlarda yaygın olarak kullanılan bir tekniktir. Amaç, bir dizi nitelik tarafından tanımlanan geniş bir nesneyi, öğrenme yöntemlerini kullanarak tanımlanmış sınıflara ayırmaktır. Hem iş hem de bilimsel veri tabanlarının giderek büyümesi nedeniyle, bu tür veri tabanlarından etkin sınıflandırma kuralları çıkarılması büyük önem taşımaktadır (Mastrogiannis ve diğ., 2009). Verilerin sınıflandırılması veya tahmin edilmesine hazırlık olarak veri önişleme; gürültüyü azaltmak veya eksik değerleri ele almak için veri temizliğini, alakasız veya gereksiz niteliklerin kaldırılması için uygunluk analizini ve verilerin daha üst düzey kavramlara genelleştirilmesi veya verilerin dönüştürülmesi süreçlerini içerebilir (Han & Kamber, 2006). Her sınıflandırıcının tahmin kalitesi problemine bağlıdır. Girdi verilerinin iyi hazırlanması ve sınıflandırıcı parametrelerinin enyilenmesinden sonra, farklı sınıflandırma teknikleri çoğu zaman benzer sonuçlar vermektedir. Çok büyük veri kümeleri için "mükemmel" sınıflandırıcı ayrı bir özellik seçim sürecine ihtiyaç duymaz ve yeterince hızlıdır (Matwin & Mielniczuk, 2016).

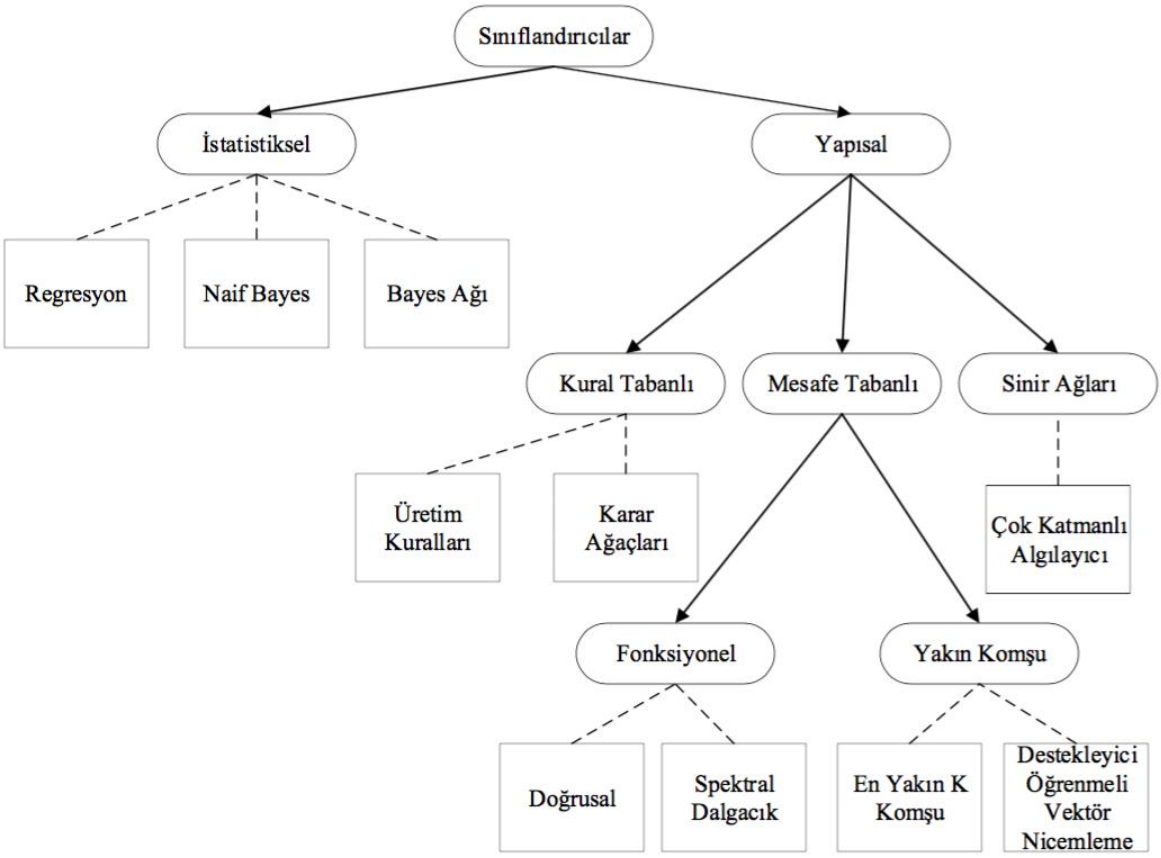
Veri madenciliği teknikleri, veri türüne ve sonuçların kullanımın amacına bağlı olarak tahmin edici ve tanımlayıcı modellere ayrılabilir. Tanımlayıcı modeller; kümeleme, özetleme, birliktelik kuralları, sıralı diziler gibi veri madenciliği tekniklerini kullanır. Tahmin edici

modeller ise sınıflandırma, eğri uydurma, zaman serileri gibi veri madenciliği tekniklerini kullanmaktadır. (Sarıman, 2011).

Sınıflandırma algoritmalarının çoğu genellikle iki aşamadan oluşur (Aggarwal, 2015):

- Eğitim aşaması: Bu aşamada, eğitim örneklerinden bir eğitim modeli oluşturulur. Bu model nitelik alanlarına göre sınıf etiketinin tahmini için kurallar kümesini içerir.
- Test aşaması: Bu aşamada; eğitim modeli, bir veya daha fazla görülmeyen test örneğinin sınıf etiketini belirlemek için kullanılır.

Veri madenciliğinde kullanılan sınıflandırma algoritmaları Şekil 2.2'de tasnif edilmiştir (Marmanis & Babenko, 2009).



Şekil 2.2: Sınıflandırma algoritmalarına genel bakış.

2.2.MAKİNE ÖĞRENMESİ

Bu alt bölümde, çalışmada önerilen yöntemlerin temelini oluşturan makine öğrenmesi ve bileşik makine öğrenmesi kavramlarından bahsedilmiştir.

2.2.1. Makine Öğrenmesi Tanımı

Makine öğrenmesi, hem yeni tekniklerin ve teorik algoritmaların sunulmasında hem de gerçek hayattaki problemlere uygulanmasında en verimli araştırma alanlarından biridir. Akıllı bir şekilde çevreye uyarlanabilen sistemlerin geliştirilmesi geniş ve pratik bir uygulama alanına sahiptir. Genellikle, bu sistemler; belirli bir algoritma ya da tekniğin performansını eniyilemekle birlikte, klasik anlamda bir "programdan" ziyade, deneysel verileri kullanarak çalışırlar (Olivas ve diğ, 2010). Makine öğrenmesi, birçok insanın ciddiye almaya başladığı bir teknoloji olan veri madenciliği için gelişen yeni bir teknolojidir. Makine öğrenmesi, veriden yapı çıkaran ve bu yapıyı doğrulayan algoritmaları konu alır. Bu algoritmalar anlaşılması zor ve karmaşık olmamalarına rağmen, tamamen açık ve kolay algoritmalar da değillerdir (Witten ve diğ., 2011).

Makine öğrenmesi, birbiriyle ilişkili iki soru üzerine odaklanmış bir disiplindir: Bu sorulardan birisi; "Deneyimle otomatik olarak gelişen bilgisayar sistemleri nasıl oluşturulabilir?", diğeri ise; "Bilgisayarlar, insanlar ve organizasyonlar dahil olmak üzere tüm öğrenim sistemlerini yöneten temel istatistik, hesaplama, bilgi ve teori kuralları nelerdir?". Makine öğreniminin incelenmesi, hem bu temel bilimsel ve mühendislik sorularının ele alınması hem de birçok uygulamada ürettiği ve uyguladığı bilgisayar yazılımları açısından oldukça önemlidir (Jordan & Mitchell, 2015).

Makine Öğrenmesinin gerçek hayat uygulamalarına aşağıdaki örnekler verilebilir (Mitchell, 2006);

• Konuşma/Ses tanıma: Konuşma tanıma uygulamalarında, konuşmayı tanımada sistemi eğitmek için makine öğrenimini kullanılmaktadır.

• Bilgisayarla görme: Yüz tanıma sistemlerinden hücrelerin mikroskop görüntülerini otomatik olarak sınıflandıran sistemlere kadar pek çok mevcut görüş sistemi makine öğrenimi kullanılarak geliştirilmektedir.

• Biyolojik gözetim: Hastalık salgınlarını tespit etmek ve takip etmek için çeşitli sağlık kurumları makine öğrenmesini kullanmaktadır.

• Robot kontrolü: Robot sistemlerinde kontrol stratejileri elde etmek için makine öğrenmesi kullanılmaktadır.

• Ampirik bilimleri hızlandırma: Birçok veriye dayalı disiplinde, bilimsel keşif sürecine yardımcı olmak için makine öğrenmesi yöntemlerinden faydalanılmaktadır.

2.2.2. Bileşik Makine Öğrenmesi Tanımı

Bileşik makine öğrenmesi; bireysel kararların yeni örnekleri sınıflandırmak için bir şekilde birleştirildiği sınıflandırıcı kümesinin bir parçası olarak tanımlanabilir (Dietterich, 2000; Tan ve diğ., 2003). Bileşik makine öğrenmesinin amacı, bir dizi "pratik yöntemin" daha doğru modeller için bir toplam sınıf tahmin kuralıyla birleştirilmesidir (Srimani & Koti, 2013).

Bileşik makine öğrenmesi, bir sınıflandırıcı grubu oluşturup daha sonra onların tahminlerinin sonucunu alarak yeni veri noktalarını sınıflandırmayı amaçlamaktadır. Her bir sınıflandırıcı bağımsız hatalar yaptığı için, çoklu sınıflandırıcının toplanması sınıflandırma doğruluğunu arttırabilir (He & Chang, 2004).

Sınıflandırma toplulukları beş popüler yaklaşım ile oluşturulabilir (Srimani & Koti, 2013):

1. İlk yaklaşımda, eğitim veri noktalarının dağılımının değiştirilmesinin sonucu, eğitim setlerinin farklı bir örneğini kullanarak topluluk içinde bir sınıflandırıcı oluşturur. Genellikle sıklıkla bu yaklaşım kullanılır ve herhangi bir sınıflandırıcı ile çalışır. Örnek olarak, Torbalama ve Hızlandırma (Boosting) algoritmaları verilebilir.

2. İkinci yaklaşımda, eğitim setindeki niteliklerin değiştirilmesinin sonucunda, veri kümesinin nitelik alanını yönetir. Ayrıca, her bir sınıflandırıcının eğitimi farklı nitelik kümeleri üzerinden yapılır. Bu kümeler, yeni oluşturulmuş niteliklerden veya eğitim verisinden olabilir. Rastgele Alt Uzay (Random Subspace) ve Rotasyon Ormanları (Rotation Forests) bu yaklaşıma ait örneklerden bazılarıdır.

3. Üçüncü yaklaşımda, farklı veri kümeleri oluşturmak için eğitimin çıktısı kullanılır.

4. Bu yaklaşım çoklu sınıf problemleri için son derece kullanışlıdır ve "Hata düzeltme kodları yaklaşımı" olarak adlandırılır.

5. Bu teknik oldukça popülerdir ve ayrım kriterini eniyileyerek, ayırıcı niteliklerin ve ayrım noktalarının seçilmesinde karar ağacı toplulukları oluşturmak için yaygın olarak kullanılır.

2.3.SINIFLANDIRMA

Filmlerin vizyon öncesi gişe hasılatının tahmini için önerilen yapay zeka metotlarından olan YSA ve DVM ve topluluk makine öğrenmesi algoritmalarından torbalamanın metodolojisi bu bölümde incelenmiştir.

2.3.1. Bireysel Sınıflandırma Algoritmaları

Tahmin modeli oluşturmada, bireysel sınıflandırma algoritmaları olarak Yapay Sinir Ağları (YSA) ve Destek Vektör Makineleri algoritmaları (DVM) kullanılmıştır.

2.3.1.1. Yapay Sinir Ağları

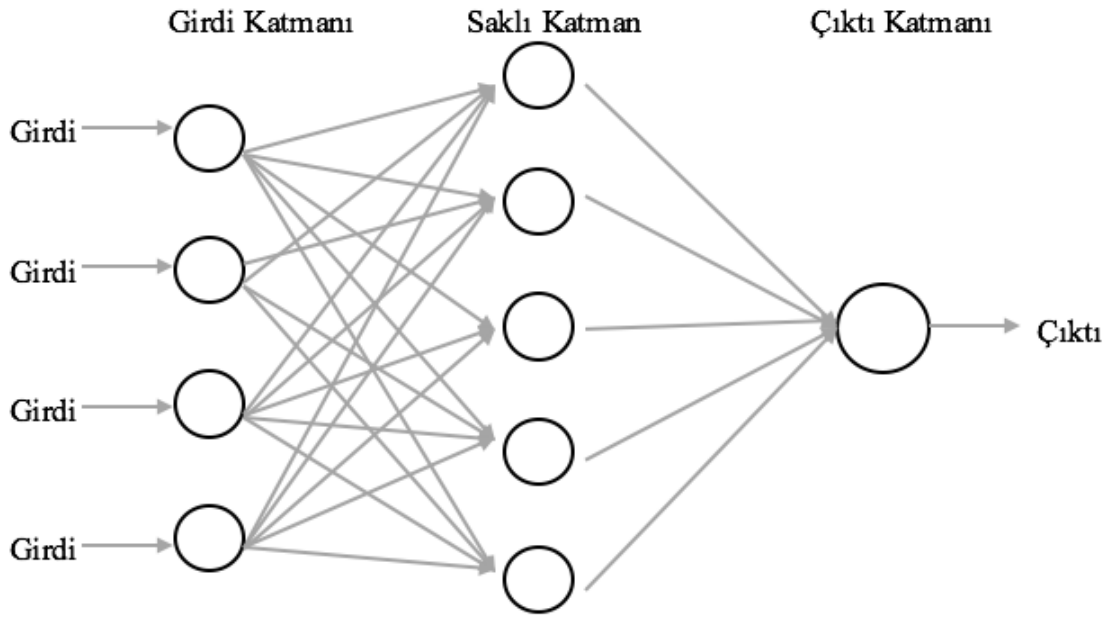
Sinir ağları yapay nöronlardan inşa edilmiştir ve kavramsal olarak, girdi değişkenlerinin her biri bir girdi nöronuna atanır ve her bir çıktı değişkeni bir çıktı nöronuna atanır. Girdi ve çıktı nöronları arasındakilere ise gizli nöronlar denir (Chakrabarti ve diğ., 2009).YSA algoritmasının kullanımı bağımlı ve bağımsız değişkenler arasındaki ilişkinin tanımının bilinmediği veya çok karmaşık olduğu uygulamalar için giderek artan bir popülerlik kazanmıştır (Remzi & Djavan, 2004).

Sinir Ağı, birbirine bağlı bir ağda birleşmiş bir dizi işlem birimidir ve biyolojik sinir ağının bazı özelliklerini sergileyen zengin bir yapısı vardır. Sinir ağının yapısı, kullanıcıya her katman düzeyinde paralel konsept uygulama fırsatı sunar. YSA'nın bir başka önemli özelliği olarak da hata toleransının olması söylenebilir. YSA algoritmaları, bilginin gürültülü ve belirsiz olduğu durumlarda kullanılmaya uygundur (Rani, 2011).

YSA sistemlerinin en kullanışlı özelliği, daha önce işlenmemiş olan bir girdi verisini doğru bir şekilde sınıflandırma yeteneğidir (Downes & Hartley, 2006).Özetle YSA, kendi kendine öğrenme becerisi olan, hataya dayanıklı, birleştirici hafızaya, yüksek paralelliğe ve karmaşık doğrusal olmayan süreçlerin üstesinden gelebilme kabiliyetine sahiptir. YSA, elde edilen

bilgiye ve problemlerin ele alınma deneyimlerine göre karmaşık sorulara rasyonel kararlar verebilir (Benfu ve diğ., 2009). YSA tahmin etme, sınıflandırma, sinyal işleme ve örüntü tanıma gibi farklı birçok alanda kullanılan en güçlü makine öğrenme algoritmalarından bir tanesidir.

Yapay Sinir ağlarının yapısı Şekil 2.3'de gösterilmiştir.



Şekil 2.3: Yapay Sinir Ağlarının yapısı.

YSA modeli aşağıdaki gibi açıklanmıştır (Erdal ve diğ., 2013):

n. katmandaki l. nöron için çıkış sinyali aşağıdaki eşitlik ile ifade edilebilir:

$$y_l^n(t) = \varphi\left[\sum_{j=1}^{p} w_{lj}^n(t) y_j^{n-1}(t) + \Psi_l^n\right] \tag{2.1}$$

Burada $\varphi(\cdot)$ aktivasyon fonksiyonunu, w_{lj}^n bağlantı ağırlığını, t zaman indisini ve $\Psi_l^n = w_{l_o}^n(t)$ ağırlıkları göstermektedir.

n- ağ katmanı için sinaptik ağırlık $w_{ji}^{n}(t)$ aşağıdaki gibi hesaplanabilir:

$$w_{ji}^{n}(t+1) = w_{ji}^{n}(t) + \Delta w_{ji}^{n}(t) \tag{2.2}$$

$l \leq n \leq N$ koşulu ile aşağıdaki şekilde revize edilebilir:

$$\Delta w_{ji}^{n}(t) = \eta \lambda_{j}^{n}(t) y_{i}^{n-1}(t) \tag{2.3}$$

$0 < \eta < 1$ koşulu ile η öğrenme hızını ve $\lambda_{j}^{n}(t) \equiv -\frac{\partial E_t}{\partial u_{j}^{n}}$ yerel hata eğrisini belirtmektedir. Geri yayılım algoritmasını ilerletmek için, momentum terimi $\propto$ eklenerek aşağıdaki eşitlik elde edilmektedir:

$$\Delta w_{ji}^{n}(t) = \eta \lambda_{j}^{n}(t) y_{i}^{n-1}(t) + \propto \Delta w_{ji}^{n}(t-1) \tag{2.4}$$

$0 < \propto < 1$ koşulunda, çıktı katmanı için yerel hata gradyanı aşağıdaki eşitlik ile hesaplanmaktadır:

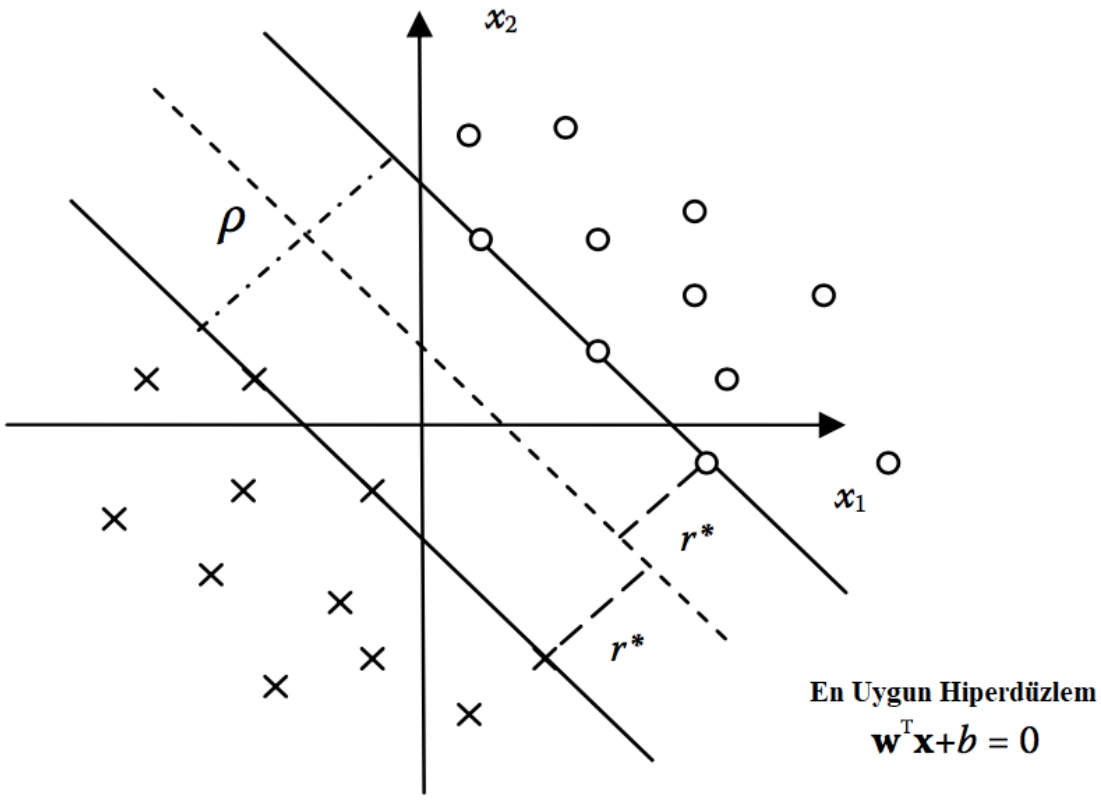
$$\lambda_{j}^{N}(t) = \left[d_j(t) - y_{j}^{N}(t)\right] \varphi\left[u_{j}^{N}(t)\right] \equiv e_j(t)\, \varphi\left[u_{j}^{N}(t)\right] \tag{2.5}$$

Burada $d_j(t)$ hedef çıkış sinyalini ve $\varphi(\cdot)$ aktivasyon fonksiyonunubelirtmektedir.

2.3.1.2. Destek Vektör Makinaları

Destek Vektör Makineleri sınıflandırmada sıklıkla kullanılan ve başarılı sonuçlar veren makine öğrenmesi algoritmalarından biridir. Bu sınıflandırma yöntemi, verinin daha yüksek bir boyuta dönüşmesiyle oluşturacağı hiperdüzlem ile iki sınıfı birbirinden ayırma prensibini temel almaktadır ve burada yüksek boyuta dönüşüm esnasında çekirdek fonksiyonları adı verilen farklı özelliklere sahip fonksiyonlar kullanılır (Kavzoğlu & Çölkesen, 2010). En hızlı DVM'lerin bile eğitim sürelerinin oldukça yavaş olmasına rağmen, DVM'ler karmaşık doğrusal olmayan karar sınırlarını modelleme yeteneklerinden dolayı oldukça doğru sonuçlar üretmektedirler ve ezberleme eğilimleri diğer yöntemlere göre daha azdır (Han & Kamber, 2006).

Doğrusal ayrılabilen veriler için Şekil 2.4'de destek vektörleri ve en uygun hiperdüzlem gösterilmiştir (Wu & Kumar, 2009).



Şekil 2.4: Doğrusal ayrılabilen veriler için DVM.

DVM algoritması aşağıdaki gibi işlemektedir (Chou ve diğ., 2014):

Uzayda doğrusal model olan $f(x,\omega)$ aşağıdaki gibi formüle edilmiştir:

$$f(x,\omega)=\sum_{j=1}^{n} w_j g_j(x)+b \tag{2.6}$$

Bu eşitlikte $g_j(x)$ girdi alanından bir dizi doğrusal olmayan dönüşümü ifade eder, b, bir önyargı terimi olarak kullanılır ve w, düzenli risk fonksiyonunu en aza indirgeyerek tahmin edilen ağırlık vektörünü ifade eder.

Aşağıdaki kayıp fonksiyonu L_ε, tahmin kalitesini ölçmek için kullanılır.

$$L_\varepsilon = L_\varepsilon\big(y, f(x,\omega)\big) = \begin{cases} 0, & if|y - f(x,\omega)| \le \varepsilon \\ |y - f(x,\omega)|, & diğer \end{cases} \tag{2.7}$$

SVM ile ε-duyarsız kayıp fonksiyonu, yeni yüksek boyutlu özellik alanı için doğrusal bir regresyon fonksiyonu hesaplar ve aynı zamanda $\|\omega\|^2$ 'nin en düşük düzeye indirgenmesiyle modelin karmaşıklığını azaltır. Bu işlev negatif olmayan boşluk değişkenleri $\xi_i$ve ξ_i^*içerir. Burada, ε-duyarsız bölgeden alınan eğitim örneklerini tanımlamak için $i = 1, \dots, n$ kullanılır.

$$min\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2.8}$$

$$koşul \begin{cases} y_i - f(x_i,\omega) \le \varepsilon + \xi_i^* \\ f(x_i,\omega) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, \dots, n \end{cases} \tag{2.9}$$

Aşağıdaki formül, eniyileme problemini ikili (dual) probleme dönüştürmek için kullanılabilir.

$$f(x) = \sum_{i=1}^{n_{SV}}(\propto_i - \propto_i^*)K(x, x_i) \;\; koşulu \; 0 \le \propto_i^* \le C, 0 \le \propto_i \le C \tag{2.10}$$

Burada, n_{SV}destek vektörlerinin sayısını belirtir. Çekirdek fonksiyonu aşağıdaki gibidir:

$$K(x, x_i) = \sum_{i=1}^{m} g_i(x) g_i(x_i) \tag{2.11}$$

Doğrusal, radyal tabaka, polinom veya sigmoid fonksiyonu gibi seçilmiş DVM çekirdek fonksiyonları eğitim sırasında fonksiyon yüzeyi boyunca destek vektörlerini tanımlamak için kullanılır.

Çekirdek fonksiyonlarındaki çekirdek parametreleri, tahmin doğruluğunu arttırmak için eniyilenmelidir.

Dört temel çekirdek fonksiyonu aşağıdaki gibidir;

Doğrusal:

$$K(x_i, x_j) = x_i^T x_j \tag{2.12}$$

Polinom:

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \qquad \gamma > 0 \tag{2.13}$$

Radyal Temelli Fonksiyon:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \qquad \gamma > 0 \tag{2.14}$$

Sigmoid:

$$K(x_i, x_j) = tanh(\gamma x_i^T x_j + r) \tag{2.15}$$

Çekirdek fonksiyonlarında bulunan γ, r ve d çekirdek parametreleridir.

Genelde mantıklı bir seçenek olan radyal temelli fonksiyon çekirdeği, doğrusal olmayan örnekleri daha yüksek boyutlu bir alana eşler; dolayısıyla doğrusal çekirdeğin aksine, sınıf etiketleri ve nitelikler arasındaki ilişki doğrusal olmadığında da başarılı sonuçlar verebilir (Hsu ve diğ., 2003).

Sıklıkla kullanılan çekirdek fonksiyonların performanslarının karşılaştırılması Tablo 2.1'de sunulmuştur (Liu ve diğ., 2014):

Tablo 2.1: DVM'deki çekirdek fonksiyonların performansları.

Çekirdek fonksiyonları	Uygulama alanı	Parametrelere duyarlılık
Doğrusal	Genel durum	Duyarsızlık
Polinom	Genel durum	Durağan olmayan
Gauss	Yerellik	σ'ya duyarlı
Üssel	Yerellik	σ'ya duyarlı
Sigmoid	Genel durum	Veri boyutuna duyarlı
Çok noktalı	Yerellik	Sabit
Dairesel	Genel durum	Sabit

Uygulama kümesine ve parametrelere duyarlılığa göre farklı çekirdek işlevleri makul bir şekilde bir araya getirilerek bir süper çekirdek işlevi oluşturulursa veri seti için çok daha iyi bir sınıflandırma başarısı elde edilebilir (Liu ve diğ., 2014).

2.3.2. Bileşik Makine Öğrenmesi Algoritmaları

Tahmin modeli oluşturmada, topluluk makine öğrenmesi algoritmalarından biri olan Torbalama algoritması kullanılmıştır.

2.3.2.1. Torbalama Yöntemi

Torbalama (Bagging) yöntemi, Breiman (1996) tarafından önerilen ve yaygın olarak kullanılan topluluk makine öğrenmesi algoritmalarından bir tanesidir. Torbalama, farklı önyükleme örnekleri tarafından üretilen sınıflandırıcıları oylama mantığına dayanmaktadır (Bauer ve Kohavi, 1999). Çoklu sınıflandırıcılar için oylama yöntemi olan Torbalama algoritmasında, nihai sonuç 1∼N alt sınıflandırıcıların sonuçlarından elde edilir (Nishimura ve diğ., 1999).

En iyi bilinen bağımsız yöntem olan torbalama, öğrenilmiş sınıflandırıcıların çeşitli çıktılarını tek bir tahmin halinde birleştirerek, geliştirilmiş bir birleştirilmiş sınıflandırıcı (I^*) oluşturup doğruluğu artırmayı amaçlamaktadır (Rokach & Maimon, 2015).

Topluluk algoritmalarından biri olan torbalama modelinin adımları aşağıdaki gibi tanımlanmıştır (Han ve Kamber, 2006):

Girdiler:

D: eğitim örneği kümesi
k: topluluktaki model sayısı
temel öğrenme modeli (Örneğin; Karar ağaçları algoritmaları, sinir ağları vb.)

Çıktı:

M:* Topluluk modeli

Yöntemin Basamakları:

1. i=1 den k 'ya kadar yap → k modelleri üret:
2. değiştirme ile D örnekleme yoluyla önyükleme örneği D_i oluştur;
3. D_i'leri kullanarak bir model M_i elde et;
4. döngüyü sonlandır

Topluluk modelini bir grupta kullanmak için, X:

1. Eğer sınıflandırma ise
2. k modelin her biri ile X'i sınıflandır;
3. oy çokluğunu döndür;
4. döngüyü sonlandır
5. Eğer tahmin ise
6. k modellerinin her biri ile X için bir değer tahmin et;
7. ortalama tahmin değerini döndür;
8. döngüyü sonlandır

Düşük gürültü seviyesinde yükselteç (boosting) algoritması genellikle torbalama ve rastgeleleştirme (randomization) algoritmalarından daha doğru sonuçlar vermektedir. Ancak, gürültü seviyesi arttıkça torbalama algoritması diğer metotlardan daha iyi performans göstermektedir (García ve diğ., 2015).

2.4.MODEL GEÇERLEME YÖNTEMLERİ

Veri setinin eğitim ve test verisi olarak ayrılmasında belirli oranlarda eğitim ve test verisi olarak bölme ve çapraz geçerleme yöntemleri kullanılmıştır.

2.4.1. Belirli Oranlarda Eğitim ve Test Verisi Olarak Bölme

Belirli oranlarda eğitim ve test verisi olarak bölme (Hold-out) yöntemi, test için belirli bir miktar ayırır ve kalanını eğitim için kullanır (gerektiğinde doğrulama için de ayırır). Uygulamada, test için verilerin üçte birini tutmak ve eğitim için kalan üçte ikisini kullanmak yaygındır (Witten ve diğ., 2011).

2.4.2. Çapraz Geçerleme

N-kat çapraz geçerlemede; veri seti, yaklaşık olarak eşit büyüklükteki *n* tane birbirinden bağımsız alt kümelere rastgele bölünür. Veri seti *n* kere eğitilir ve test edilir; her defasında 1 kat test için ve kalan *n*-1 kat eğitim için kullanılır (Rokach& Maimon, 2015). Çapraz geçerleme genellikle mevcut örneklerin az sayıda olması nedeniyle bir sınıflandırma modelinin doğrulanması için kullanılır (Westerhuis, 2008).

5-kat çapraz geçerleme için eğitim ve test verisi oluşturulması Tablo 2.2'de gösterilmiştir.

Tablo 2.2: 5-kat çapraz geçerlemenin örnek gösterimi.

	1. Kat	2. Kat	3. Kat	4. Kat	5. Kat
Veri Seti	Test	Eğitim	Eğitim	Eğitim	Eğitim
	Eğitim	Test	Eğitim	Eğitim	Eğitim
	Eğitim	Eğitim	Test	Eğitim	Eğitim
	Eğitim	Eğitim	Eğitim	Test	Eğitim
	Eğitim	Eğitim	Eğitim	Eğitim	Test

2.5.MODEL PERFORMANS DEĞERLENDİRME ÖLÇÜLERİ

Bu alt bölümde, sınıflandırma modellerinin performansının karşılaştırılmasında kullanılan ölçütlere yer verilmiştir.

2.5.1. Karışıklık Matrisi

Sınıflandırma doğruluğunu hesaplamak için karışıklık matrisi kullanılabilir. Karışıklık matrisi, her bir sınıf için kaç tane örnek atandığını gösterir (Dangare & Apte, 2012).

Karışıklık matrisi, sınıflandırma sonuçlarının matris olarak temsilidir. İki sınıflı bir tahmin probleminde, sol üst hücre, doğru olduklarında doğru olarak sınıflandırılan örneklerin sayısını gösterir (doğru pozitif) ve sağ alt hücre, yanlış olduklarında yanlış olarak sınıflandırılan örneklerin sayısını gösterir (doğru negatif) (Delen ve diğ., 2005). Karışıklık matrisi, bir sınıflandırma sistemi tarafından yapılan gerçek ve tahmin edilen sınıflandırmalar hakkında bilgi içerir. Bu tür sistemlerin performansı genellikle matristeki veriler kullanılarak değerlendirilir (Goyal & Mehta, 2012).

N sınıflı sınıflandırma problemlerinde $n \times n$ boyutunda karışıklık matrisleri elde edilir. Çok sınıflı sınıflandırma problemleri için, her bir C_j sınıfı için karışıklık matrisinin dönüşümleri ele alınır. İkili sınıflandırma için karışıklık matrisi Tablo 2.3'de gösterilmiştir (Sokolova& Lapalme, 2009).

Tablo 2.3: Karışıklık matrisi.

		Tahmin Edilmiş Sınıf	
		Pozitif	Negatif
Gerçek Sınıf	Pozitif	TP	FN
	Negatif	FP	TN

Burada; Burada TP doğru pozitif, TN doğru negatif, FP yanlış pozitif ve FN yanlış negatif olarak ifade edilmektedir.

2.5.2. Doğruluk Oranı

Doğruluk; doğru pozitif (TP) ve doğru negatif (TN) örneklerin toplam sayısının;doğru pozitif (TP), doğru negatif (TN), yanlış pozitif (FP) ve yanlış negatiflerin (FN) toplamına bölünmesiyle elde edilir.

$$Doğruluk = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.16}$$

2.5.3. Kesinlik

Performans değerlendirmede yaygın olarak kullanılan ölçütlerden biri olan kesinlik, gerçekte pozitif olan örnek sayısının pozitif olarak tahmin edilen örnek sayısına oranıdır (Caruana, & Niculescu-Mizil, 2004). Kesinlik, " İstediğimi ne ölçüde alırım?" sorusunu cevaplamaktadır (Marmanis & Babenko, 2009).

$$Kesinlik = \frac{TP}{TP + FP} \tag{2.17}$$

2.5.4. Anma

Anma; doğru pozitiflerin sayısının, doğru pozitif ve yanlış negatif örnek sayısının toplamına bölünmesiyle elde edilmektedir (Fulk & Sazonov, 2011). Anma, "Aldıklarım alabileceğim her şeyi içeriyor mu?" sorusunu cevaplamaktadır (Marmanis & Babenko, 2009).

$$Anma = \frac{TP}{TP + FN} \tag{2.18}$$

2.5.5. F-Ölçütü

F-Ölçütü ilk olarak MUC-4 konferansında bilgi çıkarma teknolojisinde değerlendirme metriği olarak gösterilmiştir (Chinchor, 1992). F-ölçümü, kesinlik ve anma ölçütlerinin harmonik ortalaması alınarak hesaplanabilir (Sasaki, 2007).

$$F - \text{Ölçütü} = \frac{2 * Anma * Kesinlik}{Anma + Kesinlik} \tag{2.19}$$

F-Ölçütü 0 (neredeyse tüm ilgili öğeler alındığında) ile 1 (hemen hemen hiçbir ilgili öğe alınmadığında) arasında değişir (Liu ve diğ., 2005).

2.5.6. Kappa İstatistiği

Kappa istatistiği, herhangi bir özel durumun doğruluğunu değerlendirmek için kullanılır, toplanan verilerin güvenilirliği ve geçerliliği arasında ayrım yapmayı sağlar (Chaurasia & Pal, 2013).

Tahmin edilen ve gözlenen değerler arasındaki tutarlılık derecesini ölçen Kappa istatistiği, sonucun gerçekten doğru bir sonuç olup olmadığını veya tesadüfen meydana gelip gelmediğini görmek için modeli değerlendirir (Azari ve diğ., 2012).

$$\varkappa = \frac{P_0 - P_e}{1 - P_e} \tag{2.20}$$

Burada, P_0 gözlenen uyumluluk oranını ifade etmektedir, P_e ise rastgele beklenen uyumluluk oranını ifade etmektedir.

Elde edilen $\varkappa$ değeri Tablo 2.4'deki gibi yorumlanabilir (Landis & Koch, 1977).

Tablo 2.4: Kappa istatistiğinin yorumu.

< 0	Hiç uyum olmaması
0,01 - 0,20	Önemsiz derecede uyum
0,21 - 0,40	Zayıf derecede uyum
0,41 - 0,60	Orta derecede uyum
0,61 - 0,80	İyi derecede uyum
0,81 - 1,00	Çok iyi derecede uyum
0,81 - 1,00	Çok iyi derecede uyum

2.6.LİTERATÜR İNCELEMESİ

Bu alt bölümde ilk olarak, bu tez kapsamında önerilen sınıflandırma ve tahmin algoritmalarının perakende sektöründe kullanıldığı önemli çalışmalar kronolojik sıra ile verilmiştir. Daha sonra, gişe hasılatı tahmin probleminin çözümüne yönelik olarak literatürde yer alan çalışmalar incelenmiş ve sınıflandırılmıştır.

Bu tez çalışması kapsamında sunulan sınıflandırma ve tahmin algoritmalarının yer aldığı önemli çalışmalar kronolojik sıra ile aşağıda verilmiştir.

Alon ve diğ. (2001) ABD toplam perakende satışlarının güçlü bir eğilime ve mevsimsel desenlere sahip olduğundan bahsetmişlerdir. Bu verilerin en iyi nasıl modellenip tahmin edilebileceğinin, zaman serileri analizinde uzun süredir devam eden bir konu olduğu vurgulanmıştır. Çalışmada Yapay Sinir Ağları; Winters Üstel Düzeltme metodu, Box-Jenkins ARIMA modeli ve çok değişkenli Regresyon gibi geleneksel yöntemlerle karşılaştırılmıştır. Performans ölçütü olarak OMYH'nin kullanıldığı karşılaştırmada sonuçlar, Yapay Sinir Ağlarının daha geleneksel istatistiksel yöntemlere göre daha iyi sonuç verdiğini ve bunu Box-Jenkins modelinin izlediğini göstermektedir. Winters modelinin basit olmasına rağmen kısmen istikrarlı ekonomik koşullar altında çok adımlı tahminlerde uygulanabilir bir yöntem olduğundan bahsedilmiştir. Çalışma, Sinir Ağı Modelinin dinamik doğrusal olmayan eğilimi, mevsimsel desenleri ve hatta bunların aralarındaki etkileşimleri yakalayabildiğini göstermektedir.

Ainslie ve diğ. (2005) çalışmalarında pazar payı modeli bağlamında gişe satışlarını incelemişlerdir. Bu inceleme, hiyerarşik Bayes çerçevesinde Sürgülü Pencere Logit modeli ve

Gama Difüzyon modeli kombinasyonu geliştirilerek gerçekleştirilmiştir. Pazar payı ve talep modelleri kurulmuş ve bu modeller Markov Chain Monte Carlokullanılarak tahmin edilmiştir. Model performanslarını değerlendirirken OMYH kullanılmıştır. Her hafta erişilebilen kayıtlı tüm seçeneklerin muhasebelerinin sadece haftalık film satışlarının uygunluğunu arttırmakla kalmayıp aynı zamanda film endüstrisinin daha zengin bir resmini gösteren parametre tahminlerine de yol açtığı görülmüştür. Film stüdyolarının ürettikleri ürünlere hakim olmasını ve bunları ne zaman destekleyeceğini ve ne zaman desteklemeyeceğini bilmelerini sağlamayı amaçlamışlardır. Ayrıca, açılış haftası gösterim sayısının etkisinin geleneksel modellerde fazla tahmin edilmiş olduğunu göstermektedir. Çalışma müşterilerin film seçiminde aktörlerin direkt etkisi, yönetmenlerin ise dolaylı etkisinin olduğunu göstermektedir. Bir filmin kendiyle aynı tür olan filmlerle aynı anda vizyona girmesinin gişe performansını olumsuz yönde etkilediği çalışmadan çıkarılan sonuçlar arasındadır.

Chang ve Lai (2005) çalışmalarında piyasaya yeni girecek kitapların satış tahmini için Kendini Örgütleyen Haritalar (KÖH) ile Durum Tabanlı Çıkarsama (DTÇ) metotlarının birleşimi olan melez bir sistem sunmuşlardır. Bu melez metodun Yapay Zekânın birçok alanında başarı ile kullanılan DTÇ sisteminin etkinliğini ve yeteneğini arttırdığı savunulmuştur. Vaka çalışması için Tayvan'da bir toptancının kitap satış verisinden yararlanılmıştır ve model yeni çıkacak kitapların satış tahmini için uygulanmıştır. Modelde girdi değişkenleri olarak yazar puanı, yayınevi puanı, kitabın türü, kitabın fiyatı ve kitabın yayın tarihi kullanılmıştır. KÖH/DTÇ melez metodu ile çok iyi bir performans gözlemlenmiştir. Sonuçlar k-Ortalamalar/DTÇ melez metodu ve geleneksel DTÇ metodu ile karşılaştırılmıştır. Karşılaştırmada, her grup için ortalama satış hacmi ve OMYH performans ölçütleri kullanılmıştır. Çalışmada KÖH/DTÇ melez metodunun daha doğru ve etkin olduğu gözlemlenmiştir.

Chang ve diğ. (2006) iade kitapların tahmini için Durum Tabanlı Çıkarsama (DTÇ) sistemi ile Genetik Algoritmayı geliştiren melez bir sistem kullanmışlardır. Yeni kitaplar için yazarın puanı, yayınevinin puanı, yayınlanma tarihinin yoğun ya da yavaş bir sezon olması, ilk üç aylık satış hacmi ve iade tarihi gibi ana faktörler tanımlanmış ve yeni çıkan kitapların benzerlik katsayısını hesaplamada ana özellik olarak uygulanmıştır. Çalışma Tayvan'da bir kitap satıcısında uygulanmıştır. Melez sistemin sonuçları Geri Yayılımlı YSA'ları, geleneksel DTÇ ve Çoklu Regresyon Analizi (ÇRA) metotları ile ortalama hata oranı performans ölçütü

kullanılarak karşılaştırılmıştır. Uygulama sonuçları; GA/DTÇ melez yönteminin, iade kitapların tahmini için uygulanan diğer metotlardan daha doğru ve etkin sonuçlar verdiğini göstermiştir.

Thomassey ve Fiordaliso (2006) Rekabet ve küreselleşmenin, Tekstil-Hazır Giyim-Dağıtım ağı aktörleri için çok doğru bir üretim ve tedarik yönetimi anlamına geldiğinden bahsetmişlerdir. Çok yönlü tekstil pazarına ve distribütörlerin ihtiyaçlarına yanıt vermek için satış tahmin sisteminin gerekli olduğuna değinmişlerdir. Günümüzde, tekstil satışlarının spesifik kısıtlamaları nedeniyle, mevcut tahmin modellerinin genellikle uygun olmadığı veya kullanılamaz durumda olduğu savunulmuştur. Çalışmada orta vadeli tahmin yapabilen kümeleme (k-ortalamalar) ve sınıflandırma (C4.5) araçlarına dayalı bir tahmin sistemi önerilmiştir. Modelin performansları önemli bir Fransız tekstil distribütörünün gerçek verileri kullanılarak değerlendirilmiştir. Önerilen sistemin; ZeroR, OneR (tek sınıf sınıflandırma), Naive Bayes ve k-EYK modelleri ile OHKK, OMYH ve MYHM (Mutlak Yüzde Hataların Medyanı) performans ölçütleri altında karşılaştırıldığında en iyi sonucu verdiği görülmüştür.

Sun ve diğ. (2008) satış tahminlerinin, birçok faktöre bağlı olan talebin değişkenliğinden dolayı zorlu bir problem olduğundan bahsetmişlerdir. Bunun, çok yönlü bir satış tahmin sisteminin kritik olduğu moda perakendeciliğinde özellikle öne çıkan bir sorun olduğu anlatılmıştır. Çalışmada, satış miktarı ile talebi etkileyen bazı önemli faktörlerin (tasarım faktörleri gibi) arasındaki ilişkiyi araştırmak için AÖM adlı bir Sinir Ağı tekniği uygulanmıştır. Modellerin performansları Hong Kong'daki bir moda perakendecisinin gerçek verileri kullanılarak, Ortalama Hata Kareleri (OHK) performans ölçütü için değerlendirilmiştir. Deneysel sonuçlar; önerilen genişletilmiş-AÖM ve AÖM yöntemlerinin, Geri Yayılımlı Sinir Ağları tabanlı uyarlamalı öğrenme oranı ile en dik iniş geri yayılım algoritması ve uyarlamalı öğrenme oranı ile momentumlu gradyan azalan geri yayılım algoritmasından daha iyi performans sergilediğini göstermiştir.

Paliwal ve Kumar (2009) hem geleneksel istatistiksel yöntemlerin hem de Sinir Ağlarının, literatürde rakip model geliştirme teknikleri olarak ele alınıldığına değinmişlerdir. Çalışmada çeşitli alanlardaki tahmin ve sınıflandırma problemleri için kullanılan İleri Beslemeli Sinir Ağlarının ve istatistiksel tekniklerin karşılaştırmalı çalışmasını içeren makalelerin kapsamlı bir incelemesi gerçekleştirilmiştir. Bu makalelerin önemli özelliklerini vurgulayan sunumlar

tablo halinde gösterilmiştir. Çalışma, farklı uygulamalarda kullanılan Sinir Ağlarının ve istatistiksel yöntemlerin yararlılıkları hakkında faydalı bilgiler vermektir.

Wu (2009) çalışmasında, Dalgacık Teorisi ve değiştirilmiş DVM temel alınarak yeni bir dalgacık-DVM yaklaşımı önermektedir. Çalışmada PSO algoritması, dalgacık-DVM modelinin en iyi parametrelerini seçmek üzere kullanılmıştır. Önerilen model; ARMA, ve PSO-DVM modelleri ile OMH, OMYH ve OHK performans ölçütleri açısından karşılaştırılmıştır. Otomobil satış serisi tahmininde uygulanan modellerin sonuçları, önerilen modele dayanan tahmin yaklaşımının etkili ve uygulanabilir olduğunu göstermiştir. Önerilen yöntem diğer yöntemlerle karşılaştırıldığında diğer yöntemlerden daha iyi sonuç verdiği görülmüştür.

Wu (2010) yaptığı çalışmada, çok boyutlu zaman serilerini tahmin etmek için bulanık Destek Vektör Makinesinin yeni bir versiyonunu sunmaktadır. Birçok tahmin probleminde sonlu örneklemeler ve belirsiz verilerle ilgili bazı problemler olduğu için, çalışmada girdi değişkenleri bulanık kapsamlı değerlendirme ile gerçek sayı olarak tanımlanmıştır. Bu girdi değişkenlerinin bulanıklık derecesini göstermek için simetrik üçgen bulanıklık tekniği kullanılmıştır. Daha sonra bulanık teori DVM ile birleştirilerek, üçgen bulanık alan üzerinde bulanık-DVM yaklaşımı önerilmiştir. Bulanık-DVM'nin en iyi parametrelerini bulmak için PSO kullanılmıştır. Önerilen yaklaşım araba satışlarının tahmini için uygulanmıştır. Satış tahminlerindeki uygulama sonuçları, bulanık-DVM modelinin geçerliliğini doğrulamıştır. Bulanık-DVM modeli; ARMA ve DVM modelleri ile OMH, OMYH ve OHK performans ölçütleri açısından karşılaştırılmıştır. Sonuçlar, önerilen yönteminin daha az örnek gerektirdiğini ve daha iyi tahmin hassaslığına sahip olduğunu göstermiştir.

Tanaka (2010) yaptığı çalışmada yeni çıkan ürünlerin satış tahmini yapabilmek için bir satış tahmin modeli önermiştir. Çalışmada; benzer ürün grupları içinde kısa ve uzun vadeli birikmiş satışlar arasındaki yüksek korelasyon kullanılarak yeni bir tahmin modeli önerilmiştir. N. günün birikmiş satış tahminlerinden yararlanarak m. günün birikmiş satışını tahmin eden bu yönteme “NM yöntemi” adı verilmiştir. Kısa ve uzun vadeli birikimlerin korelasyon veri tabanı temelli modeli, ürünlerin piyasaya sunulduğu ilk zamanların satış verilerini kullanarak uzun vadeli tahmin sağlamaktadır. Model; doğruluk, tahmini piyasaya çıkma zamanı ve geniş ürün yelpazesi olmak üzere üç nokta üzerine kurulmuştur. Vaka çalışması olarak, Japonya’da kitap ve elektronik ürünler satılan bir mağazada

gerçekleştirilmiştir. Önerilen yöntem; Hareketli Ortalama (HO) ve Üstel Düzeltme (ÜD) yöntemleri ile OMYH performans ölçütüne göre karşılaştırılmıştır. Model ürünlerin piyasaya sürülmesinden bir veya iki hafta sonra ürünlerin yaşam döngüsü boyunca satış tahminlerini elde etmeyi amaçlamaktadır. Ayrıca kitaplar için tekrar basım kararı vermede değerli bilgi sağlayan bir model olduğu savunulmuştur.

Wong ve Guo (2010) çalışmalarında, orta vadeli moda satışı tahmin problemini çözmek için veri önişleme bileşenini ve melez zekâ tahmincisini içeren bir Melez Zekâ (MZ) modeli geliştirilmiştir. MZ tahmincisinin, önce ilk satış tahminlerini oluşturmak için öğrenme algoritması tabanlı Sinir Ağı uyguladığı, sonrasında başlangıçtaki tahminlerden daha doğru tahminler elde etmek için sezgisel ince ayar işlemi kullandığı belirtilmiştir. Öğrenme algoritmasının, ağın genel performansını artırmak için geliştirilmiş uyum arama algoritması ve AÖM'yi entegre ettiğinden bahsedilmiştir. Çalışmada önerilen modelin performansını değerlendirmek için gerçek moda perakende verileri ve halka açık veri setleri üzerinde uygulamalar yapılmıştır. Evrimci Sinir Ağları (ESA), Sun ve diğ. (2008) çalışmalarında önermiş oldukları genişletilmiş AÖM, ARIMA, AR ve AR2 metotlarıyla OHKK, OMYH ve Ortalama Mutlak Ölçekli Hata (OMÖH) performans ölçütleri için karşılaştırılmıştır. Uygulama sonuçları önerilen modelin performansının moda satış tahmini için geleneksel ARIMA modeli ve daha önce geliştirilmiş iki Sinir Ağı modelinden daha üstün sonuçlar verdiğini göstermektedir.

Yu ve diğ. (2011) talep ve arz ile ilgili belirsizlikler nedeniyle satış tahminlerinin moda sektöründe çok önemli olduğundan bahsetmişlerdir. Moda ürünlerini tahmin etmeye yönelik birçok modelin literatürde var olduğu vurgulanmıştır. Yapay Zeka modellerinin ortaya çıkması ile birlikte, YSA tahminde yaygın olarak kullanıldığı belirtilmiştir. YSA modellerinin birçok geleneksel istatistiksel tahmin modelinden daha verimli ve etkili olduğu dile getirilmiştir. Bu avantajlarına rağmen, YSA'nın tahmin yapmasının nispeten daha uzun sürdüğü ifade edilmiştir. Moda endüstrisinde, çok sayıda ürün çeşidi olduğu için satış tahminlerinin zorlu olduğu ve hızlı tahmin sonuçlarına ihtiyaç duyulduğu belirtilmiştir. Bu nedenle mevcut YSA modellerinin yetersiz kaldığı düşünülmüştür. Çalışmada hem AÖM hem de geleneksel istatistiksel yöntemleri kullanan yeni bir Akıllı Hızlı Satış Tahmini (AHST) modeli önerilmiştir. Önerilen bu AHST modeli, geleneksel YSA ve istatistiksel yöntemlerden biri olan çok terimli regresyon modeli ile karşılaştırılmıştır. Modeller gerçek veri setleri

üzerinde uygulanmıştır. OHK ve tahmin süreleri performans ölçütleri ile karşılaştırıldığında önerilen AHST modelinin diğer geleneksel yöntemlerden hızlı ve doğruluk oranı yüksek sonuçlar verdiği gözlemlenmiştir.

Ni ve Fan (2011) moda perakendeciliği tahmini ile ilgili zorlukların; mevsim, bölge ve moda etkisi gibi bir takım faktörlere bağlı olduğuna ve orijinal satış kurallarında doğrusal olmayan bir değişikliğe neden olduğuna değinmişlerdir. Çalışmada moda perakendeciliği tahmininin doğruluğunu iyileştirmek için, uzun vadeli ve kısa vadeli tahminlerin birleştirildiği iki aşamalı dinamik bir tahmin modeli önerilmiştir. Model, geliştirilmiş düzeltme yöntemlerini, ana ayarlama modelini ve hata öngörme modelini birbiriyle birlikte çalıştıran ayarlama sistemi olarak tanıtılmıştır. Model gerçek zamanlı veri üzerinde uygulanmıştır. Modelin etkinliğini ölçmek için GHKK performans ölçütünden yararlanılmıştır. Uygulama, önerilen modelin moda perakende tahmini için iyi sonuçlar verdiğini göstermektedir.

Wu ve Law (2011) çalışmalarında bulanık Destek Vektör Makinesinin yeni bir versiyonunu sunmaktadırlar. Üçgen bulanık teori, Destek Vektör Makinesi ile birleştirerek; kısıt koşulları standart bulanık DVM'den daha az olan, genişletilmiş bulanık üçgen sayılar kullanılan yeni bir yaklaşım önerilmiştir. Önerilen bu yaklaşımın en iyi parametrelerini bulmak için, Parçacık Sürüsü Optimizasyonu (PSO) kullanılmıştır. Önerilen yaklaşım araba satışlarının tahmininde uygulanmıştır. Satış tahminlerindeki uygulama sonuçları, önerilen modelinin geçerliliğini doğrulamaktadır. Bu model; ARMA, DVM ve bulanık-DVM modelleri ile OMH, OMYH ve OHK performans ölçütleri açısından karşılaştırılmıştır. Karşılaştırma sonuçları, önerilen yönteminin daha az örnek gerektirdiğini ve daha iyi tahmin hassaslığına sahip olduğunu göstermiştir.

Chen ve Ou (2011) çalışmalarında; güçlü rekabet nedeniyle perakendecilerin, kârlarını artırmak ve maliyetlerini düşürmek için sürekli çaba gösterdiğinden bahsetmişlerdir. Doğru bir satış tahmin sisteminin, perakendeci için önemini vurgulamışlardır. Bu çalışmada satın alma kararlarını desteklemek için GİA ve AÖM'nin entegre edildiği Gri Aşırı Öğrenme Makinesi (GAÖM), Taguchi yöntemiyle bütünleştirilerek önerilmiştir. GİA'nın ham verilerden daha etkili faktörleri seçtiğinden ve bunların AÖM gibi sinir ağlarında girdi verileri olarak dönüştürüldüğünden bahsedilmiştir. Önerilen sistem perakende endüstrisinde gerçek satış verisiyle değerlendirilmiştir. Önerilen model GİA, Geri Yayılım Ağları entegresiyle ve GİA, Çok Katmanlı Fonksiyonel Bağlantı Ağları (ÇKFBA) entegresiyle OMS ve OHK

performans ölçütleri açısından karşılaştırıldığında en iyi sonucu önerilen GAÖM modelinin verdiği gözlemlenmiştir.

Yu ve diğ. (2012) moda rengi eğiliminin son derece değişken olmasından dolayı, gelecekteki renk trendini tahmin etmenin tasarım, üretim ve satış gibi moda endüstrisinde çok önemli ve zorlayıcı bir süreç olduğundan bahsetmişlerdir. Gelişmiş yöntemler olmadan, oldukça yüksek doğrulukla moda renk trend tahminini yapmanın çok zor olduğuna ve bu durumun moda endüstrisinde akıllı uzman sistemlerin geliştirilmesinde bir engel olduğuna değinilmiştir. Literatürde birçok çalışmada ARIMA gibi geleneksel regresyon modellerinin veya YSA ve Gri Model (GM) gibi akıllı modellerin kullanılarak renk eğilimi tahminlerinin gerçekleştirildiği söylenmiştir. Bununla birlikte, bu tahmin yöntemlerinin doğruluklarının çok çeşitli olduğu ve bu modellerin performansları hakkında literatürde tartışmalar bulunduğu ifade edilmiştir. Bu nedenle bu çalışmada; ARIMA, YSA ve GM modellerinin performansları veri gereksinimi, hesaplama hızı, doğruluk ve yöntemin sadeliği ölçütleri açısından karşılaştırılmıştır. Gerçek veri analizi ile elde edilen sonuçlar, YSA ailesinin modellerinin, özellikle Gri İlişkisel Analiz (GİA) ile AÖM'nin moda renk trendini tahmin etmek için diğer modellerden daha iyi olduğu görülmüştür.

Lee ve diğ. (2012) çalışmalarında ürün yaşam döngüsünün ilk evrelerinde ürünün talebini tahmin etmenin zor olduğunu ancak politika oluşturma ve piyasa için bir o kadar da hayati bir durum olduğu üzerinde durmuşlarıdır. Buradan yola çıkılarak çalışmada eldeki kısıtlı veri ile yeni çıkacak ürün için yapılacak tahminin doğruluğunu arttıracak bir yöntem önerilmek istenmiştir. Çalışmada müşteri çekince fiyatı ile müşterinin ürünü benimseme süresinin ilişkili olduğu varsayımı yapılmıştır. Modelde müşteri anketlerinden alınan müşteri çekince fiyatı verisini kullanarak tahmin yapılmaktadır ve tahmin Bayes kuralları kullanılarak erişilebilen satış verileri ile güncellenmektedir. Müşteri bekleme süresi olasılık yoğunluk fonksiyonu tabanlı bu yöntem geleneksel bir metot olan Analoji ile OMYH ve Bayes bilgi kriteri (BBK) performans ölçütleri için karşılaştırılmıştır. Önerilen metot iki performans ölçütü için de en iyi değerleri vermiştir. Sonuçlar müşteri çekince fiyatının yeni ürün piyasaya sürülmeden önce veya sürüldükten kısa bir süre sonra talep tahmini için kullanılabilir olduğunu göstermiştir.

Kulkarni ve diğ. (2012) çalışmalarında müşterilerin aradıkları ürün, kavram veya ilgili terimle alakalı daha çok şey öğrenmek için sıklıkla arama motorları kullandığı görüşünden yola çıkmışlarıdır. Bu çalışmada müşteri tarafından kullanılan arama terimleri verisinin müşterinin

ilgilendiği ürün, kavram ya da terimlerle ilgili değerli ölçü ve göstergeleri içerdiği öne sürülmektedir. Özellikle bazı verilerin yöneticiler için piyasaya sürülecek yeni ürünlerin potansiyelini ölçmede çok değerli oldukları savunulmuştur. Çalışmanın amacı, yayın öncesi ve sonrası arama davranışlarını modellemek ve gözlemlenen arama örüntüsünü filmlerin açılış haftası gişe satışlarıyla ilişkilendirmektir. Çalışmada, Markov Zinciri Monte Carlo Simülasyonu (MZMCS) kullanılmıştır. Bağımsız model, arama hacmi bileşeninde filme özgü eşdeğer değişken içermeyen model, hem satış hem de arama hacminde filme özgü eşdeğer değişken içeren model ve reklam etkili model olmak üzere dört model için uygunluk ve performanslar karşılaştırılmıştır. Modellerin karşılaştırılmasında, ÇBK (Çarpıklık Bilgi Kriteri) ve OYH (Ortalama Yüzde Hata) performans ölçütleri kullanılmıştır. Sonuçlar çevrimiçi aramaların ve reklam etkisinin tahmin gücünü arttırdığını göstermiştir.

Lu ve diğ. (2012) Yapay Sinir Ağlarının (YSA) satış ve talep tahmini için yararlı olduğuna değinmişlerdir. Ancak YSA'nın temel eksikliklerinden birisinin, önemli tahmin değişkenlerini tanımlayamadığı olduğu vurgulanmıştır. Çalışmada bilgisayar toptancıları için satış tahmini modelleri oluşturmada doğrusal olmayan ve parametrik olmayan regresyon yöntemi olan Çok Değişkenli Uyarlamalı Regresyon Uzanımları (ÇDURU) yöntemi kullanılmıştır. ÇDURU'nın olağanüstü değişken tarama kabiliyeti sayesinde, bilgisayar toptancıları için önemli satış tahmin değişkenlerinin elde edilebileceği vurgulanmıştır. ÇDURU'nın performansını değerlendirmek için Tayvanlı bilgisayar toptancılarından alınan gerçek satış verileri kullanılmıştır. Deneysel sonuçlar ÇDURU modelinin, Geri Yayılımlı Sinir Ağları, DVM, Serebellar Modeli Telâffuz Denetleyicisi Sinir Ağları, AÖM, ARIMA modeli, çok değişkenli DR modeli ve dört tane iki aşamalı tahmin modeli (ÇDURU-DVR, ÇDURU-Geri yayılımlı sinir ağları, ÇDURU-AÖM ve ÇDURU-Serebellar Modeli Telâffuz Denetleyicisi Sinir Ağları) ile ortalama mutlak sapma (OMS), OHKK, OMYH ve ortalama yüzde hata karelerinin karekökü (OYHKK) gibi farklı performans kriterleri açısından karşılaştırıldığında en iyi sonucu verdiğini göstermektedir. Ayrıca, ÇDURU tahmin sonuçlarının, seçilen tahmin değişkenleri ile satış tutarları arasındaki ilişkiler hakkında yararlı bilgiler sağladığı ve bu nedenle de, uygun satış kararları ve stratejilerinin uygulanması için önemli etkilere sahip olduğu belirtilmiştir.

Yu ve diğ. (2013) bilgi teknolojilerindeki gelişmelerin birçok açıdan hayatımızı değiştirdiğinden bahsetmişlerdir. İnsanların internetten haber ve hikâye aradıkları bir eğilimin

var olduğunu ifade etmişlerdir. Bu koşullar altında, geleneksel medya şirketlerinin gazete ve dergi gibi basılı satışlarını her zamankinden daha iyi tahmin etmesinin ciddi bir durum olduğu söylenmiştir. Gazete ve dergilerin satış tahminleri için yapılan önceki çalışmalarda örnek veri setlerine dayanan regresyon modelleri oluşturma üzerine odaklanıldığından bahsedilmiştir. Ancak, bu tür regresyon modellerinin aşırı uyum probleminden zarar görebileceği belirtilmiştir. İstatistikteki son teorik çalışmaların aşırı uyum probleminin üstesinden gelmek için DVR gibi yeni bir yöntem önerdiği belirtilmiştir. Geleneksel regresyon modellerinin aksine, DVR'nin amacının minimum deneysel risk yerine minimum yapısal riski elde etmek olduğu öne sürülmüştür ve bu nedenle; çalışmada DVR, satış tahmin problemine uygulanmıştır. Uygulama DVR'nin bu tür problemlerde üstün bir yöntem olduğunu göstermiştir.

Lee ve diğ. (2014) çalışmalarında yeni bir ürünün piyasaya sürülmeden önce talep tahmininin yapılması için Bass modeli ve istatistiksel ve makine öğrenmesi algoritmaları tabanlı yeni bir yaklaşım sunmuşlardır. Çalışmada istatistiksel ve makine öğrenmesi algoritmaları ürünler piyasaya girmeden önce Bass modelinin 2 parametresini tahmin etmek için kullanılırken, Bass modeli ürünlerin yayılma sürecini açıklamak için kullanılmıştır. Öncelikle ürün özellikleri ve ürün yayılımını içeren ve birincisi girdi olarak, diğeri çıktılar olarak algılanan 2 tip veri tabanı oluşturulmuştur. ÇDR, k-EYK, YSA, DVR, SRA ve Gauss Süreç Regresyonu olmak üzere altı tek tahmin algoritması kullanılmış ve tahmin gücü arttırmak için bir kolektif tahmin modeli oluşturulmuştur. Parametreleri Bass modeli tarafından belirlenen; en iyi doğruluğa sahip ÇDR, kolektif modellerinden ÇDR-Gauss Süreç Regresyonu ve Analoji modelleri OMH ve OHKK performans ölçütlerine göre karşılaştırılmıştır. En iyi sonuçları ÇDR-Gauss Süreç Regresyonu kolektif modelinin verdiği görülmüştür. Önerilen bu model 3D Tv'ler için talep tahmini yapmada uygulanmıştır. Deneysel sonuçlar; çoğu tek tahmin modelinin, geleneksel Analoji yönteminden daha iyi olduğunu ve kolektif modelin, tahmin kesinliğini daha da geliştirdiğini göstermiştir.

Choi ve diğ. (2014) hızlı modanın, moda perakendeciliğinde yaygın olarak benimsenen bir strateji olduğundan bahsetmişlerdir. Hızlı moda altında, işlevsel kararların sıkı bir çizelgeyle yapılması gerektiği ve ilgili tahmin yönteminin, sınırlı bir süre içinde, çok sınırlı verilerle çalışması gerektiği vurgulanmıştır. Bu makalede, Aşırı Öğrenme Makinesi ve Gri Model gibi araçları bir araya getiren akıllı tahmin algoritması geliştirilmiştir. Çalışma gerçek bir veri

setine uygulanmıştır. Önerilen bu hızlı moda tahmini modeli GM ve GM-genişletilmiş AÖM modelleri ile OMYH performans ölçütü açısından karşılaştırılmıştır. Önerilen algoritmanın verilen zaman ve veri kısıtlamaları altında oldukça iyi tahmin yapabildiği gösterilmektedir. Yapay bir veri kümesiyle daha yapılan analiz, talep eğiliminin eğimi büyük veya mevsimsel döngünün varyansı büyük olduğunda, önerilen algoritmanın özellikle iyi performans sergilediğini göstermektedir. Bu çalışmanın sonuçları, gelecekte hızlı moda operasyonları için gerçek zamanlı satış tahminleri gerçekleştirmeye yardımcı olabilecek temelleri ortaya koymaktadır.

Castillo ve diğ. (2016) çalışmalarında yayınevlerinin karşılaştıkları en önemli problemlerden biri olan piyasaya yeni çıkacak bir kitaptan kaç tane basılması gerektiği problemi üzerinde durmuşlardır. Çalışmada, çok az ya da çok fazla basım yapmanın yayınevine negatif ekonomik bir etki yapacağı bu nedenle de doğru tahminin çok önemli olduğu vurgulanmıştır. Bunun üzerine çalışmada kitapların henüz daha satışa başlamadan önce doğru miktarlarda basılması için toplam satışının tahmini problemi ele alınmıştır. 7 yıl boyunca İspanya da yayımlanmış kitaplar için tüm satış verisini içeren gerçek bir veri seti kullanılmıştır. Çalışma üç aşamada yürütülmüştür. Bunlardan ilki veri görüntüleme teknikleri yardımıyla verinin incelendiği aşamasıdır. Diğer aşama verinin özellik seçimi aşamasıdır ve bu aşamada satışta daha etkin olan değişkenlerin hangileri olduğunu belirlemek için farklı teknikler kullanılmıştır. Son aşama ise tahmin aşamasıdır, bu aşamada kitap satışları için tahmin modeli oluşturmada; M5A (M5 Model ağaçları), RO, k-EYK, DVM, DR (Doğrusal Regresyon), ÇKA ve AÖM algoritmaları uygulanmıştır. Uygulanan algoritmaların performansı OMH, OHKK, GMH ve GHKK ile değerlendirilmiştir. En iyi sonuçları veren tahmin algoritmasının M5A olduğu gözlemlenmiştir.

Schneider ve Gupta (2016) tüketici incelemelerinde yer alan sayısal ve metin verileri kullanarak yeni ve mevcut ürünlerin satış tahmini problemini ele almışlardır. Çalışmada, mevcut yaklaşımların birçoğunun metin verilerin ciddi anlamda manüel olarak ön işlemesini gerektirdiğinden, bu yöntemlerin uygulanmasının oldukça pahalı ve zor olduğundan bahsedilmiştir. Buna karşı olarak, önerilen yaklaşımda, az önişleme ve ayrıştırma gerektiren, verimli ve ölçeklenebilir bir yöntem olan Kelime Çantası (KÇ) metodu kullanılmıştır. KÇ modellerini kötü etkileyen çok boyutluluk sıkıntısının üstesinden gelmek için Rastgele Projeksiyonlar (RP) yaklaşımı önerilmiştir. RP yaklaşımlarının hesaplama açısından basit,

esnek ve hızlı olduğuna ve çekici istatistiksel özelliklerinin olduğuna değinilmiştir. Önerilen yaklaşım niteliklere dayalı regresyon modeli ile tüketici incelemelerini kullanarak Amazon.com'daki satışların tahmini için uygulanmıştır. Model, mevcut ve yeni tanıtan tablet bilgisayarlar için bir haftalık ileriye yönelik satış tahminlerini üretmek için uygulanmıştır. OHKK ve OMYH performans ölçütlerine göre yapılan karşılaştırmalar, önerilen yaklaşımın tahmin performansının, gerek tüketici incelemelerinin metin içeriğini dikkate almayan gerekse metin içeriğe sahip bir Destek Vektör Makinesi modelinden daha güçlü ve daha iyi olduğunu göstermektedir. Ayrıca, yaklaşımın ürün kategorileri arasında kolaylıkla tekrarlanabilir ve daha geniş veri kümelerine kolaylıkla ölçeklenebilir olduğundan bahsedilmiştir.

Gişe hasılatı tahmini ile ilgili literatürde yer alan çalışmalar kronolojik sıraya göre aşağıdaki gibi özetlenebilir:

Elberse ve Eliashberg (2003) çalışmalarında, yeni bir ürünün iç pazarda ilk kez piyasaya sürüldüğü ve sadece dış pazarlarda daha sonraki bir aşamada başlatıldığı ve ürünün talebinin ve tedarikinin zaman içinde birbirine yüksek ölçüde bağlı olduğu bir ortamı incelemek için bir model geliştirmişlerdir. Çalışmada, müşterilerin ve perakendecilerin davranışları arasındaki karşılıklı ilişki ve etkenlerin dinamik eşanlı denklemler modeli geliştirilmiştir. Gelire ve yayınlanmaya (ekrana) bağlı denklem sistemi oluşturulmuştur. Bulgular, seyirci ve katılımcı davranışının içyapısı ve eşzamanlılığını göz önüne almanın önemini vurgulamıştır. Ayrıca, genellikle izleyicileri doğrudan etkilediği varsayılan film nitelikleri ve reklam harcamaları gibi değişkenlerin, çoğunlukla gelirleri dolaylı olarak etkilediği görülmüştür.

Sharda ve Delen (2006) çalışmalarında henüz vizyona girmemiş filmlerin gişe hasılatını tahmin etmek için sinir ağlarını kullanmışlardır. Modelde tahmin problemi, gişe hasılatının nokta tahmini yerine 9 kategoriden oluşan bir sınıflandırma problemine dönüştürülmüştür. Tahmin sonuçlarını değerlendirmek için performans ölçütü olarak ortalama doğru tahmin oranı (ODTO) kullanılmıştır. Bu çalışmadaki Sinir Ağları modelinin sınıflandırmadaki başarısı daha önce literatürde önerilen Çoklu Lojistik Regresyon (ÇLR), Diskriminant Analizi (DA), Sınıflandırma ve Regresyon Ağacı (SRA) algoritmaları ile karşılaştırılmış ve Sinir Ağlarının çok daha iyi sonuçlar verdiği görülmüştür.

Delen ve diğ. (2007) Sharda ve Delen'in 2006'da gerçekleştirdiği çalışmadaki modeli kullanan ve sonuçları görselleştiren bir web tabanlı karar destek sistemi oluşturmuşlardır.

Dellarocas ve diğ. (2007) çevrimiçi ürün inceleme forumlarının artan popülariteleri ile birlikte, firmaların karar destekleri için yeni ve önemli bilgi kaynakları haline geldiğinden bahsetmişlerdir. Çalışma, eğlence endüstrisinin benzersiz özelliklerinden bazılarını yakalayan ve vizyona daha yeni giren filmlerin hasılat tahmini bağlamında performanslarını test eden yeni difüzyon modelleri ailesi önermektedir. Önerilen modellerden ilki A Modeli; pazarlamanın, yayınlanma sayısının, profesyonel eleştiri incelemelerinin, erken gişe hasılatının ve çevrimiçi inceleme verilerinin mevcut olduğu durumlarda kullanılmak üzere hazırlanmıştır. B Modeli, erken gişe verilerinin mevcut olmadığı durumlarda kullanılmak üzere hazırlanmıştır. C Modeli, B modeline benzemekte ancak herhangi bir çevrimiçi inceleme verisi kullanmamaktadır. D Modeli ise yalnızca çevrimiçi gözden geçirme verilerine dayanmaktadır. Önerilen modeller OMYH performans ölçütüyle karşılaştırıldığında en iyi sonucu A Modelinin verdiği görülmüştür. Çalışma önceden pazarlama, sinema salonu bulunurluğu ve profesyonel eleştiri yazıları içeren bir karşılaştırma modeli için çevrimiçi ürün inceleme metriklerinin eklenmesinin tahmin doğruluğunu önemli ölçüde arttırdığını göstermektedir.

Duan ve diğ. (2008) internetteki ağızdan ağıza pazarlamanın (AAP) perakende sektöründe müşterilerin alım kararını nasıl etkilediğinin, giderek üzerinden durulan bir konu olduğunu vurgulamışlardır. Perakende satışlarının öncüsü ve sonucu olarak çevrimiçi AAP efektini ayırdıkları, dinamik eşanlı denklemler sistemi vasıtasıyla süreç tanımlanmıştır. Yaklaşım hem filmlerin gişe hasılatını hem de AAP değerliliğinin AAP hacmini önemli derecede etkilediğini göstermek için film endüstrisinde uygulanmıştır. Tek denklem kullanılan en küçük kareler yöntemi ile iki denklem sisteminin kullanıldığı üç aşama en küçük kareler yöntemi, logaritmik doğrusal fonksiyonun tahmin katsayısına göre karşılaştırılmıştır. AAP hacminin gişe hasılatı performansını arttırdığı gözlemlenmiştir. Bu pozitif geri bildirim mekanizması, perakende gelirlerini üretmek ve sürdürmek için AAP'nin ne kadar önemli olduğunu vurgulamıştır.

Zhang ve diğ. (2009) henüz vizyona girmemiş olan filmlerin gişe gelirlerini önceden tahmin etmek amacıyla çok girdi ve çok çıktı ile Çok Katmanlı Geri Yayılımlı Yapay Sinir Ağları algoritması kullanılarak bir tahmin modeli geliştirmişlerdir. Tüm filmler için gişe gelirleri altı

farklı kategoriye ayrılmıştır ve amaç filmin sınıfını doğru bir şekilde tahmin etmektir. Girdi verilerinin seçimlerinde piyasa araştırmalarından yararlanılmıştır ve ağırlıklar istatistiksel metotlar kullanılarak belirlenmiştir. Tahmin modelinin performansının ölçülmesinde 6-kat çapraz geçerleme kullanılmıştır. Çok Katmanlı Algılayıcı (ÇKA) metot ile karşılaştırma yapıldığında Çok Katmanlı Geri Yayılımlı Yapay Sinir Ağı tahmin modelinin daha tatmin edici sonuçlar verdiği görülmüştür. Bu problemin çözümünde Çok Katmanlı Geri Yayılımlı Yapay Sinir Ağı tahmin modelinin kullanılmasının daha güvenilir ve efektif olduğu sonucuna ulaşılmıştır.

Delen ve Sharda (2010) bu çalışmalarında Yapay Sinir Ağları, Karar Ağaçları, Destek Vektör Makineleri algoritmaları ve bilgi birleştirme yöntemleri gibi veri madenciliği yöntemlerini kullanarak henüz vizyona girmemiş filmlerin gişe hasılatının finansal performansını tahmin etmeye çalışmışlardır. Problemde gişe hasılatının nokta tahmininin yapılması yerine problem 9 sınıftan oluşan bir sınıflandırma problemi içinde ele alınmıştır. MPAA film reytingleri, aylara bağlı rekabet, oyuncu popülerliği, film türü, özel efektler, devam filmi olup olmadığı ve filmin kaç kez oynatıldığı değişkenleri girdi değişkenleri olarak kullanılmıştır. DVM, YSA, SRA, kolektif-RO, kolektif-HA ve kolektif-Bilgi birleştirme metotları performans ölçütü olarak ODTO kullanılarak karşılaştırılmıştır. En iyi performansı kolektif-Bilgi birleştirme metodunun verdiği görülmüştür.

Asur ve Huberman (2010) son yıllarda sosyal medyanın sosyal paylaşım ve içerik paylaşımı için yaygın ve önemli hale geldiğinden ama buna rağmen bu web sitelerinden üretilen içeriğin büyük oranda kullanılmadığından bahsetmişlerdir. Buradan yola çıkılarak çalışmalarında sosyal medya içeriğinin gerçek dünya sonuçlarını tahmin etmek için nasıl kullanılabileceğini göstermişlerdir. Filmlerin gişe hasılatı tahminini yapmak için Twitter'daki sohbetler kullanılmıştır. Çalışmada; sosyal medyada ürünün konuşulma sıklığını ifade eden dikkat oranı parametresi, filmin yayımlanmasından sonra önem kazanan ve tahminlerin doğruluğunu arttırdığı gözlemlenen sosyal medyada yaygınlaşan görüşleri içeren ikinci bir parametre ve filmin yayınlanma sayısını içeren dağıtım parametresi olmak üzere üç parametreli doğrusal regresyon modeli kullanılmıştır. Modelin etkinliğinin ölçülmesinde ayarlanmış OMYH performans ölçütünden yararlanılmıştır. Belirli konularda tweetlerin oluşturulma oranından yararlanarak kurulan bu basit modelin, piyasaya dayalı tahminlerden daha iyi performans

gösterebileceği anlatılmıştır. Ayrıca, çalışma Twitter'dan alınan duyguların (yorumların) sosyal medyanın öngörü gücünü artırmak için nasıl kullanılabileceğini göstermektedir.

Kim ve diğ. (2015) yaptıkları çalışmada Sosyal Ağ Servisi (SAS) verisi ve makine öğrenmesi algoritmaları kullanarak sinema filmlerinin gişe hasılatını tahmin eden yeni bir yaklaşım sunmuşlardır. Çalışmada; tahmin algoritmalarının ve sinema alanında kullanılan açıklayıcı değişkenlerin kapsamlı araştırması yapılmıştır. Erken dönemde tahminin önemi nedeniyle birikimsiz ve birikimli gişe hasılatı kazancını tahmin için; önce, bir hafta sonra ve iki hafta sonra olmak üzere üç ardışık tahmin modeli geliştirilmiştir. Görüntülenme ile ilgili bilgilere ek olarak sosyal ağlarda bahsedilme sayısı ve onların haftalık eğilimleri girdi değişkenleri olarak kullanılmıştır. Önemli girdi değişkenlerini tanımlamak için Genetik Algoritma (GA) kabul edilmiştir. Tahmin modellerini oluşturmak için ÇDR, DVR, GSR ve k-EYK olmak üzere dört regresyon algoritması kullanılmıştır. Dört ayrı regresyon modeline ek olarak, tahmin doğruluğunu artırmak için DVR, GSR ve k-EYK olmak üzere üç makine öğrenmesi tabanlı regresyon algoritmasının kombinasyonuna dayanan bir başka tahmin yapılmıştır. İlk olarak, algoritmaların her biri, aynı eğitim verilerini kullanarak eğitilmiştir ve sonrasında, kombinasyon modelinin tek bir tahmin oluşturması için tahminler eşit ağırlıklı ortalamalar kullanılarak birleştirilmiştir. Çalışmada; performans değerlendirmek için OMYH ve OHKK kullanılmıştır. Uygulama sonuçları Sosyal Ağ Servisi (SAS) verisinin, makine öğrenmesi tabanlı algoritmaların ve onların kombinasyonlarının üç ardışık tahmin modelinin hepsinde tahmin doğruluklarını fark edilir derecede iyileştirdiğini göstermiştir.

Ghiassi ve diğ. (2015) çalışmalarında, film sektöründe gişe hasılatı belirlemek için yapılan çalışmaların çoğunda mevcut modellerin ilk hafta verilerini kullanarak tahmin yaptığına değinmiştir ve kendi çalışmalarında henüz vizyona girmemiş filmlerin gişe hasılatını tahmin etmeye çalışmışlardır. Çalışmada; vizyon öncesi zamanındaki veriler kullanılarak Dinamik Yapay Sinir Ağları (DYSA) yöntemi uygulanmıştır. Gişe hasılatları 9 sınıfa ayrılarak ele alınmıştır. Daha önceki benzer çalışmalarda kullanılan değişkenlerden bazıları çıkarılarak reklam bütçesi, mevsimsellik gibi değişkenler eklenmiştir. DVM, YSA, SRA, RO ve HA gibi mevcut modellerle ODTO açısından karşılaştırma yapıldığında; kurulan model ve uygulanan Dinamik Yapay Sinir Ağları yönteminin daha efektif olduğu ve doğruluk oranını arttırdığı görülmüştür.

Hur ve diğ. (2016) gişe hasılatı tahmin etmenin film distribütörlerinin karar verme süreci için ilgi çekici ve zor bir iş olduğundan bahsetmiş ve gişe hasılatını doğru tahmin etmek için daha önce yapılan çalışmalarda değişken çeşitliliğinin azlığı ve tahmin algoritmalarının basitliği yüzünden tatmin edici sonuçlara ulaşılamadığını vurgulamışlardır. Çalışmada tahmin doğruluğunu arttırmak amacıyla, eleştiri/yorum duyarlılığından faydalanarak ve doğrusal olmayan makine öğrenmesi algoritmaları kullanarak yeni bir gişe hasılatı tahmin modeli önerilmiştir. Girdi değişkenlerinin kaynağı olarak sinema, seyirci ve dış faktörler olmak üzere üç farklı faktör kullanılmıştır. Doğrusal bir model olan ÇDR ve SRA, YSA, DVR olmak üzere 3 farklı makine öğrenmesi tabanlı algoritma gişe hasılatı ve tahmin algoritmaları arasındaki doğrusal olmayan ilişkiyi yansıtmak için kullanılırken geleneksel tahmin algoritmalarına ek olarak eleştiri yazılarından izleyici duyarlılığı girdi değişkeni olarak kullanılmıştır. Makine öğrenmesi tabanlı tahmin algoritmaları için değişken ağırlıklarını bulma amacıyla Bağımsız Alt Uzay (BAU) ve Rastgele Alt Uzay (RAU) metodu uygulanmıştır. Önerilen modeller tahmin başarısını değerlendirmek için OHKK ve OMYH performans ölçütleri kullanılmıştır. Tahmin sonuçları, makine öğrenmesi tabanlı algoritmaların, mevcut veriler yetersiz olduğunda veya uzun vadeli tahminlerde doğruluk oranı daha yüksek sonuçlar verdiğini göstermiştir.

Filmlerin gişe hasılatının tahmini problemi üzerine yapılan çalışmaların sınıflandırılması Tablo 2.5'de sunulmuştur.

Tablo 2.5: Gişe hasılatı tahmini problemine ilişkin literatür özeti.

Yayın Bilgisi	**Kullanılan Değişkenler**	**Metot**	**Veri Boyutu**	**Performans Ölçütü**
Elberse ve Eliashberg, 2003	Ağızdan ağıza iletişim Yeni vizyona giren filmler için ayrılan yer (oynatma alanı) Devam filmler için ayrılan yer (oynatma alanı) İzleyicilerin dikkatini çekmek için rekabet Mevsimsellik Yerli pazar performansı Yerli ve yabancı yayın arasında geçen süre Haftalık gelir Haftalık sahnelenme sayısı İlk haftanın beklenen geliri İlk haftadan sonra beklenen gelir Üretim bütçesi Oyuncu popülaritesi Yönetmen popülaritesi Reklam harcamaları Eleştirel incelemeler Ana dağıtımcı	Dinamik Eşanlı Denklem Modeli	537	tahmin katsayısı Standart hata p-değeri R^2

Tablo 2.5 (Devam): Gişe hasılatı tahmini problemine ilişkin literatür özeti.

Yayın Bilgisi	Kullanılan Değişkenler	Metot	Veri Boyutu	Performans Ölçütü
Sharda ve Delen, 2006	MPAA film reytingleri Aylara bağlı rekabet Oyuncu popülaritesi Film türü Özel efektler Devam filmi olup olmadığı Filmin kaç kez oynatıldığı	İki Saklı Katmanlı Sinir Ağları	834	ODTO
Dellarocas ve diğ., 2007	Vizyon öncesi pazarlama maliyetleri Açılış haftası filmin gösterim sayısı Tanınmış oyuncu varlığı İyi çıkış yapan bir film olup olmadığı MPAA film reytingleri Film türü Mesleki eleştirmenlerin aritmetik ortalama değerlendirmeleri Hafta sonu açılışındaki kullanıcı oylarının aritmetik ortalaması Hafta sonu açılışındaki toplam oy sayısı Açılış hafta sonu değerlendiricilerinin yaş gruplarına göre dağılımının entropisi Açılış hafta sonu değerlendiricilerinin cinsiyet dağılımının entropisi Hafta sonu açılışında gişe hasılatı gelirleri	Difüzyon modelleri (Markov Zinciri Monte Carlo Simülasyonu)	80	OMYH
Duan ve diğ., 2008	Günlük gösterim sayısı Günlük gelir Yorum sayısı Kullanıcı notlarının kümülatif ortalaması Günlük ortalama kullanıcı notu Vizyona girdiğinden beri geçen gün sayısı Hafta sonu olup olmadığını gösteren değişken	Eşanlı Denklemler	71	logaritmik doğrusal fonksiyonun tahmin katsayısı
Zhang ve diğ., 2009	Ülke gücü Yönetmen popülaritesi Oyuncu popülaritesi Reklam Film türü Vizyona girdiği ay Vizyona girdiği gün Hangi Özel gün/Festival'e denk geldiği Aynı anda vizyona girdiği film sayısına göre rekabet durumu Filmin kaç sinemada gösterime girdiği Filmin kaç kez oynatıldığı	Çok Katmanlı Geri Yayılımlı Yapay Sinir Ağları	241	6-kat çapraz geçerleme
Delen ve Sharda, 2010	MPAA film reytingleri Aylara bağlı rekabet Oyuncu popülaritesi Film türü Özel efektler Devam filmi olup olmadığı Filmin kaç kez oynatıldığı	DVM YSA SRA RO HA Füzyon	2632	ODTO
Asur ve Huberman, 2010	Dikkat isteme oranı Düşüncelerin ve incelemelerin kutupsallığı Dağıtım parametresi	ÇDR	24	AOMYH R^2

Tablo 2.5 (Devam): Gişe hasılatı tahmini problemine ilişkin literatür özeti.

Yayın Bilgisi	Kullanılan Değişkenler	Metot	Veri Boyutu	Performans Ölçütü
Kim ve diğ., 2015	P. Haftadaki izleyici sayısı P. Haftadaki gişe hasılatı P. Hafta başındaki oynatılma sayısı P. Haftada toplam oynatılma sayısı P. Hafta başındaki koltuk sayısı P. Haftada toplam koltuk sayısı P. Haftada sosyal ağlarda toplam konuşulma sayısı P. Haftada sosyal ağlarda toplam coşkulu konuşulma sayısı P. Haftada sosyal ağlarda toplam pozitif konuşulma sayısı P. Haftada sosyal ağlarda toplam pozitif konuşulma sayısı	GA(değişken seçimi) ÇDR DVR Gauss Süreç Regresyonu k-EYK	212	OMYH OHKK
Ghiassi ve diğ., 2015	MPAA film reytingleri Devam filmi olup olmadığı Filmin kaç kez oynatıldığı Üretim bütçesi Reklam harcamaları Filmin süresi Mevsimsellik	DYSA	4390	ODTO
Hur ve diğ., 2016	(P + 1). Haftanın başında her film için oynatılma oranı (P + 1). Haftada ortalama film izleyici sayısı P. haftaya kadar her film için ortalama puanlama P. haftaya kadar her film için kayıtlı reyting hacmi P. haftaya kadar her film için olumlu görüşlerin toplamı P. haftaya kadar her film için tarafsız görüşlerin toplamı P. haftaya kadar her film için olumsuz görüşlerin toplamı P. haftaya kadar her film için olumlu görüşlerin oranı P. haftaya kadar her film için tarafsız görüşlerin oranı P. haftaya kadar her film için olumsuz görüşlerin oranı Kore filmi olup olmadığı Hollywood filmi olup olmadığı MPAA film reytingleri Yönetmen popülaritesi Satış gücü Dağıtıcı firma gücü Oyuncu popülaritesi P. Haftaya kadar her film için izleyicilerin toplamı	ÇDR SRA YSA DVR BAU RAU	606	OHKK OMYH

Filmlerin vizyona girmeden önce gişe hasılatı tahmininin 9 kategoriden oluşan sınıflandırma problemi olarak ele alındığı bu çalışmada, yukarıda söz edilen ve yaygın olarak kullanılan makine öğrenmesi esaslı sınıflandırma algoritmalarından YSA ve DVM algoritmaları kullanılmıştır. Literatürdeki çalışmalardan farklı olarak topluluk makine öğrenmesi algoritmalarından biri olan Torbalama algoritması gişe hasılatı tahmin problemi için ilk kez uygulanmıştır. Aynı zamanda yine literatürden farklı olarak problemde doğruluk oranını iyileştirmek amacıyla önerilen yöntemlerin parametre değerleri eniyilenmeye çalışılmıştır.

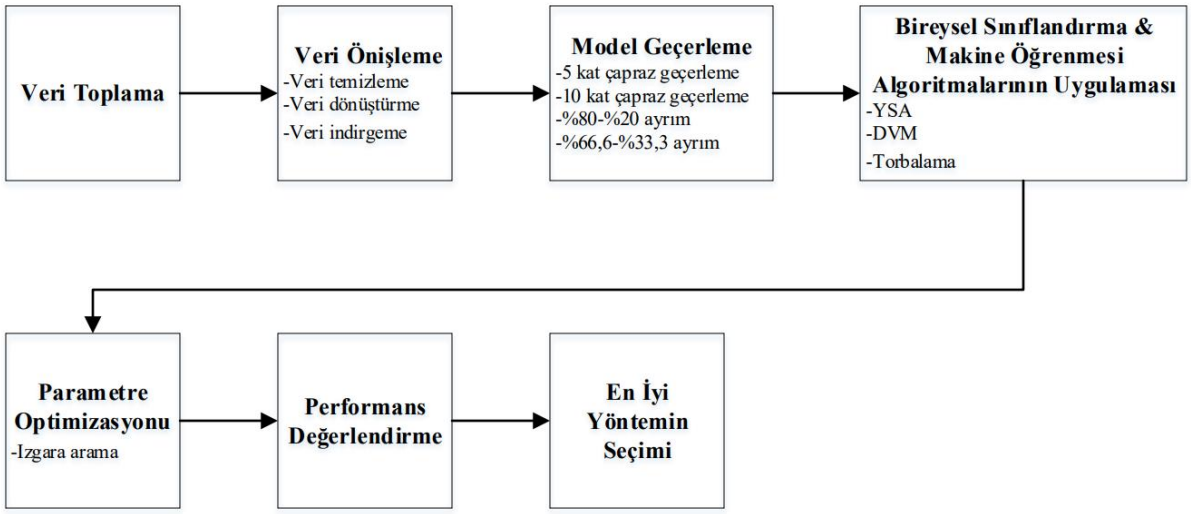
3. MALZEME VE YÖNTEM

3.1.PROBLEMİN TANIMI

Film endüstrisi; günümüzde yönetmenler, yapımcılar, senaristler, oyuncular, teknik ekipler, dağıtımcı firmalar, salonlar ve izleyicilerden oluşan geniş bir sektör haline gelmiştir. Gelişen teknoloji ve değişen dünya koşullarından diğer birçok sektör olduğu gibi film endüstrisi de etkilenmektedir ve film endüstrisi içinde kaçınılmaz bir rekabet ortamı oluşmaktadır. Bu rekabet ortamında sürdürülebilirliğini devam ettirmek isteyen ve piyasada tutunmak isteyen film üreticileri için, bir filme yatırım yapma ve gerekli bütçeyi ayırma kararı ciddi ve bir o kadar da stratejik bir karar haline gelmektedir. Bu çalışmanın amacı; yatırım kararlarına temel oluşturma noktasında film üreticileri başta olmak üzere, teknik ekip, senaristler ve oyuncular gibi sürece dahil olan tüm paydaşlar için büyük önem arz eden ve filmin piyasadaki yankısını gösteren "Filmin gişe hasılatının ne olacağı?" problemini makine öğrenmesi algoritmaları kullanarak tahmin etmeye çalışmaktır.

Tahmin modeli oluşturmada, geleneksel yapay zekâ metotlarından olan YSA ve DVM algoritmaları ile topluluk makine öğrenmesi algoritmalarından biri olan Torbalama algoritması kullanılmıştır. Oluşturulan tahmin modellerinin doğruluğunu iyileştirmek için modellerin parametreleri eniyilenmeye çalışılmıştır.

Gişe hasılatı tahmin probleminin çözümünde uygulanan adımları gösteren akış şeması Şekil 3.1'deverilmiştir.



Şekil 3.1: Çalışmanın akış şeması.

3.2.VERİ SETİNİN OLUŞTURULMASI

Uygulamada kullanılacak veri setinin oluşturulması için öncelikle "Box-office Türkiye[1]" nin internet sitesinden filmlere ait isim, bütçe ve gişe hasılatı bilgileri alınmıştır. Ardından filmlerin isimlerine göre IMDb[2]'nin internet sitesinden OMDb API[3] 'si kullanılarak EK 1'de gösterilen C# kodu ile filmlere ait seçilmiş bazı veriler çekilmiştir. IMDb'nin sitesinden çekilen veriler ilgili filmin yılı, MPAA reytingi, vizyon tarihi, süresi, türü, yönetmeni, yıldız oyuncuları, dili, ülkesi ve metascore'udur. Verilerin bir araya getirilmesinden sonra yaklaşık 3700 filme ait 13 farklı nitelik değeri bilgisi elde edilmiştir. Girdi değişkeni olarak yıl, mpaa, mevsimsellik, süre, tür, yönetmen, oyuncu1, oyuncu2, dil, ülke, metascore ve bütçe değişkenleri kullanılmıştır. Çıktı değişkeni ise gişe hasılatı olarak belirlenmiştir.

Kullanılan değişkenlere ait açıklamalar aşağıda sunulmaktadır:

- Yıl

İlgili filmin çekildiği yıl bilgisini içermektedir.

[1]https://boxofficeturkiye.com.

[2]https://www.imdb.com.

[3]http://www.omdbapi.com.

- MPAA

Amerika'da Amerika Sinema Filmleri Derneği (Motion Picture Association of America) film tarafından kullanılan bir film derecelendirme sistemidir ve seyirciye yayınlanan filmin içeriği hakkında bilgiler vermektedir. Tablo 3.1'de veri setinde kullanılan MPAA derecelendirmeleri gösterilmektedir.

Tablo 3.1: MPAA film derecelendirme sistemi.

Film Derecelendirme Sistemi[4]		
Kodu	**Adı**	**Açıklaması**
G	Genel izleyici	Her yaştan seyirci izleyebilir.
PG	Ebeveyn rehberliği önerilir	Çocuklar için uygun olamayan materyaller içerebilir.
PG-13	Ebeveyn dikkatle uyarılır	Bazı materyaller 13 yaşından küçük çocuklar için uygun olmayabilir.
R	Kısıtlı/Yasak	17 yaşın altındakiler için ebeveyn veya yetişkin birinin eşlik etmesi gereklidir.

- Mevsimsellik

Aylara göre ortalama gişe hasılatları hesaplanarak her aya izlenme oranına göre bir değer verilmiştir. Filmlerin vizyona girmiş olduğu ayın değeri filmin mevsimsellik derecesini ifade etmektedir. Tablo 3.2'de aylara göre mevsimsellik değerleri gösterilmiştir.

Tablo 3.2: Filmlerin aylara göre mevsimsellik değerleri.

Aylar	Box-Office	Derece
Ocak	19898107850	3
Şubat	20781671337	4
Mart	27491091483	7
Nisan	18361430852	2
Mayıs	49888655589	8
Haziran	56800025740	12
Temmuz	53353050974	11
Ağustos	26972042043	6
Eylül	17161343628	1
Ekim	22840506892	5
Kasım	50122945581	9
Aralık	52915005566	10

[4] https://www.mpaa.org/film-ratings/.

- Süre

Filmin süresinin dakika cinsinden değerini ifade etmektedir.

- Tür

Filmin konusuna göre hangi türe ait olduğu bilgisini içermektedir.

- Yönetmen

Filmin yönetmeninin popülerliğini ifade eden bir değerdir. Yönetmen isimlerinin Google'da aranma sayıları Google trends'den alınmıştır. Yönetmenlerin Google'da aranma sayısına göre literatürde yer alan çalışmalara benzer şekilde 1 en az, 6 en çok popülerliği temsil etmek üzere 1 ile 6 arasında derecelendirme yapılmıştır.

- Oyuncu1

Filmin birinci yıldız oyuncusunun popülerliğini ifade eden bir değerdir. Yönetmen popülerliği değerinde olduğu gibi Google aranma sayıları Google trends'den alınarak elde edilmiştir. Birinci yıldız oyuncunun Google'da aranma sayısına göre literatürde yer alan çalışmalara benzer şekilde 1 en az, 6 en çok popülerliği temsil etmek üzere 1 ile 6 arasında derecelendirme yapılmıştır.

- Oyuncu2

Filmin ikinci yıldız oyuncusunun popülerliğini ifade eden bir değerdir. Yönetmen ve oyuncu1 popülerlik değerlerinde olduğu gibi Google aranma sayıları Google trends'den alınarak elde edilmiştir. Filmin ikinci yıldız oyuncusunun Google'da aranma sayısına göre literatürde yer alan çalışmalara benzer şekilde 1 en az, 6 en çok popülerliği temsil etmek üzere 1 ile 6 arasında derecelendirme yapılmıştır.

- Dil

Filmin orijinal dilinin ne olduğu bilgisini içermektedir.

- Ülke

Filmin hangi ülkede yapıldığı bilgisini içermektedir.

- MetaScore

Basından, saygın kuruluşlardan alınan skorların ağırlıklı ortalaması ile elde edilmiş 100 üzerinden bir skor değeridir. Film henüz vizyona girmeden önce, filmin konuşulma düzeyinden ve popülerliğinden elde edilen bir puandır.

- Bütçe

Filmlerin yapımı için ayrılan bütçeye ait sınıf değerini ifade etmektedir. En büyük ve en küçük bütçe arasında 9 eş sınıf oluşturulmuş ve her bir sınıfa bir harf değeri atanmıştır.

- Gişe Hasılatı

Filmlerin gişe hasılatlarına ait sınıf değerini ifade etmektedir. Literatürde bu problemi ele alan diğer çalışmalarda olduğu gibi filmlerin gişe hasılatları "en düşük" den "en yüksek" e kadar değişen dokuz sınıfta kategorize edilmiştir (Sharda & Delen, 2006; Delen & Sharda, 2010; Ghiassi ve diğ., 2015).

Değişkenlere ait tanım, tür ve değer bilgileri Tablo 3.3'de özet olarak gösterilmiştir.

Tablo 3.3: Veri setindeki değişkenler.

No	Değişken	Değişkenin Tanımı	Değişkenin türü	Değişkenin Değeri
1	Yıl	Filmin vizyona girdiği yıl	Nümerik	1927, ..., 2016
2	MPAA	Filmin MPAA reytingi	Kategorik	G, PG, PG-13, R
3	Mevsimsellik	Filmin vizyona girdiği ayın gişe hasılatlarına göre derecesi	Nümerik	1, 2, 3, ..., 12
4	Süre	Filmin süresi	Nümerik	20, 52, 63, ..., 213
5	Tür	Filmin türü	Kategorik	Aksiyon, Animasyon, Macera, Biyografi, Komedi, Suç, Belgesel, Dram, Aile, Fantastik, Tarih, Korku, Gizemli, Müzikal, Romantik, Bilim Kurgu, Polisiye
6	Yönetmen	Filmin yönetmenin popülerliği	Kategorik	0, 1, 2, ..., 6
7	Oyuncu1	Filmin 1. başrol oyuncusunun popülerliği	Kategorik	0, 1, 2, ..., 6

Tablo 3.3 (Devam): Veri setindeki değişkenler.

No	Değişken	Değişkenin Tanımı	Değişkenin türü	Değişkenin Değeri
8	Oyuncu2	Filmin 2. başrol oyuncusunun popülerliği	Kategorik	0, 1, 2, ..., 6
9	Dil	Filmin orijinal dili	Kategorik	İngilizce, Almanca, ...
10	Ülke	Filmin yapıldığı ülke	Kategorik	İngiltere, Amerika, ...
11	MetaScore	Filmin konuşulma puanı	Kategorik	1, 2, ..., 100
12	Bütçe	Film için harcanan bütçe	Kategorik	A, B, C, ..., I
13	Gişe Hasılatı	Filmden elde edilen gişe hasılatı	Kategorik	A, B, C, ..., I

Problemde kullanılmak üzere hazırlanan veri setinin ufak bir kesiti Şekil 3.2'de gösterilmiştir.

Yıl	MPAA	Mevsimsellik	Süre	Tür	Yönetmen	Oyuncu1	Oyuncu2	Dil	Ülke	Meta-Score	Bütçe	Box-Office
2010	G	12	103	Animation	3	5	5	English	USA	92	E	D
2006	G	2	94	Animation	4	5	4	English	Canada	47	B	A
2016	PG-13	11	123	Action	4	6	5	English	UK	58	C	B
1997	R	6	135	Action	4	5	5	English	USA	49	B	A
2016	PG	2	106	Adventure	5	4	5	English	UK	77	D	D

Şekil 3.2: Veri setinin özet görünümü.

3.3.MAKİNE ÖĞRENMESİ ALGORİTMALARININ UYGULANMASI

Veri seti üzerinde gerekli veri önişleme adımlarının uygulanmasından sonra, 10 kat çapraz geçerleme, 5 kat çapraz geçerleme, %80 eğitim-%20 test ve %66,6 eğitim-%33,3 test yöntemleri ile eğitim ve test verileri oluşturulmuştur. Gişe hasılatı tahmin probleminin çözümü için oluşturulan eğitim ve test verileri üzerinde literatürde de yaygın kullanıma sahip olan ve güçlü sonuçlar veren geleneksel yapay zeka algoritmalarından YSA ve DVM algoritmaları, bileşik makine öğrenmesi algoritmalarından Torbalama algoritması uygulanmıştır. Önerilen algoritmaların uygulanması MATLAB 2017a yazılımı kullanılarak gerçekleştirilmiştir.

YSA algoritmasına ait MATLAB kodu EK 2'de gösterilmiştir. İlk aşamada MATLAB yazılımının varsayılan parametre değerleri kullanılmıştır.

YSA algoritmasının öğrenme oranı ve gizli katman sayısı parametreleri için varsayılan değerler Tablo 3.4'de gösterildiği gibidir.

Tablo 3.4: YSA algoritmasının varsayılan parametre değerleri.

Algoritma	Parametreler		
YSA	varsayılan	Öğrenme oranı	0,01
		Gizli katman sayısı	1

DVM algoritmasına ait MATLAB kodu EK 3'de gösterilmiştir. İlk aşamada MATLAB yazılımının varsayılan parametre değerleri kullanılmıştır. DVM algoritmasının çekirdek fonksiyonu, maliyet ve kutu-kısıtı parametreleri için varsayılan değerler Tablo 3.5'de gösterildiği gibidir.

Tablo 3.5: DVM algoritmasının varsayılan parametre değerleri.

Algoritma	Parametreler		
DVM	varsayılan	Çekirdek fonksiyonu	doğrusal
		Maliyet	1
		Kutu-kısıt	1

Torbalama algoritmasına ait MATLAB kodu EK 4'de gösterilmiştir. İlk aşamada MATLAB yazılımının varsayılan parametre değerleri kullanılmıştır. Torbalama algoritmasının zayıf öğrenici sayısı ve maliyet parametreleri için varsayılan değerler Tablo 3.6'da gösterildiği gibidir.

Tablo 3.6: Torbalama algoritmasının varsayılan parametre değerleri.

Algoritma	Parametreler		
Torbalama	varsayılan	Zayıf öğrenici sayısı	100
		Maliyet	1

3.4.PARAMETRE ENİYİLEMESİ İLE ÖNERİLEN YÖNTEMLERİN PERFORMANSLARININ İYİLEŞTİRİLMESİ

Sınıflandırma modellerinin parametre değerlerinin eniyilenmesi modellerin başarımını önemli ölçüde iyileştirdiği literatür çalışmalarında görülmüştür. Bu doğrultuda bu çalışmada önerilen sınıflandırma modellerinin en iyi parametre değerlerini bulmak için ızgara arama algoritması kullanılmıştır. Parametre değerleri eniyilenirken amaç fonksiyonu olarak doğruluk oranı kullanılmıştır.

Izgara Arama Algoritması

Izgara arama algoritması, arama ve işleme sürecini başarılı bir şekilde gerçekleştiren bir eniyileme tekniğidir. Arama süreci, başlangıç fonksiyonundan hedef durumuna kadar, mesafe fonksiyonu olarak adlandırılan "minimum maliyet" haritası oluşturur. İşleme süreci, hedeften başlangıç yapılandırmasına kadar bir geri dönüştür (Petres ve diğ., 2007).

Izgara arama algoritmasının temel ilkesi, belirli bir alan aralığını ızgaraya bölmek ve daha sonra modelin en uygun parametrelerini bulmak için ızgaradaki tüm noktaları çaprazlayarak dikkatlice incelemektir. Izgara arama algoritması kullanıldığında, eniyileme için her parametre nispeten bağımsızdır, bu yüzden de paralel optimizasyon gerçekleştirilebilmektedir (Wenhai & Jiandong, 2014; Huang ve diğ., 2012).

Çalışmada tahmin modellerini geliştirmek amacıyla eniyilenmeye çalışılan parametreler aşağıda verilmiştir:

YSA'nın parametreleri;

Birçok parametreye sahip YSA algoritması için öğrenme oranı ve gizli katman sayısı parametreleri enyilenmeye çalışılmıştır. YSA algoritmasının en iyi parametre değerlerini arayan MATLAB kodu EK 5'de gösterilmiştir.

Öğrenme oranı

Öğrenme oranı, her bir eğitim döngüsünde giriş-çıkış eğitim modellerinin hata fonksiyonunun azaltılmasının sağlanmasıyla belirlenir (Tang, 1996).Öğrenme oranı parametresi, ağın özel

uygulamasına bağlı değildir ve tipik olarak 0 ile 1 arasında bir değerdir (Marmanis& Babenko, 2009).

Gizli katman sayısı

YSA'da ilk katman ağın giriş nöronlarını içerir ve problemin eğitim verileri, bu giriş nöronları aracılığıyla ağa sunulur. Son katman ise, giriş nöronlarında sunulan herhangi bir durum için sinir ağının tahminlerini içerir. Bu giriş ve çıkış katmanları arasında genellikle bir veya daha fazla "gizli" katman bulunur (Torgo, 2011).Literatürde YSA uygulamasında gizli katman sayısının 1 ile 20 arasındaki değerler için performans ölçeğine etkisinin incelendiği çalışmalar görülmüştür (Nassif, 2018).

DVM'nin parametreleri;

DVM algoritması için çekirdek fonksiyonu, maliyet ve kutu kısıtı parametreleri enyilenmeye çalışılmıştır. DVM algoritmasının en iyi parametre değerlerini arayan MATLAB kodu EK 6'da gösterilmiştir.

Çekirdek fonksiyonu

Çekirdek fonksiyonunun seçimi DVM'nin performansını belirleyen önemli bir parametredir (Lin & Liu, 2007).

DVM için dört temel çekirdek fonksiyonu bulunmaktadır:

- Doğrusal Fonksiyon
- Polinomial Fonksiyon
- Radyal Temelli Fonksiyon
- Gauss Fonksiyon

Maliyet

Maliyet parametresi (C), ayırıcı hiperdüzlemin düzgünlüğü ile eğitim verilerinin yanlış sınıflandırılması arasındaki dengeyi kontrol eder (Cortes & Vapnik, 1995; Güldoğan ve diğ., 2017).Karmaşıklıkla hata terimi arasındaki dengeyi kontrol eden C (maliyet) parametresi, düzenleyici sabit olarak da adlandırılmaktadır (Tolun, 2015).

Literatüre bakıldığında DVM algoritması için maliyet parametresi genellikle 1 ile 1000 arasında ya da 1, 3, 10, 50, 100 gibi değerler almaktadır (Alsufyani, 2017; Purver, 2014).

Kutu Kısıtı

Kutu-kısıtı, marjı ihlal eden gözlemler üzerinde uygulanan maksimum cezayı kontrol eden bir parametredir ve aşırı uyum göstermeyi önlemeye yardımcı olur. Kutu kısıtlamasını arttırmak, DVM sınıflandırıcısında daha az destek vektörü atar. Bununla birlikte, kutu kısıtının artırılması daha uzun eğitim sürelerine yol açabilir (Mathworks, 2018).

Torbalama algoritmasının parametreleri;

Bileşik makine öğrenme algoritmalarından biri olan Torbalama algoritması için zayıf öğrenici sayısı ve maliyet parametreleri enyilenmeye çalışılmıştır. Torbalama algoritmasının en iyi parametre değerlerini arayan MATLAB kodu EK 7'de gösterilmiştir.

Zayıf öğrenici sayısı

Bileşik makine öğrenme algoritmaları, güçlü bir öğrenici sağlamak için birçok zayıf öğrenciyi bir arada kullanmaktadır (Farzi & Bolandi, 2016).

Doğruluk oranı üzerinde farklı etkiler gösterebilen zayıf öğrenici sayısı için literatürde 5, 10, 50, 100 ve 100, 1000, 10000 değerlerinin kullanıldığı görülmüştür (Atasever & Özkan, 2012; Pang ve diğ., 2013).

Maliyet

Torbalama yöntemi gibi maliyete duyarlı algoritmalarda, maliyetler yanlış sınıflandırma toplamını en aza indirmeyi amaçlamaktadır (Hido ve diğ., 2009). Literatürde genellikle 1, 10, 50, 100, 1000, 10000 gibi değerler verilmiş olan yanlış sınıflandırma maliyetini arttırmak yanlış sınıflandırma oranını azaltacaktır (Lomax & Vadera, 2016).

Önerilen algoritmaların yukarıda bahsedilen parametre değerlerini en iyilemek için literatüre de uygun olarak parametre aralıkları belirlenmeye çalışılmıştır. Parametre aralıkları belirlenirken varsayılan değerlerin de aralığın içine dahil edilmesine dikkat edilmiştir.

Çalışmada, makine öğrenmesi algoritmalarının performansını arttırmak için Tablo 3.7'de verilen parametre aralık değerleri için Izgara arama algoritması kullanılarak eniyileme işlemi gerçekleştirilmiştir.

Tablo 3.7: Çalışmada kullanılan algoritmaların parametrelerinin değer aralıkları.

Algoritma	Parametreler	Parametre için değer aralığı
YSA	Öğrenme oranı	0,05; 0,10; 0,15; … ; 0,95
	Gizli katman sayısı	1; 2; … ; 25
DVM	Çekirdek fonksiyonu	doğrusal, polinomial, gauss, radyal tabanlı
	Maliyet	1×10^{-3},1×10^{-2}, … , 1×10^{6}
	Kutu-kısıt	1×10^{-6},1×10^{-5}, … , 1×10^{5}
Torbalama	Zayıf öğrenici sayısı	100; 200; 300; … ; 1000
	Maliyet	1×10^{-3},1×10^{-2}, … , 1×10^{6}

4. BULGULAR

Bu bölümde, farklı model geçerleme yöntemleri için önerilen sınıflandırma algoritmalarının varsayılan ve eniyilenmiş parametre değerleri ile performans analizi sonuçları sunulmuştur.

10-kat çapraz geçerleme ile eğitim ve test verisi olarak ayrılmış veri seti için algoritmaların varsayılan ve eniyilenmiş parametre değerleri için performans ölçütleri açısından değerlendirmeleri Tablo 4.1'de gösterilmiştir.

Tablo 4.1: Sınıflandırma modellerinin performans analizi (çapraz geçerleme 10-kat ile).

Algoritmalar	Parametreler			Doğruluk Oranı	Kesinlik	Anma	F-Ölçütü	Kappa İstatistiği
YSA	varsayılan	Öğrenme oranı	0,01	0,7460	0,6884	0,7460	0,6906	0,1986
		Gizli katman sayısı	1					
	eniyilenmiş	Öğrenme oranı	0,9	0,9241	0,9452	0,9241	0,9330	0,4851
		Gizli katman sayısı	4					
DVM	varsayılan	Çekirdek fonksiyonu	doğrusal	0,9091	0,9685	0,9091	0,9367	0,2265
		Maliyet	1					
		Kutu-kısıt	1					
	eniyilenmiş	Çekirdek fonksiyonu	doğrusal	0,9173	0,9559	0,9173	0,9343	0,3670
		Maliyet	10					
		Kutu-kısıt	10					
Torbalama	varsayılan	Zayıf öğrenici sayısı	100	0,9173	0,9590	0,9173	0,9356	0,3484
		Maliyet	1					
	eniyilenmiş	Zayıf öğrenici sayısı	100	0,9208	0,9595	0,9208	0,9375	0,3838
		Maliyet	1000					

Tablo 4.1'de en iyi doğruluk oranına sahip olan YSA algoritmasının en iyilenmiş parametre değerleri ile uygulanmasında 3700 film verisinden 3384 tanesi doğru tahmin edilmiştir. Sınıf dağılımları test verisinde bulunan 9 kategori arasında farklılık göstermektedir.

5-kat çapraz geçerleme ile eğitim ve test verisi olarak ayrılmış veri seti için algoritmaların varsayılan ve eniyilenmiş parametre değerleri için performans ölçütleri açısından değerlendirmeleri Tablo 4.2'de gösterilmiştir.

Tablo 4.2: Sınıflandırma modellerinin performans analizi (çapraz geçerleme 5-kat ile).

Algoritmalar	Parametreler			Doğruluk Oranı	Kesinlik	Anma	F-Ölçütü	Kappa İstatistiği
YSA	varsayılan	Öğrenme oranı	0,01	0,8979	0,9020	0,8979	0,8973	0,4466
		Gizli katman sayısı	1					
	eniyilenmiş	Öğrenme oranı	0,3	0,9238	0,9443	0,9238	0,9329	0,4766
		Gizli katman sayısı	5					
DVM	varsayılan	Çekirdek fonksiyonu	doğrusal	0,9110	0,9711	0,9110	0,9390	0,2409
		Maliyet	1					
		Kutu-kısıt	1					
	eniyilenmiş	Çekirdek fonksiyonu	doğrusal	0,9153	0,9568	0,9153	0,9336	0,3427
		Maliyet	0,1					
		Kutu-kısıt	10					
Torbalama	varsayılan	Zayıf öğrenici sayısı	100	0,9153	0,9599	0,9153	0,9352	0,3257
		Maliyet	1					
	eniyilenmiş	Zayıf öğrenici sayısı	100	0,9208	0,9610	0,9208	0,9387	0,3830
		Maliyet	0,1					

Tablo 4.2'de en iyi doğruluk oranına sahip olan YSA algoritmasının en iyilenmiş parametre değerleri ile uygulanmasında 3700 film verisinden 3383 tanesi doğru tahmin edilmiştir. Sınıf dağılımları test verisinde bulunan 9 kategori arasında farklılık göstermektedir.

%80 eğitim-%20 test yöntemleri ile eğitim ve test verileri oluşturulmuş veri seti için algoritmaların varsayılan ve eniyilenmiş parametre değerleri için performans ölçütleri açısından değerlendirmeleri Tablo 4.3'de gösterilmiştir.

Tablo 4.3: Sınıflandırma modellerinin performans analizi (%80 eğitim-%20 test).

Algoritmalar	Parametreler			Doğruluk Oranı	Kesinlik	Anma	F-Ölçütü	Kappa İstatistiği
YSA	varsayılan	Öğrenme oranı	0,01	0,7400	0,6234	0,7400	0,6708	0,2166
		Gizli katman sayısı	1					
	eniyilenmiş	Öğrenme oranı	0,8	0,9227	0,9422	0,9227	0,9313	0,4733
		Gizli katman sayısı	1					
DVM	varsayılan	Çekirdek fonksiyonu	doğrusal	0,9085	0,9390	0,9085	0,9228	0,3626
		Maliyet	1					
		Kutu-kısıt	1					
	eniyilenmiş	Çekirdek fonksiyonu	doğrusal	0,9147	0,9618	0,9147	0,9361	0,3252
		Maliyet	0,001					
		Kutu-kısıt	1					
Torbalama	varsayılan	Zayıf öğrenici sayısı	100	0,9160	0,9619	0,9160	0,9361	0,3203
		Maliyet	1					
	eniyilenmiş	Zayıf öğrenici sayısı	200	0,9188	0,9597	0,9188	0,9374	0,3395
		Maliyet	1					

Tablo 4.3'de en iyi doğruluk oranına sahip olan YSA algoritmasının en iyilenmiş parametre değerleri ile uygulanmasında 3700 film verisinden 3379 tanesi doğru tahmin edilmiştir. Sınıf dağılımları test verisinde bulunan 9 kategori arasında farklılık göstermektedir.

%66,6 eğitim-%33,3 test yöntemleri ile eğitim ve test verileri oluşturulmuş veri seti için algoritmaların varsayılan ve eniyilenmiş parametre değerleri için performans ölçütleri açısından değerlendirmeleri Tablo 4.4'de gösterilmiştir.

Tablo 4.4: Sınıflandırma modellerinin performans analizi (%66,6 eğitim-%33,3 test).

Algoritmalar	Parametreler			Doğruluk Oranı	Kesinlik	Anma	F-Ölçütü	Kappa İstatistiği
YSA	varsayılan	Öğrenme oranı	0,01	0,8157	0,8009	0,8157	0,7861	0,2934
		Gizli katman sayısı	1					
	eniyilenmiş	Öğrenme oranı	0,05	0,9227	0,9370	0,9227	0,9292	0,4933
		Gizli katman sayısı	3					
DVM	varsayılan	Çekirdek fonksiyonu	doğrusal	0,9102	0,9814	0,9102	0,9435	0,2240
		Maliyet	1					
		Kutu-kısıt	1					
	eniyilenmiş	Çekirdek fonksiyonu	doğrusal	0,9148	0,9672	0,9148	0,9379	0,3124
		Maliyet	0,001					
		Kutu-kısıt	1					
Torbalama	varsayılan	Zayıf öğrenici sayısı	100	0,9168	0,9623	0,9168	0,9374	0,3208
		Maliyet	1					
	eniyilenmiş	Zayıf öğrenici sayısı	200	0,9218	0,9648	0,9218	0,9407	0,3471
		Maliyet	100000					

Tablo 4.4'de en iyi doğruluk oranına sahip olan YSA algoritmasının en iyilenmiş parametre değerleri ile uygulanmasında 3700 film verisinden 3379 tanesi doğru tahmin edilmiştir. Sınıf dağılımları test verisinde bulunan 9 kategori arasında farklılık göstermektedir.

5. TARTIŞMA VE SONUÇ

Yaklaşık 250 yıllık bir geçmişe sahip olan sinema filmleri, yıllar içerisinde gelişimini tamamlayarak; günümüzde birçok paydaştan oluşan geniş bir sektör haline gelmiştir. Günümüzde birçok sektörde olduğu gibi teknolojinin sürekli olarak gelişmesi, değişen dünya, endüstrileşmenin hızla ve her alanda artması film endüstrisini de etkilemiştir. Bu doğrultuda, bir filme yatırım yapma ve gerekli bütçeyi ayırma kararı önemli ve bir o kadar da stratejik bir karar haline gelmektedir. Mevcut karar destek sistemlerinin yetersiz kaldığı ve yüksek güvenilirlikle sonuçlar veremediği bu tarz problemlerde makine öğrenmesi algoritmaları kullanarak üretim öncesi güçlü tahminler yapmak ilgi çekici bir konu haline gelmiştir.

Bu çalışmada; yatırım kararlarına temel oluşturma noktasında, vizyona girmeden önce filmlerin gelir tahmininin yapılması için makine öğrenmesi teknikleri kullanılarak tahmin modelleri oluşturulmuştur. Tahmin modeli oluşturmada, geleneksel yapay zekâ metotlarından olan YSA ve DVM ile topluluk makine öğrenmesi algoritmalarından biri olan Torbalama algoritması kullanılmıştır. Oluşturulan tahmin modellerinin doğruluğunu iyileştirmek için modellerin parametreleri Izgara Arama algoritması kullanılarak eniyilenmeye çalışılmıştır. Çalışmada, hem parametre eniyilemenin modelin tahmin başarısına etkisini gözlemlemek hem de geleneksel yapay zekâ algoritmaları ile topluluk makine öğrenmesi algoritmasının başarısını karşılaştırmak hedeflenmiştir. Bu hedef doğrultusunda tahmin modellerinin uygulanmasında, 10 kat çapraz geçerleme, 5 kat çapraz geçerleme, %80 eğitim-%20 test ve %66,6 eğitim-%33,3 test yöntemleri ile eğitim ve test verileri oluşturulmuştur ve tahmin modellerinin doğruluk oranı, kesinlik, anma, F-ölçütü ve kappa istatistiği performans ölçütlerinin altında varsayılan ve eniyilenmiş parametre değerleri için karşılaştırılmalı analizine Tablo 4.1, Tablo 4.2, Tablo 4.3, ve Tablo 4.4'de yer verilmiştir.

10-kat çapraz geçerlemenin kullanıldığı durumda önerilen sınıflandırma modellerinin performans analizi için Tablo 4.1 incelendiğinde, 0,9241 doğruluk oranı ile en yüksek başarıyı YSA algoritmasının verdiği görülmektedir. Diğer performans ölçütleri açısından incelediğimizde; DVM 0,9685 değeri ile kesinlik ölçütü için en iyi performansa sahiptir. YSA 0,9241'lik değeri ile anma 0,4851'lik değeri ile kappa istatistiği ölçütleri için en iyi

performansı sergilemiştir. Son olarak F-ölçütü için en iyi değeri 0,9375 ile Torbalama algoritması vermektedir.

5-kat çapraz geçerlemenin kullanıldığı durumda önerilen sınıflandırma modellerinin performans analizi için Tablo 4.2 incelendiğinde, 0,9238 doğruluk oranı ile en yüksek başarıyı YSA algoritmasının verdiği görülmektedir. Diğer performans ölçütleri açısından bakacak olursak; DVM 0,9711 değeri ile kesinlik ölçütü için ve 0,9390 değeri ile F-ölçütü için en iyi performansa sahiptir. YSA 0,9238'lik değeri ile anma ve 0,4766'lık değeri ile ise kappa istatistiği ölçütleri için en iyi performansı sergilemiştir.

%80 eğitim-%20 test verisi olarak bölündüğü durumda önerilen sınıflandırma modellerinin performans analizi için Tablo 4.3incelendiğinde, 0,9227 doğruluk oranı ile en yüksek başarıyı YSA algoritmasının verdiği görülmektedir. Diğer performans ölçütleri açısından baktığımızda ise; Torbalama algoritması 0,9616 değeri ile kesinlik ölçütü için 0,9374'lük değeri ile de kappa istatistiği ölçütü için en iyi performansa sahiptir. Geri kalan anma ve kappa istatistiği ölçütleri için sırasıyla 0,9227 ve 0,4733 değerleri ile en iyi performansı sergileyen algoritma YSA olmuştur.

%66,6 eğitim-%33,3 test verisi olarak bölündüğü durumda önerilen sınıflandırma modellerinin performans analizi için Tablo 4.4incelendiğinde, 0,9227 doğruluk oranı ile en yüksek başarıyı YSA algoritmasının verdiği görülmektedir. Diğer performans ölçütleri açısından bakıldığında; 0,9814 değeri ile kesinlik ölçütü için ve 0,9435 değeri ile F-ölçütü için en iyi performansa sahip olan algoritma DVM'dir. Anma ve kappa istatistiği ölçütleri için YSA algoritmasının0,9227 ve 0,4933 değerleri ile en iyi performansı sergilediği görülmüştür.

Sınıflandırma modellerinin incelenmesi için sunulan dört özet tabloya da baktığımızda YSA algoritmasının gişe hasılatının tahmini problemi için oldukça iyi sonuçlar verdiğini söyleyebiliriz. Bununla birlikte YSA gibi bireysel bir makine öğrenme algoritması olan DVM ve bileşik makine öğrenme algoritması olan Torbalama algoritmaları da yüksek doğruluk oranına sahip sonuçlar sunmaktadır. Bireysel makine öğrenmesi algoritmalarından olan YSA'nın topluluk makine öğrenmesi algoritmalarından olan Torbalama algoritması ile karşılaştırılabilir derecede iyi sonuçlar verdiği gözlemlenmiştir.

Önerilen yöntemlerin performans değerlerinin arttırılması için parametre eniyilemenin modeller üzerindeki etkisini inceleyecek olursak, dört farklı model geçerleme yöntemi ile

sunulan üç sınıflandırma modeli için de parametre eniyilemenin doğruluk oranını kesin olarak arttırıcı etki gösterdiğini söyleyebiliriz.

Gelecekte çalışmanın geliştirilmesi için gişe hasılatının tahmin probleminde Aşırı Öğrenme Makinesi gibi farklı sınıflandırma algoritmaları kullanılarak sonuçlar iyileştirilmeye çalışılabilir. Daha güçlü sonuçlar elde edebilmek için melez yaklaşımlar önerilebilir. Bu çalışmada iyi sonuç verdiği gözlenen YSA ve Torbalama Algoritmasının ve DVM ve Torbalama algoritmasının melez yaklaşımları problemin çözümü için önerilebilir. Ayrıca literatürde şuana kadar sınıflandırma problemi olarak ele alınmış olan gişe hasılatının tahmin problemi tamamen nümerik değerlerin kullanılmasıyla sınıflandırma probleminden ziyade nümerik bir tahmin problemi olarak çözülebilir.

KAYNAKLAR

Aggarwal, C. C., 2015, *Data mining: the textbook*, Springer, Cham, Switzerland, ISBN 978-3-319-14141-1.

Ainslie, A., Dreze, X., Zufryden, F., Drèze, X., & Zufryden, F., 2005, Modeling Movie Life Cycles and Market Share, *Marketing Science*, 24(3), 508–517.

Alon, I., Qi, M., & Sadowski, R. J., 2001, Forecasting aggregate retail sales: A comparison of artifcial neural networks and traditional methods, *Journal of Retailing and Consumer Services*, 8(3), 147–156.

Alsufyani, A., 2017, A comparison of classification algorithms for Event Related Potentials. *In Computer Applications and Information Processing Technology (CAIPT), 4th International Conference on*, (pp. 1-5), 2017 Ağustos, IEEE.

Asur, S., & Huberman, B. A., 2010, Predicting the future with social media, *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology*, 492–499.

Atasever, Ü. H., & Özkan, C., 2012, Arazi Örtüsünün Belirlenmesinde Torbalama-Karar Ağaçları Yönteminin Kullanımı, *IV. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu(UZAL-CBS 2012)*, 16-19 Ekim 2012, Zonguldak.

Azari, A., Janeja, V. P., & Mohseni, A., 2012, Predicting hospital length of stay (PHLOS): A multi-tiered data mining approach, *In Data Mining Workshops (ICDMW)*, pp. 17-24.

Bauer, E., & Kohavi, R., 1999, An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, *Machine learning*, 36(1-2), 105-139.

Benfu, Y., Hongmei, S., Ye, S., Xiuhui, L., & Bin, Z., 2009, Study on the artificial neural network in the diagnosis of smear negative pulmonary tuberculosis, *Computer Science and Information Engineering*, 5, pp. 584-588.

Borzemski, L., 2007, Internet path behavior prediction via data mining: Conceptual

framework and case study, *Journal of Universal Computer Science*, 13(2), 287-316.

Breiman, L., 1996, Bagging predictors, *Machine learning*, 24(2), 123-140.

Caruana, R., & Niculescu-Mizil, A., 2004, Data mining in metric space: an empirical analysis of supervised learning performance criteria, *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 69-78.

Castillo, P. A., Mora, A. M., Faris, H., Merelo, J. J., García-Sánchez, P., Fernández-Ares, A. J., … García-Arenas, M. I., 2016, Applying Computational Intelligence Methods for Predicting the Sales of Newly Published Books in a Real Editorial Business Management Environment, *Knowledge-Based Systems*, 115, 133–151.

Chakrabarti, S., Cox, E., Frank, E., Güting, R. H., Han, J., Jiang, X., ... & Pyle, D., 2009,*Data mining: know it all*. Morgan Kaufmann, San Francisco, ISBN: 978-0-12-374629-0.

Chang, P. C., & Lai, C. Y., 2005, A hybrid system combining self-organizing maps with case-based reasoning in wholesaler's new-release book forecasting, *Expert Systems withApplications.*

Chang, P. C., Lai, C. Y., & Lai, K. R., 2006, A hybrid system by evolving case-based reasoning with genetic algorithm in wholesaler's returning book forecasting, *Decision Support Systems*, 42(3), 1715–1729.

Chaurasia, V., & Pal, S., 2013, Early prediction of heart diseases using data mining techniques, *Caribbean Journal of Science and Technology*, Vol.1, 208-217.

Chen, F. L., & Ou, T. Y., 2011, Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry, *Expert Systems with Applications*, 38(3), 1336–1345.

Chinchor N., 1992, MUC-4 Evaluation Metrics, *in Proc. of the Fourth Message Understanding Conference*, pp. 22–29.

Choi, T. M., Hui, C. L., Liu, N., Ng, S. F., & Yu, Y., 2014, Fast fashion sales forecasting with limited data and time, *Decision Support Systems*, 59(1), 84–92.

Chou, J. S., Tsai, C. F., Pham, A. D., & Lu, Y. H., 2014, Machine learning in concrete strength simulations: Multi-nation data analytics, *Construction and Building Materials*, *73*, 771–780.

Cichosz, P., 2015, *Data Mining Algorithms: Explained Using R*, John Wiley & Sons, UK, ISBN: 9781118332580.

Cortes, C., & Vapnik, V., 1995, Support-vector networks, *Machine learning*, 20(3), 273-297.

Dangare, C. S., & Apte, S. S., 2012, Improved study of heart disease prediction system using data mining classification techniques, *International Journal of Computer Applications*, 47(10), 44-48.

Delen, D., & Sharda, R., 2010, Predicting the financial success of hollywood movies using an information fusion approach, *Industrial Engineering Journal*, 21(1), 30-37.

Delen, D., Sharda, R., & Kumar, P., 2007, Movie forecast Guru: A Web-based DSS for Hollywood managers, *Decision Support Systems*, 43(4), 1151–1170.

Delen, D., Walker, G., & Kadam, A., 2005, Predicting breast cancer survivability: a comparison of three data mining methods, *Artificial intelligence in medicine*, 34(2), 113-127.

Delgado, M., Sánchez, D., MartıN-Bautista, M. J., & Vila, M. A., 2001, Mining association rules with improved semantics in medical databases, *Artificial Intelligence in Medicine*, 21(1-3), 241-245.

Dellarocas, C., Zhang, X., & Awad, N. F., 2007, Exploring the value of online product reviews in forecasting sales: The case of motion pictures, *Journal of Interactive Marketing*, 21(4), 23–45.

Dietterich, T. G., 2000, Ensemble methods in machine learning, *In International workshop on multiple classifier systems*, (pp. 1-15), Springer, Berlin, Heidelberg.

Downes, A., & Hartley, P., 2006, Using an artificial neural network to assist roll design in cold roll-forming processes,*Journal of Materials Processing Technology*, 177(1-3), 319-322.

Duan, W., Gu, B., & Whinston, A. B., 2008, The dynamics of online word-of-mouth and product sales-An empirical investigation of the movie industry, *Journal of Retailing*, 84(2), 233–242.

Elberse, A., & Eliashberg, J., 2003, Demand and Supply Dynamics for Sequentially Released Products in International Markets: The Case of Motion Pictures, *Marketing Science*, 22(3), 329–354.

Erdal, H. I., Karakurt, O., & Namli, E., 2013, High performance concrete compressive strength forecasting using ensemble models based on discrete wavelet transform, *Engineering Applications of Artificial Intelligence*, *26*(4), 1246–1254.

Farzi, R., & Bolandi, V.,2016, Estimation of organic facies using ensemble methods in comparison with conventional intelligent approaches: a case study of the South Pars Gas Field, Persian Gulf, Iran,*Modeling Earth Systems and Environment*, 2(2), 105.

Fernandez, G., 2010, *Statistical data mining using SAS applications*, CRC Press, Boca Raton, ISBN-13: 978-1-4398-1076-7 (Ebook).

Fulk, G. D., & Sazonov, E., 2011, Using sensors to measure activity in people with stroke, *Topics in stroke rehabilitation*, 18(6), 746-757.

García, S., Luengo, J., & Herrera, F., 2015, *Data preprocessing in data mining*, Springer, Cham, Switzerland, ISBN 978-3-319-10246-7.

Ghiassi, M., Lio, D., & Moon, B., 2015, Pre-production forecasting of movie revenues with a dynamic artificial neural network, *Expert Systems with Applications*, 42(6), 3176–3193.

Goyal, A., & Mehta, R., 2012, Performance comparison of Naïve Bayes and J48 classification algorithms, *International Journal of Applied Engineering Research*, 7(11), 2012.

Gupta, S., Kumar, D., & Sharma, A., 2011, Data mining classification techniques applied for breast cancer diagnosis and prognosis, *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(2), 188-195.

Güldoğan, E., Arslan, A. K., & Yağmur, J., 2017, Çeşitli Çekirdek Fonksiyonları ile Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bir

Klinik Uygulama, *Firat Tip Dergisi*, 22(3).

Han, J. & Kamber, M., 2006, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, ISBN 13: 978-1-55860-901-3.

Hand, D. J., 2007, *Principles of data mining*, Drug safety, 30(7), 621-622.

He, B., & Chang, K. C. C., 2004, Making holistic schema matching robust: An ensemble framework with sampling and voting, *Technical Report UIUCDCS-R-2004-2451*, Dept. of Computer Science, UIUC.

Hido, S., Kashima, H., & Takahashi, Y., 2009, Roughly balanced bagging for imbalanced data, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2(5- 6), 412-426.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J., 2003, A practical guide to support vector classification, Department of Computer Science, National Taiwan University.

Huang, Q., Mao, J., & Liu, Y., 2012, An improved grid search algorithm of SVR parameters optimization. *In Communication Technology (ICCT)*, pp. 1022-1026, IEEE.

Hur, M., Kang, P., & Cho, S., 2016, Box-office forecasting based on sentiments of movie reviews and Independent subspace method, *Information Sciences*, 372, 608–624.

Jordan, M. I., & Mitchell, T. M., 2015, Machine learning: Trends, perspectives, and prospects, *Science*, 349(6245), 255-260.

Jun, D. B., Kim, D. S., & Kim, J., 2011, A Bayesian Dymimic Model for Forecasting Movie Viewers, *KAIST Business School Working Paper Series*, No. 2011-003.

Kavzoğlu, T., Çölkesen, İ., 2010, Destek Vektör Makineleri İle Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi, *Harita Dergisi*, 144(7), 73-82.

Kim, T., Hong, J., & Kang, P., 2015, Box office forecasting using machine learning algorithms based on SNS data, *International Journal of Forecasting*.

Kotsiantis, S., & Kanellopoulos, D., 2006, Association rules mining: A recent overview, *GESTS International Transactions on Computer Science and Engineering*, 32(1), 71-82.

Kulkarni, G., Kannan, P. K., & Moe, W., 2012, Using online search data to forecast new product sales, *Decision Support Systems*, 52(3), 604–611.

Landis, J. R., & Koch, G. G., 1977, The measurement of observer agreement for categorical data, *Biometrics*, 159-174.

Lee, H., Kim, S. G., Park, H. woo, & Kang, P., 2014, Pre-launch new product demand forecasting using the Bass model: A statistical and machine learning-based approach, *Technological Forecasting and Social Change*, 86, 49–64.

Lee, J., Lee, C. Y., & Lee, K. S., 2012, Forecasting demand for a newly introduced product using reservation price data and Bayesian updating, *Technological Forecasting and Social Change*, 79(7), 1280–1291.

Lin, S. L., & Liu, Z., 2007, Parameter selection in SVM with RBF kernel function. *Journal-Zhejiang University of Technology*, 35(2), 163.

Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G., 2005, Selecting thresholds of occurrence in the prediction of species distributions, *Ecography*, 28(3), 385-393.

Liu, L., Shen, B., & Wang, X., 2014, Research on kernel function of support vector machine. In Advanced Technologies, *Embedded and Multimedia for Human-centric Computing* (pp. 827-834). Springer Netherlands.

Lomax, S., & Vadera, S., 2016, A cost-sensitive decision tree learning algorithm based on a multi-armed bandit framework, *The Computer Journal*, 60(7), 941-956.

Lu, C. J., Lee, T. S., & Lian, C. M., 2012, Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks, *Decision Support Systems*, 54(1), 584–596.

Madhulatha, T. S.,2012, An overview on clustering methods,*(IOSR) Journal of Engineering*, 2(4),719-725.

Maimon, O., & Rokach, L., 2010, *Data mining and knowledge discovery handbook(Second Edition)*, Springer, New York, ISBN: 978-0-387-09822-7.

Marmanis, H., & Babenko, D., 2009, *Algorithms of the intelligent web*, Manning, Greenwich, ISBN 978-1-933988-66-5.

Mastrogiannis, N., Boutsinas, B., & Giannikos, I., 2009, A method for improving the accuracy of data mining classification algorithms, *Computers & Operations Research*, 36(10), 2829-2839.

Mathworks, 2018, *fitcsvm: Train binary support vector machine (SVM) classifier*, https://www.mathworks.com/help/stats/fitcsvm.html#bt7oo83-5, [Ziyaret Tarihi: 24 Nisan 2018].

Matwin, S., & Mielniczuk, J., 2016, *Challenges in Computational Statistics and Data Mining* (Vol. 605), Springer International Publishing, Cham, Switzerland, ISBN 978-3-319-18780-8.

Mitchell, T. M., 2006, The discipline of machine learning (Vol. 9), *Carnegie Mellon University, School of Computer Science*, Machine Learning Department.

Nassif, N. A. A., 2018, Predicting the seam efficiency of sewn blended fabrics using ANN and linear regression models,*International Design Journal*, 8(1), 123-129.

Ni, Y., & Fan, F., 2011, A two-stage dynamic sales forecasting model for the fashion retail, *Expert Systems with Applications*, 38(3), 1529–1536.

Nishimura, H., Kobayashi, M., Maruyama, M., & Nakano, Y., 1999, Off-line character recognition using HMM by multiple directional feature extraction and voting with bagging algorithm, *ICDAR'99, Proceedings of the Fifth International Conference*, 22-22 Sept. 1999, Bangalore, India, pp. 49-52.

Olivas, E., Guerrero, J., Sober, M., Benedito, J., & Lopez, A., 2010, *Handbook of research on machine learning applications and trends: Algorithms. Methods and Techniques*, IGI Global, USA, ISBN 978-1-60566-766-9.

Paliwal, M., & Kumar, U. A., 2009, Neural networks and statistical techniques: A review of

applications, *Expert Systems with Applications*, 36(1), 2–17.

Pang, J., Huang, Q., Yin, B., Qin, L., & Wang, D., 2013, Stochastic boosting for large-scale image classification, *In Image Processing (ICIP), 20th IEEE International Conference on*, (pp. 3274-3277),2013 September, IEEE.

Petres, C., Pailhas, Y., Patron, P., Petillot, Y., Evans, J., & Lane, D., 2007, Path planning for autonomous underwater vehicles, *IEEE Transactions on Robotics*, 23(2), 331-341.

Piramuthu, S., 2004, Evaluating feature selection methods for learning in data mining applications, *European journal of operational research*, 156(2), 483-494.

Purver, M., 2014, A simple baseline for discriminating similar languages,*Association for Computational Linguistics*.

Rani, K. U., 2011, Analysis of heart diseases dataset using neural network approach,*International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 1(5).

Remzi, M., & Djavan, B., 2004, Artificial neural networks in Urology 2004,*European Urology Supplements*, 3(3), 33-38.

Rokach, L., & Maimon, O., 2015, *Data mining with decision trees: theory and applications (2nd Edition)*, World scientific, Singapore, ISBN:978-9814590075.

Sarıman, G., 2011, Veri madenciliğinde kümeleme teknikleri üzerine bir çalışma: k-means ve k-medoids kümeleme algoritmalarının karşılaştırılması, Süleyman *Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 15(3).

Sasaki, Y., 2007, The truth of the F-measure, *Teach Tutor mater*, 1(5).

Sawhney, M. S., & Eliashberg, J., 1996, A parsimonious model for forecasting gross box-office revenues of motion pictures, *Marketing Science*, 15(2), 113-131.

Schneider, M. J., & Gupta, S., 2016, Forecasting sales of new and existing products using consumer reviews: A random projections approach, *International Journal of Forecasting*, 32(2), 243–256.

Sharda, R., & Delen, D., 2006, Predicting box-office success of motion pictures with neural networks, *Expert Systems with Applications*.

Sokolova, M., & Lapalme, G., 2009, A systematic analysis of performance measures for classification tasks, *Information Processing & Management*, 45(4), 427-437.

Srimani, P. K., & Koti, M. S., 2013, Medical diagnosis using ensemble classifiers-a novel machine-learning approach, *Journal of Advanced Computing*, 1, 9-27.

Sun, Z.-L., Choi, T.-M., Au, K.-F., & Yu, Y., 2008, Sales forecasting using extreme learning machine with applications in fashion retailing, *Decision Support Systems*, 46(1), 411–419.

Tan, A. C., Gilbert, D., & Deville, Y., 2003, Multi-class protein fold classification using a new ensemble machine learning approach, *Genome Informatics*, 14, 206-217.

Tanaka, K., 2010, A sales forecasting model for new-released and nonlinear sales trend products, *Expert Systems with Applications*.

Tang, Y., 1996, Active control of SDF systems using artificial neural networks,*Computers & structures*, 60(5), 695-703.

Thomassey, S., & Fiordaliso, A., 2006, A hybrid sales forecasting system based on clustering and decision trees, *Decision Support Systems*, 42(1), 408–421.

Tolun, S., 2015, *Destek vektör makineleri: banka başarısızlığının tahmini üzerine bir uygulama*, Doktora Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü.

Torgo, L., 2011,*Data mining with R: learning with case studies*, CRC press, Boca Raton, ISBN: 978-1-4398-1018-7.

Wenhai, W., &Jiandong, D., 2014, October, Short-term wind power forecasting based on Maximum Correntropy Criterion, *In Power System Technology (POWERCON)*, pp. 2800-2805, IEEE.

Westerhuis, J. A., Hoefsloot, H. C., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J., ... & van Dorsten, F. A., 2008, Assessment of PLSDA cross validation, *Metabolomics*, 4(1),

81-89.

Witten, I. H., Frank, E., & Hall, M. A., 2011, *Data Mining: Practical machine learning tools and techniques-3rd ed.*, Morgan Kaufmann, USA, ISBN 978-0-12-374856-0.

Wong, W. K., & 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(2), 614–624.

Wu, Q., & Law, R., 2011, The forecasting model based on fuzzy novel ν-support vector machine, *Expert Systems with Applications*, 38(10), 12028–12034.

Wu, Q., 2009, The forecasting model based on wavelet ν-support vector machine, *Expert Systems with Applications*, 36(4), 7604–7610.

Wu, Q., 2010, Regression application based on fuzzy ν-support vector machine in symmetric triangular fuzzy space, *Expert Systems with Applications*, 37(4), 2808–2814.

Wu, X., & Kumar, V., 2009, *The top ten algorithms in data mining*, CRC Press, Boca Raton, ISBN-13: 978-1-4200-8964-6.

Yu, X., Qi, Z., & Zhao, Y., 2013, Support vector regression for newspaper/magazine sales forecasting, *Procedia Computer Science*, 17, 1055–1062.

Yu, Y., Choi, T. M., & Hui, C. L., 2011, An intelligent fast sales forecasting model for fashion products, *Expert Systems with Applications*, 38(6), 7373–7379.

Yu, Y., Hui, C. L., & Choi, T. M., 2012, An empirical study of intelligent expert systems on forecasting of fashion color trend, *Expert Systems with Applications*, 39(4), 4383–4389.

Zhang, L., Luo, J., & Yang, S., 2009, Forecasting box office revenue of movies with BP neural network, *Expert Systems with Applications*, 36(3 PART 2), 6580–6587.

EKLER

EK 1. OMDb API'si ile IMDb'nin İnternet Sitesinden İstenen Verinin Çekilmesi için C# Kodu

```
using System;
usingSystem.Collections.Generic;
usingSystem.ComponentModel;
usingSystem.Data;
usingSystem.Drawing;
usingSystem.Linq;
usingSystem.Text;
usingSystem.Windows.Forms;
usingSystem.IO;
usingSystem.Xml;

namespace WindowsFormsApplication1
{
publicpartialclassForm1 : Form
{
string[] tut = newstring[3];
string movietitle;
string year;
string rated;
string released;
string runtime;
string genre;
string director;
string writer;
string actors;
string language;
string country;
string metascore;
string imdbRating;
int i;

public Form1()
{
            InitializeComponent();
}

privatevoid Form1_Load(object sender, EventArgs e)
{

}

privatevoid button1_Click(object sender, EventArgs e)
{
string dosya_yolu = @"C:\Users\Public\Documents\dosya.txt";
FileStream fs = newFileStream(dosya_yolu, FileMode.Open, FileAccess.Read);
StreamReader sw = newStreamReader(fs);
string[] lines = System.IO.File.ReadAllLines(@"C:\Users\Public\Documents\dosya.txt");
```

```
string dosya_yoluu = @"C:\Users\Public\Documents\yaz.txt";
FileStream fss = newFileStream(dosya_yoluu, FileMode.OpenOrCreate, FileAccess.Write);
StreamWriter sww = newStreamWriter(fss);

for (i = 0; i <=1380; i++)
{

string    url    =    String.Format("http://www.omdbapi.com/?t={0}&y=&plot=full&r=xml",
lines[i]);
System.Threading.Thread.Sleep(2000);
XmlDocument xml = newXmlDocument();
xml.Load(url);

if (xml.SelectNodes(@"root/movie").Count <= 0)
{
goto atla;
}

foreach (XmlElement item inxml.SelectNodes(@"root/movie"))
{
                    movietitle = item.GetAttribute("title");
                    year = item.GetAttribute("year");
                    rated = item.GetAttribute("rated");
                    released = item.GetAttribute("released");
                    runtime = item.GetAttribute("runtime");
                    genre = item.GetAttribute("genre");
                    director = item.GetAttribute("director");
                    writer = item.GetAttribute("writer");
                    actors = item.GetAttribute("actors");
                    language = item.GetAttribute("language");
                    country = item.GetAttribute("country");
                    metascore = item.GetAttribute("metascore");
                    imdbRating = item.GetAttribute("imdbRating");
}


sww.WriteLine("{0};{1};{2};{3};{4};{5};{6};{7};{8};{9};{10};{11};{12}",     movietitle,
year, rated, released, runtime, genre, director, writer, actors, language, country,
metascore, imdbRating);
sww.Flush();
atla: ;

}
sww.Close();
fss.Close();

sw.Close();
fs.Close();
}

}
}
```

EK 2. Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Yapay Sinir Ağları Modeli Kodları

```
function [Dogruluk,CMatris]=NNoptsiz_CV(data)

inputs= data(:,1:41);
targets= data(:,42);

k=10;
    cvFolds = crossvalind('Kfold', targets, k);
    net=newff(inputs',targets');
for i = 1:k
        testIdx = (cvFolds == i);
        trainIdx = ~testIdx

        net=train(net,inputs(trainIdx,:)',targets(trainIdx,:)')
        net.trainParam

end
     outputs = round((net(inputs')));
     errors = gsubtract(targets',outputs);
     performance= perform(net,targets',outputs);
     [C, order]=confusionmat(outputs,targets');
     Dogruluk= sum(diag(C))/sum(sum(C));
     CMatris=C;

view(net)
end
```

EK 3.Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Destek Vektör Makinaları Modeli Kodları

```
function [CMatris,Dogruluk]=SVMoptsiz_CV(meas,groups)
X=meas;
Y=groups;

k=10;
indices = crossvalind('Kfold',Y,k);
cp = classperf(Y);

for i = 1:k
test = (indices == i);
    train = ~test;

t = templateSVM();
    Mdl = fitcecoc(X(train,:), Y(train),'Learners',t);
    pred = predict(Mdl, meas(test,:));

    classperf(cp,pred,test);
end

CMatris=cp.CountingMatrix;
Dogruluk=cp.CorrectRate;
```

EK 4.Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Torbalama Modeli Kodları

```
function [Dogruluk,CMatris]=Bagoptsiz_CV(meas,groups)
X=meas;
Y=groups;

k=10;
indices = crossvalind('Kfold',Y,k);
cp = classperf(Y);

for i = 1:k
test = (indices == i); train = ~test;

    Mdl = fitcensemble(X(train,:), Y(train),'Method','Bag');
    pred = predict(Mdl, meas(test,:));

  cp = classperf(cp, pred, test);
end

CMatris=cp.CountingMatrix;
Dogruluk=cp.CorrectRate;
```

EK 5. Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Yapay Sinir Ağları Modeli En İyi Parametre Arama Kodları

```
function [Dogruluk_opt,CMatris,nopt,lrateopt]=NNoptli_CV(data)

inputs= data(:,1:41);
targets= data(:,42);

Dogruluk_opt=0;

for n=1:25
for lrate=0.05:0.05:0.95
k=10;
        cvFolds = crossvalind('Kfold', targets, k);

        net=newff(inputs',targets',n);

for i = 1:k
            testIdx = (cvFolds == i);
            trainIdx = ~testIdx

            net.trainParam.lr =lrate;
            net.trainParam.epochs = 1200;
            net=train(net,inputs(trainIdx,:)',targets(trainIdx,:)');
end
        outputs = round((net(inputs')));
        errors = gsubtract(targets',outputs);
        performance= perform(net,targets',outputs);
        [C, order]=confusionmat(outputs,targets');
        Dogruluk= sum(diag(C))/sum(sum(C));

if Dogruluk>Dogruluk_opt;
           Dogruluk_opt=Dogruluk;
           CMatris=C;
           nopt=n;
           lrateopt=lrate;
end
end
end
view(net)
```

EK 6. Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Destek Vektör Makinaları Modeli En İyi Parametre Arama Kodları

```
function [CMatris,Dogruluk,ktut,ctut,boxtut]=SVMoptli_CV(meas,groups)
X=meas;
Y=groups;
Dogruluk=0;
kernel={'linear','polynomial','gaussian','rbf'};

k=10;
indices = crossvalind('Kfold',Y,k);
cp = classperf(Y);

for ks=1:4
for cs=1:10
for box=1:12
for i = 1:k
test = (indices == i);
                train = ~test;

                kerneltext = sprintf('%s',kernel{ks});
                bc=0.0001*(10^cs);
                K=length(unique(Y(train)));
                C=(bc.*ones(K))-(bc.*eye(K));
                boxc=1e-7*(10^box);

t = templateSVM('BoxConstraint',boxc,'KernelFunction',kerneltext);
                Mdl = fitcecoc(X(train,:), Y(train),'Learners',t,'Cost',C);
                pred = predict(Mdl, meas(test,:));

                classperf(cp,pred,test);
end

            Deger=cp.CorrectRate

if Deger>Dogruluk
                Dogruluk=Deger;
                CMatris=cp.CountingMatrix;
                ktut = sprintf('%s',kernel{ks});
                ctut=bc;
                boxtut=boxc;
end
            cp = classperf(Y);
end
end
end
```

EK 7. Veri Setinin 10-Kat Çapraz Geçerleme ile Oluşturulduğu Torbalama Modeli En İyi Parametre Arama Kodları

```
function [Dogruluk,CMatris,Ntut,ctut]=Bagoptli_CV(meas,groups)
X=meas;
Y=groups;
Dogruluk=0;

k=10;
indices = crossvalind('Kfold',Y,k);
cp = classperf(Y);

for Num=1:10
for cs=1:10
for i = 1:k
            bNum=Num*100;
test = (indices == i); train = ~test;
            bc=0.0001*(10^cs);
            K=length(unique(Y(train)));
            C=(bc.*ones(K))-(bc.*eye(K));

t=templateTree();
            Mdl = fitcensemble(X(train,:),
Y(train),'Method','Bag','Learners',t,'NumLearningCycles',bNum,'Cost',C);
            pred = predict(Mdl, meas(test,:));

            cp = classperf(cp, pred, test);
end
        Deger=cp.CorrectRate
if Deger>Dogruluk
            CMatris=cp.CountingMatrix;
            Dogruluk=Deger;
            Ntut = bNum;
            ctut=bc;
end
        cp = classperf(groups);
end
end
```

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Özge Hüsniye NAMLI
Doğum Yeri	Bakırköy/ İstanbul
Doğum Tarihi	11.06.1992
Uyruğu	☑T.C.❒ Diğer:
Telefon	+90 538 377 47 67
E-Posta Adresi	ozge.namli@hotmail.com
Web Adresi	http://www.tau.edu.tr/profil/ozge_husniye_namli/



Eğitim Bilgileri	
Lisans	
Üniversite	Celal Bayar Üniversitesi
Fakülte	Mühendislik Fakültesi
Bölümü	Endüstri Mühendisliği
Mezuniyet Yılı	09.07.2015

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

Makale ve Bildiriler

Namlı Ö. H., Özcan T., "Makine Ögrenmesi Algoritmaları Kullanarak Gişe Hasılatının Tahmini", *Yönetim Bilisim Sistemleri Dergisi*, 3(2), 130-143, 2017.

Erdoğan Z., Namlı Ö. H., Akarsu C. H., "Öğrenci Başarısının Bileşik Makine Öğrenme Teknikleri Kullanılarak Tahmini", *International Symposium On Industry 4.0 and Applications*,Karabük, Türkiye, 12-14 Ekim 2017, pp. 150-155.

Namlı Ö. H., "Karar Agaçları Algorıtmalarının Sınıflandırmadaki Başarısı ve Bankacılık Tele-Pazarlama Sektöründe Uygulanması". *Uluslararası Katılımlı 16. Üretim Arastırmaları Sempozyumu "4.Sanayi Devriminde Üretim"*, İstanbul, Türkiye,12-14 Ekim 2016, pp. 845-850.

Akarsu C. H., Erdogan Z., Namlı Ö. H., "Türkıye'de Gsm Operatötü Seçiminde Bulanık Karar Verme Yaklaşımı Olarak AHP Bütünlesık TOPSIS Yöntemi" *Uluslararası Katılımlı 16. Üretim Arastırmaları Sempozyumu "4.Sanayi Devriminde Üretim"*, İstanbul, Türkiye,12-14 Ekim 2016, pp. 27-32.