T.C. İSTANBUL KÜLTÜR ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ TEKNİKLERİNİN BÜTÇE VERİMLİLİĞİNE UYGULANMASI ÜZERİNE BİR ÇALIŞMA

DOKTORA TEZİ

Göksel Kıvanç DEMİREL

Anabilim Dalı: İŞLETME

Programı: İŞLETME

Tez Danışmanı: Prof. Dr. Ali ŞEN

İSTANBUL-2021

T.C. İSTANBUL KÜLTÜR ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ TEKNİKLERİNİN BÜTÇE VERİMLİLİĞİNE UYGULANMASI ÜZERİNE BİR ÇALIŞMA

DOKTORA TEZİ

Göksel Kıvanç DEMİREL

Anabilim Dalı: İŞLETME

Programı: İŞLETME

Tez Danışmanı: Prof. Dr. Ali ŞEN

Jüri: Prof. Dr. Uğur YOZGAT

Dr. Öğr. Üyesi Hatice Nazan ÇAĞLAR

Dr. Öğr. Üyesi Murat Taha BİLİŞİK

Dr. Öğr. Üyesi Özge Nalan BİLİŞİK

İSTANBUL-2021

ÖNSÖZ

Bu çalışmam boyunca bilgi birikimi ve tecrübesiyle hiçbir desteğini, ilgisini esirgemeyip beni cesaretlendiren ve yön gösteren, değerli akademisyen ve tez danışmanım Sayın Prof. Dr. Ali Şen hocama,

19 yıl önce İstanbul Üniversitesi'nde tanışma şerefine eriştiğim, bana olan inancını hep hissettiren değerli akademisyen Prof. Dr. Mahmut Paksoy hocama,

Tez izleme jürimde de yer alan ve Doktora öğrenimim boyunca bana eşsiz katkılarda bulunmuş olan Sayın Dr. Öğretim Üyesi Murat Taha Bilişik beye,

Akademik kariyerim konusunda beni daima cesaretlendiren ve desteklerini esirgemeyen Dr. Cumali Ergün beye,

Bu çalışma ve iş hayatımın her evresinde kendilerinden ödünç aldığım zaman konusunda gösterdikleri destek ve sonsuz hoşgörüden dolayı değerli eşim Ayşegül ve sevgili çocuklarımız Defne ve Mustafa'ya en derin şükranlarımı sunuyorum.

Bu çalışmamı evladı olmaktan gurur duyduğum, Rahmetli babam Mustafa Demirel' e ithaf ediyorum.

Göksel Kıvanç DEMİREL

İÇİNDEKİLER

TABLO LİSTESİ

ŞEKİL LİSTESİ

KISALTMALAR

AI : Yapay Zeka

AUC : Area Under Curve

BP : Backpropagation

CART : Classification and Regression Trees

CHAID : Chi-Squared Automatic Interaction Detector

CPU : Central Process Unit

DID : Dual Information Distance

DN : Doğru Negatif

DP : Doğru Pozitif

GBM : Gradient Boosted Machine

HHS : Households

IBM : International Business Machines

KNIME : Konstanz Information Miner

MARS : Multivariate Adaptive Regression Splines

MİB : Metropolitan İstatistiksel Alanı

ML : Makine Öğrenmesi

MLP : Multilayer Perceptron

NP : Nondeterministic Polynomial

QUEST : Quick, Unbiased, Efficient Statistical Tree

RO : Rastgele Orman

ROC : Receiver Operating Characteristics

SKU : Stock Keeping Unit

SLIQ : Supervised Learning in Quest

SPRINT **:** Scalable Parallelizable Induction of Decision Trees

SPSS **:** Statistical Package for the Social Sciences

SQL **:** Structured Query Language

TPR **:** Temporary Price Production

UPC **:** Unit Product Code

WEKA **:** Waikato Environment for Knowledge Analysis

YGFİ **:** Yalnızca Geçici Fiyat İndirimi

YN **:** Yanlış Negatif

YP **:** Yanlış Pozitif

YSA **:** Yapay Sinir Ağı

Enstitüsü : Lisansüstü Eğitim
Anabilim Dalı : İşletme
Programı : İşletme Doktora
Tez Danışmanı : Prof. Dr. Ali ŞEN
Tez Türü ve Tarihi : Doktora - Haziran 2021

ÖZET

Makine Öğrenmesi Tekniklerinin Bütçe Verimliliğine Uygulanması Üzerine Bir Çalışma

Göksel Kıvanç DEMİREL

Bu tez çalışmasının amacını satış ve promosyon bilgilerinden elde edilen büyük bir veri seti üzerinde daha az sayıdaki özellik sayısının kullanılarak pazarlamanın etkin olabileceği müşteri kitlesinin çeşitli makine öğrenmesi algoritmaları aracılığıyla seçilerek skorlanması oluşturmuştur. Bu amaç doğrultusunda, geçmiş yıllardan beri toplanılan müşteri bazlı pazarlama verileri üzerinden daha önce pazarlama bütçesi için ayrılmış ve başarılı olunmuş kitle işaretlenerek, deney veri seti farklı makine öğrenmesi algoritmaları KNIME 4.2.1 programında analiz edilmiştir. Makine öğrenmesi algoritmalarının matematiksel ve veri işleme farklılıkları dikkate alınarak, veri seti üzerinde karar ağacı, rastgele orman, çok katmanlı algılayıcı ve gradyan artırma algoritmaları kullanılarak performans analizleri gerçekleştirilmiştir. Bu algoritmaların sonuçları doğruluk, duyarlılık, kesinlik, seçicilik ve F-Ölçütü oranları belirlenerek birbirleriyle karşılaştırılmıştır. Pazarlamanın etkin olabileceği müşteri kitlesinin belirlenmesi için kullanılan dört algoritmadan elde edilen doğruluk değerleri birbirlerine yakın olduğundan bu algoritmaların herbirinin kullanışlı olduğu söylenebilir. Ancak en iyi performansı sergileyen algoritmanın gradyan artırma makineleri olduğu görülmüş ve bu algoritmanın pazarlamanın etkin olabileceği kitlenin tespit edilmesinde kullanımı tavsiye edilmiştir. Diğer taraftan, çalışmada özellik sayısının azaltıldığı ve karmaşıklık durumlarının en aza indirgendiği modellerin performans ölçümünden yüksek doğruluk değerleri elde edilmiştir. Bu bağlamda, yapılan çalışma en önemli ve doğru seçilmiş özelliklerden oluşturulan bir modele dayalı analizlerin yapılabileceğini göstermesi

açısından değerlidir. Gerçek dünya verilerinin kullanıldığı bu çalışmada, izlenilen yöntemin ve elde edilen bulguların veri bilimine katkı sağlaması ve rehberlik etmesi beklenmektedir. Bu çalışmanın altyapı ve kapsamının geliştirilmesi, şirketlerin finansal öngörüler sağlayabildiği bütçe destek sistemlerinin oluşturulabilmesine olanak sağlayacaktır.

Anahtar Kelimeler: Makine Öğrenmesi, Özellik Mühendisliği, Karar Ağacı, Rastgele Orman, Çok Katmanlı Algılayıcı, Gradyan Artırma Makineleri, Pazarlama Bütçesi.

University : Istanbul Kültür University
Institute : Institute of Graduate Education
Department : Business
Programme : Business, PhD
Supervisor : Prof. Dr. Ali ŞEN
Degree Awarded and Date : Ph.D. - June 2021

ABSTRACT

A Study on the Application of Machine Learning Techniques to Budget Efficiency

Göksel Kıvanç DEMİREL

The aim of this thesis study is to scoring the customers who can be effective in marketing by using various machine learning algorithms by using a smaller number of features on a big data set obtained from sales and promotion information. For this purpose, the customer group that has been reserved for the marketing budget and has been successful is marked on the customer-based marketing data collected for the past years and the data set was analyzed by different machine learning algorithms in KNIME 4.2.1 program. Considering the mathematical and data processing differences of machine learning algorithms, performance analyzes were performed on the data set using decision tree, random forest, multilayer perceptron and gradient boosted algorithms. The results of these algorithms were compared with each other by determining the accuracy, sensitivity, precision, specifity and F-measure ratios. It can be said that each of these algorithms is useful, since the accuracy values obtained from the four algorithms used to determine the customer base where marketing can be effective are close to each other. However, it is seen that the algorithm with the best performance is gradient boosted machines and it is recommended to use this algorithm to determine the customers where marketing can be effective. On the other hand, in terms of applied analysis, after the variable significance analysis, high accuracy values were obtained in model performance measurement by reducing the number of features and minimizing the complexity. On the other hand, high accuracy values were obtained from the performance measurement of the models in which the number of features in the study was reduced and the complexity situations were minimized. In

this context, the study is valuable in that it shows that analyzes can be made based on a model created from the most important and correctly selected features.It is expected that the method followed and the findings obtained in this study, in which real world data is used, will contribute and guide data science. As a result of the development of the infrastructure and scope of this work, budget support systems will be created in which companies can provide financial predictions. Improving the infrastructure and scope of this work will allow companies to create budget support systems that can provide financial forecasts.

Keywords: Machine Learning, Feature Engineering, Decision Tree, Random Forest, Multilayer Perceptron, Gradient Boosted Machines, Marketing Budget.

1. GİRİŞ

Daha iyi hesaplama gücü ve daha fazla depolama kaynağı ile zenginleştirilmiş bir "veri çağında" yaşıyoruz. Bu verilerin her geçen gün artış göstermesi tüm verilerin anlamlandırılmasındaki zorlukları da beraberinde getirmektedir. İşletmeler; veri bilimi, veri madenciliği ve makine öğrenmesinden gelen kavram ve metodolojilerden faydalanarak oluşturdukları akıllı sistemler ile verileri anlamlandırmaya çalışmaktadır. Makine öğrenmesi, verilere anlam kazandıran algoritma uygulaması ve bilimi olarak bilgisayar biliminin en heyecan verici alanlarından biridir.

Makine öğrenmesi (ML), bilgisayar sistemlerinin, insanların yaptığı gibi verilere anlam verebildiği bilgisayar bilimi alanıdır. Basit bir deyişle, makine öğrenmesi, bir algoritma veya yöntem kullanarak ham verilerden modelleri çıkaran bir yapay zeka (AI) türüdür. Makine öğrenmesinin ana odak noktası, bilgisayar sistemlerinin açıkça programlanmadan veya insan müdahalesi olmadan deneyimlerden öğrenmesine olanak sağlamasıdır. Makine öğrenmesi, deneyim yoluyla otomatik olarak gelişen bilgisayarların nasıl oluşturulacağı sorusunu ele alır. Bilgisayar bilimi ile istatistiğin kesişme noktasında ve yapay zeka ile veri biliminin merkezinde yer alan, günümüzün en hızlı büyüyen teknik alanlarından biridir. Makine öğrenmesindeki son gelişmeler, hem yeni öğrenme algoritmalarının ve teorilerinin geliştirilmesi hem de çevrimiçi verilerin kullanılabilirliğindeki ve düşük maliyetli hesaplamalardaki artışın sürdürülmesinde kullanılır. Veri yoğun makine öğrenmesi yöntemlerinin benimsenmesi bilim, teknoloji ve ticarette kullanılabilir. Bu da sağlık hizmetleri, üretim, eğitim, finansal modelleme ve pazarlama gibi yaşamın birçok alanında daha fazla veriye dayalı karar vermeye yol açar.

Makine öğrenmesi, birbiriyle ilişkili iki soruya odaklanan bir disiplindir. Bu sorulardan ilki, deneyim yoluyla otomatik olarak gelişen bilgisayar sistemleri nasıl inşa edilebildiği; ikincisi, bilgisayarlar, insanlar ve organizasyonlar dahil tüm öğrenme sistemlerini yöneten temel istatistiksel hesaplama-bilgi-teorik yasaların neler olduğudur. Makine öğrenmesi çalışmaları, hem bu temel bilimsel ve mühendislik

sorularını ele almak hem de ürettiği ve birçok uygulamada kullandığı son derece uygulamalı bilgisayar yazılımı için önemlidir.

Makine öğrenmesi, laboratuvar çalışmalarından yaygın ticari kullanımdaki uygulamalı teknolojilere kadar son yirmi yılda dikkat çekici bir şekilde ilerleme kaydetmiştir. Yapay zeka içerisinde makine öğrenmesi; bilgisayar görüşü, konuşma dilini tanıma, doğal dil işleme, robot kontrolü ve diğer uygulamalarda yazılım geliştirmek için tercih edilen bir yöntem olarak ortaya çıkmıştır. Yapay zeka sistemleri geliştiricilerinin çoğu, birçok uygulamada bir sistemin tüm olası girdileri için istenen yanıtı önceden tahmin ederek manuel olarak programlamaktan çok, istenen girdi-çıktı davranışının örneklerini göstererek eğitmenin çok daha kolay olabileceğini kabul etmektedir. Makine öğrenmesinin etkisi, bilgisayar biliminde ve tüketici hizmetleri, karmaşık sistemlerdeki arızaların teşhisi ve lojistik zincirlerinin kontrolü gibi veri yoğun konularla ilgili birçok endüstriyel alanda hissedilmektedir. Deneysel verileri yeni yollarla analiz etmek için makine öğrenmesi yöntemleri geliştirildiğinden, deneysel bilimlerde biyolojiden kozmolojiye ve sosyal bilimlere kadar benzer şekilde geniş bir etki yelpazesi bulunmaktadır.

Bir öğrenme problemi, bir çeşit eğitim deneyimi yoluyla bir görevi yerine getirirken bazı performans ölçütlerini iyileştirme sorunu olarak tanımlanabilir. Farklı makine öğrenmesi problemlerinde sergilenen çok çeşitli veri ve problem türlerini kapsayacak şekilde çok çeşitli makine öğrenmesi algoritmaları geliştirilmiştir (Hastie vd., 2009; Murphy, 2012). Kavramsal olarak, makine öğrenmesi algoritmaları performans ölçütlerini optimize eden bir program bulmak için eğitim deneyiminin rehberliğinde geniş bir aday program alanında arama yapıyor olarak görülebilir. Makine öğrenmesi algoritmaları, kısmen aday programları temsil etme şekillerine ve kısmen de bu programlar alanında arama yapma şekillerine göre büyük farklılıklar gösterir. Pek çok algoritma, görevin bir işlevde somutlaştırıldığı işlev yaklaştırma problemlerine odaklanır ve öğrenme problemi, işlevin bilinen girdi-çıktı çiftlerinden oluşan bir örneklemden oluşan deneyimle bu işlevin doğruluğunu iyileştirmeyi amaçlamaktadır. Bazı durumlarda, işlev açıkça parametreleştirilmiş bir işlevsel form olarak temsil edilir; diğer durumlarda, işlev kapalıdır ve bir arama süreci, bir çarpanlara ayırma, bir optimizasyon prosedürü veya simülasyon tabanlı bir prosedür yoluyla elde edilir. İşlev kapalı olduğunda bile genellikle parametrelere veya diğer

ayarlanabilir serbestlik derecelerine bağlıdır ve eğitim, performans ölçüsünü optimize eden bu parametreler için değerler bulmaya karşılık gelir.

Öğrenme algoritması ne olursa olsun, temel bir bilimsel veya uygulamalı problemin amacı, belirli öğrenme algoritmalarının kapasitesini ve herhangi bir öğrenme probleminin doğasında olan zorluğu karakterize etmektir. Bu bağlamda, öğrenme algoritmasının belirli bir türden ve eğitim verilerinden ne kadar doğru bir şekilde öğrenebildiği, öğrenme algoritmasının modelleme varsayımlarındaki hatalara veya eğitim verilerindeki hatalara karşı ne kadar sağlam olduğu, verilen bir eğitim verisi hacmine sahip bir öğrenme problemi için başarılı bir algoritma tasarlamanın mümkün olup olmadığı göz önüne alınması gereken önemli durumlardır.

Bir çalışma alanı olarak makine öğrenmesi, bilgisayar bilimi, istatistik ve zaman içinde otomatik iyileştirme ve belirsizlik altında çıkarım ve karar verme ile ilgili çeşitli diğer disiplinlerin dönüm noktasında yer alır. İlgili disiplinler arasında insan öğreniminin psikolojik çalışması, evrim çalışması, uyarlanabilir kontrol teorisi, eğitim uygulamalarının incelenmesi, sinirbilim, örgütsel davranış ve ekonomi bulunmaktadır.

Geçtiğimiz on yılda, ağa bağlı ve mobil bilgi işlem sistemlerinin büyük miktarda veriyi toplama ve taşıma becerisinde hızlı bir büyüme görüldü. Bu tür verileri toplayan bilim adamları ve mühendisler, bu tür veri kümelerinden yararlı içgörüler, tahminler ve kararlar elde etme sorununa çözümler için makine öğrenmesine yönelmişlerdir. Aslında, verilerin tüm boyutu, hesaplama ve istatistiksel konuları harmanlayan ölçeklenebilir prosedürler geliştirmeyi gerekli kılmaktadır. Ancak sorun, modern veri setlerinin yalnızca boyutundan daha fazlasıdır. Bu verilerin çoğu tanecikli ve kişiselleştirilmiş bir doğaya sahiptir. Mobil cihazlar ve gömülü veri işleme, insanların kişilikleri hakkında büyük miktarda verinin toplanmasına izin verir ve makine öğrenmesi algoritmaları, hizmetlerini her bireyin ihtiyaçlarına ve koşullarına göre özelleştirmek için bu verilerden öğrenebilir. Dahası, bu kişiselleştirilmiş hizmetler birbirine bağlanabilir, böylece birçok kişiden gelen verilerin zenginliğinden ve çeşitliliğinden yararlanırken yine de her birinin ihtiyaçlarına ve koşullarına göre özelleştirilen genel bir hizmet ortaya çıkar. Hizmetleri ve üretkenliği iyileştirmek için büyük miktarda veriyi yakalama ve madenciliği yapma yönündeki bu eğilimin örnekleri ticaret, bilim ve hükümetin birçok alanında bulunabilir. Tarihsel tıbbi kayıtlar, hangi hastaların hangi tedavilere en iyi yanıt vereceğini keşfetmede, geçmiş

trafik verileri trafik kontrolünü iyileştirmede ve tıkanıklığı azaltmada, tarihsel suç verileri yerel polisi belirli zamanlarda belirli yerlere tahsis etmeye yardımcı olmada kullanılırken; biyoloji, astronomi, sinirbilim ve diğer veri yoğun deneysel bilimlerdeki ilerlemeyi hızlandırmak için büyük deneysel veri setleri yakalanır ve düzenlenir.

Makine öğrenmesi, tahmin modellerini potansiyel olarak iyileştirebilecek ve yönetim kararlarının alınmasına yardımcı olabilecek yenilikçi bir yöntemdir. Veri tabanlarından doğru tahmin modelleri oluşturmaya dayanan doğrudan pazarlama, bu tür uygulamalardan yararlanabilecek alanlardan birisidir. Çoğu işletme doğrudan pazarlamayı bir dağıtım stratejisi olarak benimsediğinden, bu kanaldaki harcamalar son yıllarda artmış ve bu da doğrudan pazarlamacılar için satışları artırmak, maliyetleri düşürmek ve kârlılığı artırmak için tüketici yanıt modelini en önemli öncelik haline getirmiştir (Cui vd., 2006). Araştırmacılar, tüketici satın alımlarını tahmin etmeye yönelik geleneksel yaklaşımlara ilaveten yakın zamanda, büyük veri setleri için birçok farklı avantajı olan makine öğrenmesi yöntemlerini uygulamaktadırlar. Talebi tahmin etmek ve satış etkenlerini anlamak, perakende analitiğindeki en önemli görevlerden biridir. Satışların fiyat duyarlılığının yanı sıra promosyonların gerçek etkililiğinin belirlenmesi, satışları belirleyen ilişkilerin karmaşıklığı nedeniyle zorlu sorunlar olmaya devam etmektedir. Tanıtılmayan ürünlerin tanıtılan ürünler tarafından engellenmesi, bir promosyonun etkinliğinin büyük olasılıkla önceki haftalardaki promosyon faaliyetine bağlı olması, belirli ürünlerin ortak promosyonları diğer bazı ürünlerin satışları üzerinde sinerji etkisi yaratması gibi durumlar ortaya çıkabilir. Sonuç olarak, bir promosyonun yarattığı satış artışının hiçbiri veya sadece bir kısmı perakendeciler için anlamlıdır (Gedenk, 2018). Daha belirgin bir biçimde, tüm promosyonların %50'den fazlasının perakendeciler için kârlı olmadığı gösterilmiştir (Ailawadi vd., 2007). Bir promosyon üretici için faydalı olsa bile, bir perakendeci için olumsuz bir etkiye neden olma olasılığı yüksektir. Çünkü promosyon yapılmamış ürünlerden, tipik olarak daha düşük marjlı promosyonlu ürünlere geçiş gerçekleşir. Promosyonları etkin bir şekilde yönetmek için, promosyonların satışlar üzerindeki etkilerinin değerlendirilmesi ve nicelendirilmesi gerekir. Bu nedenle iyi bir satış yanıt modeline ihtiyaç duyulur.

Literatürde satış tahmini için pazarlamada matematiksel modeller ve talep tahmininin fiyat ve promosyon esnekliği üzerine kullanılan klasik modellerden biri SCAN * PRO modeli ve varyasyonlarında hem sıradan en küçük kareler hem de daha

gelişmiş ekonometrik yöntemler kullanılarak tahmin edilir. Wittink ve arkadaşları (1988) tarafından önerilen modelin varyasyonları literatürde tartışılmış ve karşılaştırılmıştır (Andrews vd., 2008). SCAN * PRO modeli ve uzantıları satışları, bir marka için fiyatın kendi ve çapraz marka etkilerine, özellikli reklam, koridor görüntüleri, hafta etkileri ve mağaza etkilerine ayrıştırır. Bazı rakip modellerde geçerli hafta fiyatları ve promosyonların etkisinin geleneksel regresyon yöntemlerine dayalı olmasına rağmen, model dönemler arası ve kategoriler arası etkileri tipik bir mağaza geniş bir yelpazede çok sınırlı bir ölçüde açıklar. Talep modelleme üzerine yapılan diğer çalışmaların çoğunda da yirmi kadar özellik içeren dar bir SCAN * PRO modeli kullanılmıştır (Haupt vd., 2014). Büyük veri ve makine öğrenmesi algoritmalarının gücü son on yılın bazı satış tahmin belgelerinde gösterilmiş olsa da, özelliklerin sayısı genellikle çok fazla olmamıştır (Sun vd., 2008; Ali vd. 2009; Ferreira vd., 2015; Yang ve Zhang, 2018). Talep tahmini için büyük boyutlu verilerle uğraşmanın önemini vurgulayan birkaç çalışmadan biri, Ma ve arkadaşları tarafından yapılmıştır (Ma vd., 2016). Bu çalışmada, özellik seçimi için dört aşamalı bir yaklaşımın önerildiği çapraz ürün ve dönemler arası etkiler için hesaplamanın önemi vurgulanmıştır. Daha sonra, bu yaklaşımları çok dönemli kâr maksimizasyonu içeren kategoriler için bir optimizasyon modeli oluşturmak üzere bir optimizasyon algoritmasıyla birleştirilmiştir (Ma ve Fildes, 2017). Bu yöntem, ilgili ürün kategorilerini belirlemenin öznel bir aşamasını içermektedir. Bu adım, modelin hesaplama açısından daha verimli olmasına izin verse de, yerine geçme ve tamamlama modellerini büyük ölçüde sınırlayabilir. Yerine geçme ve tamamlama modellerinin aynı veya bir şekilde ilgili kategorilerdeki stok tutma birimleri (SKU-Stock Keeping Unit) için daha güçlü olmasını beklemek mantıklı olsa da, SKU'ların en azından biraz ilgili kategorilerden olması durumunda herhangi bir SKU'nun fiyat ve promosyon özelliklerinin diğer herhangi bir SKU'nun satışını etkileyebileceğini varsayarak daha fazla esnekliğe izin verilebilir. Örneğin, bir kişi çok fazla indirimli dondurma satın alırsa, çikolata satın alma olasılığı daha düşük olabilir. Çünkü o hafta için çok fazla sağlıksız yiyecek aldığını veya dondurma satın aldıktan sonra çok az parasının kaldığını düşünebilir. Bu kısıtlayıcı olmayan varsayım, hesaplama verimliliği açısından olumsuz görünse de, günümüzün büyük veri teknolojileri bu tür varsayımların yapılmasına ve modellerde çok sayıda öngörücünün dahil edilmesine izin vermektedir. Günümüzün büyük veri teknolojilerinin petabaytlarca perakende verisi ile çalışmaya izin verdiğini hesaba katarsak (Bradlow vd., 2017), ilgili potansiyel öngörücülerin sayısı hakkında gittikçe

daha az endişelenebilir. Bu nedenle, yüzlerce ve hatta binlerce özelliği etkili bir şekilde işleyen ve her bir özelliğe rastgele ormanlar (RO) ve gradyan artırma makineleri (GBM) gibi modellemede kullanılma fırsatı veren algoritmalar, bazı özelliklerden kurtulmayı gerektiren yöntemlerden daha umut verici görünmektedir. Potansiyel öngörücülerin sayısını öznel olarak sınırlandırmak çoğu zaman yardımcı olabilirken, Ma ve arkadaşlarının önerdiği yaklaşımın daha önemli bir sınırlaması, fonksiyonel formun esnekliğini esasen doğrusal otoregresif dağıtılmış gecikme modelleri kullanarak kısıtlamasıdır (Ma vd., 2016). Bu nedenle, çok boyutlu verilerle çalışmayı amaçlayan tüm değişikliklere rağmen, Ma ve Fildes'in temel ekonometrik tanımlaması halen çok aşamalı bir özellik seçim prosedürü ile elde edilen geleneksel bir derin öğrenme modelidir (Ma ve Fildes, 2017). Aslında, ekonomistler ve pazarlama bilimcileri parametre tahminlerinin kolay yorumlanmasına izin veren geleneksel ekonometrik modelleri tercih etmeye devam etmektedirler. Aynı zamanda, büyük süpermarket çeşitliliğinin kaçınılmaz olarak veri oluşturma sürecinde giderek artan karmaşık etkileşimlere ve doğrusal olmayanlıklara yol açtığını varsaymak mantıklıdır. Bu, ilgili özelliklerin sayısı çok olduğunda parametrik regresyon modellerinin kullanılmasıyla açıklanamamaktadır. Örneğin, stoklama nedeniyle, çoğu ürün için diyelim ki t-2 dönemindeki bir promosyon, t dönemindeki bir promosyonun etkisini zayıflatacaktır, ancak bu gecikmeli promosyonun etkisi muhtemelen t-3 ve t-1 dönemlerinde promosyonlar ve geçici fiyat indirimleri olup olmadığına göre hafifletilecektir. Geleneksel doğrusal modeller, özellikle sayısı birkaç düzineyi aşan özellikler açısından tüm bu tür etkileşimleri ve doğrusal olmayanlıkları hesaba katacak kadar esnek değildir. Parametrik regresyonlar hala yaygın olarak kullanılsa da, "ekonometri için büyük veri hilelerinin" ekonomi ve iş dünyasındaki birçok sorun için uygulanabilir olduğunu göstermektedir (Bajari vd., 2015; Varian, 2014).

Çalışma kapsamında, bir perakende analitik şirketi olan Dunnhumby tarafından akademik amaçlarla "Frat'ta kahvaltı: Bir zaman serisi analizi" başlığı altında sunulan haftalık satış verileri kullanılmıştır (Dunnhumby, 2019). Bu veri seti, dört seçilmiş kategoride (ağız bakım suyu, çubuk kraker, dondurulmuş pizza ve kahvaltılık gevrek) en çok satış yapan ilk üç markanın en iyi beş ürünü ile ilgili satış ve promosyon bilgilerini içermektedir. Veri seti içerisine dahil edilen stok tutma birimi (SKU) ve ürün kategorilerinin sayısı açısından biraz sınırlı olmasına rağmen, bu veri seti bilimsel açıdan benzersiz özellikler taşımaktadır. Çünkü veri seti önde gelen bir

perakende veri sağlayıcısının gerçek yaşamı yansıtan, yüksek kalitede ve halka açık verilerini içermektedir. Tez çalışması kapsamında yararlanılan veri seti literatürde birkaç çalışmada daha kullanılmıştır (Antipov ve Pokryshevskaya, 2020; Tareq vd., 2020; Li, 2018). Antipov ve Pokryshevskaya'nın yapmış olduğu çalışmada, yorumlanabilir makine öğrenimi yöntemlerini kullanarak, kavramsal modelden çeşitli yordayıcı gruplarının önemi ile ilgili eyleme geçirilebilir içgörüler elde edilmiştir. Aynı zamanda, pazarlama yöneticilerinin özellik ilişkilendirmesi için yaklaşık Shapley değerleri kullanılarak tahminleri bireysel bağlayıcı değişkenlerin etkilerine ayırmanın ne kadar yararlı olabileceği gösterilmiştir. Yapılan çalışmada, satış modellemesi için gradyan artırma makineleri, rastgele orman ve elastik ağlar makine öğrenmesi algoritması olarak kullanılmıştır. Satış tepkisi modellemesi için modelden bağımsız yorumlanabilir makine öğrenmesi tekniklerinin yararlılığı, yorumlanabilir ancak daha az esnek olan geleneksel ekonometrik modellerin ağırlıklı olarak kullanıldığı yerlerde gösterilmiştir. Tareq ve arkadaşlarının yapmış olduğu çalışmada, online pazar için bir dinamik öneri sistemi modeli önerilmiştir. Önerilen teknik; pazar sepeti analizini, sürekli ürün madenciliğini, en çok satan ürünleri ve müşteri kişiselleştirmesini entegre ederek müşterilerin derecelendirme ve geri bildirim sorunlarının üstesinden gelmek için akıllı bir çözüm modeli sağlamaktadır. Li'nin yaptığı kavramsal çalışmada, veri setindeki toplam 55 ürünün haftalık satışları bir araya getirilmiştir. İlk olarak ham satışlar üzerinden bir günlük dönüşüm sağlanmış ve çapraz zamanda otomatik bağıntıyı ortadan kaldırmak için önceki haftanın günlük satışlarında bir özbağlanım (otoregresyon) gerçekleştirilmiştir.

Talebi tahmin etmek ve satış etkenlerini anlamak, pazarlama ve bütçe planlamasının en önemli görevlerden biridir. Geleneksel olarak az sayıda öngörücüye modeller, satış modellemesinde ağırlıklı olarak kullanılmasına rağmen, bütçe verimliliğine veya planlamasına yönelik büyük veriler kullanılarak yapılan çalışma sayısı oldukça kısıtlıdır. Makine öğrenmesi tekniklerinin kullanıldığı büyük boyutlu veriler ile bütçe verimliliğinin önemini vurgulayan literatürde birkaç çalışma bulunmaktadır (Ma vd., 2016; Ma ve Fildes, 2017). Bu nedenle, müşteri bazlı pazarlama ve satış verileri üzerinden birbiriyle ilişkili çok sayıdaki özellik ve değişkenin göz önüne alındığı, pazarlama bütçesinin potansiyel müşterilere harcanabildiği, gerçek dünyadaki talebin tahmin edilebildiği ve satış etkenlerinin anlaşılabildiği gerçekçi bir modele ihtiyaç duyulmaktadır. Bu çalışma, herhangi bir

bütçe planlamasında kullanılabilecek müşteri bazlı pazarlama verilerine dayalı bir pazarlama bütçesi modellemesi için kavramsal bir model önermektedir. Bu bağlamda, büyük verili satış ve promosyon bilgilerinin işlenerek pazarlamanın etkin olabileceği müşteri kitlesinin belirlenmesine yönelik ortaya konulan kavramsal model literatürde yapılmış olan ilk çalışma özelliği taşımanın yanı sıra, çalışma kapsamında uygulanan adımlar bütçe planlaması ve verimliliği için örnek oluşturmaktadır.

Bir veri setinin daha fazla sayıda olan gözlem sınıfına çoğunluk sınıfı, daha fazla sayıda olan gözlem sınıfına azınlık sınıfı adı verilmektedir (Chawla vd., 2004). Genel olarak azınlık sınıfının çalışma alanları daha çok öneme sahip olmakla birlikte, azınlık sınıflarının doğru olarak yüksek oranlarda tahmin edilmesi beklenmektedir (Liu vd., 2009). Makine öğrenmesi algoritmalarının doğruluğunun ve güvenilirliğinin belirlenmesinde ele alınan deneysel verilerin eğitim ve test veri seti şeklinde ikiye ayrılması son derece önemlidir. Çalışmada kullanılan veri setinde yer alan 538.643 adet pazarlama verisinin %70'i (377.050 adet) modelin eğitiminde, geri kalan %30'u (161.593 adet) ise modelin test edilmesinde kullanılmıştır. Bu çalışmada, pazarlama bütçesinin tüm müşterilere harcanması yerine yalnızca potansiyel müşterilere harcanmasına olanak sağlayan bir yapay zeka modeli kullanılarak pazarlama bütçesinin daha verimli kullanılmasını öngören bir yöntem önerilmiştir. Modelin kurulum işlemi eğitim veri seti üzerinde yapılmış ve tabakalı örnekleme yöntemi ile model sonuçları genelleştirilmiştir. Daha sonra, test veri setinin yardımı ile kurulan modelin performansı değerlendirilmiştir. Çalışmada, makine öğrenmesi algoritmalarından karar ağacı, rastgele orman, çok katmanlı algılayıcı ve gradyan artırma algoritmaları kullanılarak çeşitli sınıflandırma modelleri ortaya konulmuştur. Ortaya konulan modellerin sonuçları birbirleriyle karşılaştırılmış ve en iyi performansı sergileyen algoritma sınıflandırma yöntemi olarak önerilmiştir. Deney veri seti üzerinde farklı makine öğrenmesi algoritmalarının analizinde KNIME 4.2.1 programı kullanılmıştır.

Bu çalışmanın sağladığı katkılar aşağıda sunulmuştur:

- ✓ Pazarlamanın etkin olabileceği müşteri kitlesinin belirlenmesi için kullanılan makine öğrenmesi algoritmalarına ait model performans ölçümüne ilişkin doğruluk, duyarlılık, kesinlik, seçicilik ve F-Ölçütü oranları birbirleriyle

karşılaştırılmış ve en iyi performansı sergileyen algoritmanın gradyan artırma makineleri olduğu görülmüştür.

- ✓ Kullanılan farklı makine öğrenmesi algoritmalarına ait model performans ölçümündeki doğruluk oranlarının birbirlerine yakın değerlerde olmasının yanı sıra, 1 değerine çok yakın yüksek değerde oldukları tespit edilmiştir. Doğruluk oranlarının birbirlerine yakın ve yüksek değerlerde olması çalışma kapsamında yararlanılan dört algoritmanın da pazarlama için ayrılacak bütçenin verimliliğinin arttırılması için kullanışlı olduğunu göstermektedir.

- ✓ Bu çalışmada kullanılan pazarlama verilerinin müşterilere ürünlerin gösterilip gösterilmediği ile ilgili "reklam (display)" değişkeni temel alınmış ve diğer değişkenlere ait özelliklerin doğru şekilde seçilmesi makine öğrenmesi algoritmalarının deney veri setine uygulanmasındaki karmaşıklık durumlarını azaltmış ve yüksek doğruluk oranlarının elde edilmesini sağlamıştır.

- ✓ Uygulamalı analiz açısından değişken önem analizinden sonra özelliklerin sayısı azaltılarak yüksek doğruluk değerleri elde edilmiştir.

- ✓ Gerçek dünya sisteminden elde edilen verilerin bütçeleme kararlarındaki satış tepkisinin modellemesine olumlu katkılar sunduğu görülmüştür.

Bu tez çalışmasında, gerçek dünya verilerinden elde edilen ve yapılan satışlar üzerinden birkaç özellik grubunun etkisi kullanılarak pazarlama için ayrılacak bütçenin verimliliğinin arttırılması amaçlanmaktadır. Bu amaç doğrultusunda genel bir çerçeve belirlenmiş ve bu çerçeve pazarlamanın etkin olabileceği kitle üzerinden farklı özelliklerdeki makine öğrenmesi teknikleri ile seçilerek skorlanmıştır. Bu bağlamda, belirlenen genel çerçeveye göre geçmiş yıllarda elde edilen müşteri bazlı pazarlama verileri üzerinden daha önce pazarlama bütçesi için ayrılmış (kampanyaya katılan müşteriler, çağrı merkezinin aradığı ve ürün satabildiği müşteriler vb.) ve başarılı olunmuş kitle işaretlenerek farklı makine öğrenmesi teknikleri ile modeller ortaya konulmuştur. Kullanılan her bir tekniğin matematiksel ve veri işleme farklılıklarının araştırma problemine farklı katkılar sunacağı öngörülerek, veri seti için karar ağacı, rastgele orman, çok katmanlı algılayıcı ve gradyan artırma algoritmalarından yararlanılmıştır. Ortaya konulan modellerden en başarılı olanı seçilerek, bu model için bir sonraki dönemin pazarlama bütçesini ilgilendiren tüm

müşteriler girdi olarak verilmiş, model üzerinden geçen veriler etiketlenerek skorlanmıştır. Bu sayede, en yüksek skor ve başarılı etiketini alan müşteriler hedeflenmiştir.

Yapılan çalışmanın amaçları doğrultusunda makine öğrenmesi, bütçe planlamasında makine öğrenmesinin yeri, makine öğrenmesinde tahmin ve analiz yöntemleri, özellik mühendisliği, veri etiketleme, çalışma kapsamında kullanılan makine öğrenmesi algoritmaları, model performans değerlendirmeleri ile ilgili kısımlar "Kavramsal Temeller" bölümü içerisinde anlatılmıştır. Modellerin analizinde kullanılan yazılım programı, verilerin elde edilmesinde yararlanılan veri toplama araçları ve veri seti ile ilgili bilgiler "Materyal ve Yöntem" bölümünde detaylandırılmıştır. Verilerin analizleri, pazarlama için ayrılacak bütçenin verimliliğinin arttırılmasına yönelik makine öğrenmesi algoritmalarının sonuçları, bu algoritmaların birbirleriyle karşılaştırılması neticesinde çalışmada kullanılan veri seti için en yüksek skoru alan modelin seçimi "Bulgular" bölümünde ayrıntılandırılmıştır. Son olarak, yapılan çalışma ile ilgili genel değerlendirmelere ve daha sonraki araştırmalar için yapılan önerilere "Sonuç ve Öneriler" bölümünde yer verilmiştir.

2. KAVRAMSAL TEMELLER

2.1. Makine Öğrenmesi

Makine öğrenmesi (ML), temelde insanların yaptığı gibi bilgisayar sistemlerinin verilere anlamlar verebildiği bilgisayar bilimi alanıdır. Bir başka deyişle makine öğrenmesi, bir algoritma veya yöntem kullanarak ham verilerden modelleri çıkaran bir yapay zeka türüdür. Makine öğrenmesinin temel odak noktası, bilgisayar sistemlerinin açıkça programlanmadan veya insan müdahalesi olmadan deneyimlerden öğrenmesine olanak sağlamaktır. Makine öğrenmesi algoritmaları, yeni çıktı değerlerini tahmin etmek için girdi olarak geçmiş verileri kullanır. Veri biliminde bir algoritma, istatistiksel işlem adımlarının bir dizisidir. Makine öğrenmesinde algoritmalar, yeni verilere dayalı kararlar ve tahminler yapmak için büyük miktarlardaki veri modellerini ve özelliklerini bulmak üzere "eğitilir". Algoritma ne kadar iyi olursa, daha fazla veri işledikçe kararlar ve tahminler de o ölçüde doğru olur.

Makine öğrenmesi, deneyimlerden öğrenen ve zaman içinde karar verme veya tahmin doğruluğunu iyileştiren uygulamalara odaklanır. Günümüzde, makine öğrenmesi örneklerine hemen hemen yaşamın her yerinde rastlanabilmektedir. Dijital asistanlar, sesli komutlarımıza yanıt olarak internet ağları üzerinde arama yapar ve isteklerimizi yerine getirir. İnternet siteleri, daha önce satın aldıklarımıza, izlediklerimize veya dinlediklerimize göre ürünler, filmler ve şarkılar önerir. Spam algılayıcılar, istenmeyen e-postaların gelen kutularımıza ulaşmasını engeller. Tıbbi görüntü analiz sistemleri, doktorların gözden kaçırmış olabilecekleri tümörleri tespit etmelerine yardımcı olur. Tüm bu gelişmelerde makine öğrenmesinin rolü büyüktür. Elbette gelişen program, yazılım, algoritma ve teknoloji sayesinde insanoğlu makine öğrenmesinden bugün elde ettiğinden daha fazlasını isteyecek ve bunun gerekliliklerini yerine getirecektir. Büyük veri daha da büyüdükçe, bilgi işlem daha güçlü ve uygun fiyatlı hale geldikçe ve veri bilimcileri daha yetenekli algoritmalar geliştirmeye devam ettikçe, makine öğrenmesi kişisel ve iş hayatımızda daha fazla verimlilik sağlayacaktır.

Avantajları söz konusu olduğunda, makine öğrenmesi, işletmelerin müşterilerini daha derinlemesine anlamalarına yardımcı olabilir. Makine öğrenmesi

algoritmaları, müşteri verilerini toplayarak ve bunları zaman içindeki davranışlarla ilişkilendirerek ilişkilendirmeleri öğrenebilir ve ekiplerin ürün geliştirme ve pazarlama girişimlerini müşteri talebine göre uyarlamasına yardımcı olabilir. Ancak makine öğrenmesinin bazı dezavantajları da bulunmaktadır. Her şeyden önce pahalı olabilir. Makine öğrenmesi projeleri, genellikle yüksek maaşlara hükmeden veri bilimcileri tarafından yönlendirilir. Bu projeler aynı zamanda yüksek maliyetli olabilecek yazılım altyapısı gerektirir.

Bir makine öğrenmesi uygulaması (veya modeli) oluşturmanın dört temel adımı vardır. İlk adım, bir eğitim veri setinin seçilmesi ve hazırlanmasıdır. Eğitim verileri, makine öğrenmesi modelinde tasarlanan sorunu çözmek için kullanacağı verileri temsil eden bir veri kümesidir. Bazı durumlarda eğitim verileri, modelin tanımlaması gereken özellikleri ve sınıflandırmaları belirtmek için veri etiketlemesi yapılır. Diğer veriler etiketlenmediğinden, modelin bu özellikleri çıkarması ve kendi başına sınıflandırmaları ataması gerekecektir. Her iki durumda da, eğitim verilerinin uygun şekilde hazırlanması, rastgele hale getirilmesi, kopyalarının çıkarılması ve eğitimi etkileyebilecek dengesizlikler veya yanlılıklar açısından kontrol edilmesi gerekir. Aynı zamanda, veriler uygulamayı eğitmek için kullanılacak eğitim alt kümesi ile test etmek ve iyileştirmek için kullanılan değerlendirme alt kümesi şeklinde iki alt gruba bölünmelidir.

İkinci adımda, eğitim veri kümesinde çalıştırılacak bir algoritma seçilir. Algoritma türü, eğitim veri setindeki veri türüne (etiketli veya etiketsiz) ve miktarına ve çözülecek problemin türüne bağlıdır. Regresyon algoritmaları, karar ağaçları ve örnek tabanlı algoritmalar etiketli veriler için yaygın olarak kullanılan makine öğrenmesi algoritmalarıdır. Kümeleme algoritmaları, ilişkilendirme algoritmaları ve sinir ağları etiketlenmemiş veriler için kullanılan algoritmalardır.

Üçüncü adımda, modeli oluşturmak için algoritma eğitilir. Algoritmayı eğitmek yinelemeli bir süreçtir. Değişkenleri algoritma aracılığıyla çalıştırmayı, çıktıyı üretmesi gereken sonuçlarla karşılaştırmayı, algoritma içinde daha doğru bir sonuç verebilecek ağırlıkları ve yanlılıkları ayarlamayı ve algoritma çoğu zaman doğru sonucu döndürene kadar değişkenleri tekrar çalıştırmayı içerir. Ortaya çıkan eğitimli doğru algoritma, makine öğrenmesi modelidir. Bu noktada, “algoritma” ve “model” makine öğrenmesinde dikkat edilmesi gereken önemli bir ayrım olup, birbirinin yerine yanlış şekilde kullanılmamalıdır.

Son adım, modelin kullanılması ve geliştirilmesidir. Bu adım, modeli yeni verilerle ve en iyi durumda, zaman içinde doğruluğu ve etkinliği iyileştirmek için kullanmaktır. Yeni verilerin nereden geldiği çözülmekte olan soruna bağlı olacaktır.

Makine öğrenmesi algoritmaları uzun yıllar piyasada olmasına rağmen, yapay zeka uygulamalarının öne çıkmasıyla yeniden popülerlik kazanmıştır. Özellikle derin öğrenme modelleri, günümüzün en gelişmiş yapay zeka uygulamalarını güçlendirmektedir. Makine öğrenmesi platformları, kurumsal teknolojinin en rekabetçi alanları arasında yer almaktadır. Veri işleme, etiketleme, sınıflandırma ve analiz ile ilgili ana konular, veri sunumunun ve depolamanın optimizasyonu, hızlı bilgi erişim algoritmalarının oluşturulması ve tavsiye sistemlerinin tasarımı ile ilgilidir. Dahası, her şirket veya birey verilerini farklı şekilde kullandığı ve analiz ettiği için, makine öğrenmesi modelleri için verilerin etiketlemesinde benzersiz mekanizmalara yatırım yapmalıdır. Amazon, Google, Microsoft ve IBM başta olmak üzere birçok şirket, müşterileri için veri toplama, veri hazırlama dahil olmak üzere veri sınıflandırma, model oluşturma, eğitim ve uygulama dağıtımı gibi makine öğrenmesi faaliyetlerinde yarışmaktadırlar. Makine öğrenmesinin iş operasyonları için öneminin artmaya devam etmesi ve yapay zekanın kurumsal ortamlarda daha pratik hale gelmesi, şirketler açısından makine öğrenmesindeki rekabetin önümüzdeki yıllarda daha da artacağı tahmin edilmektedir.

2.2. Bütçe Planlamasında Makine Öğrenmesinin Yeri

Bütçeleme, planlama ve tahmin süreci, bir organizasyonun finansal performans yönetiminde kilit rol oynayan unsurlardır. Bu süreçler; gelecekteki gelirleri, giderleri ve nakit akışını doğru ve verimli bir şekilde üretmeyi amaçlamanın yanı sıra, bir işletmeye bütçeleme ve stratejik planlama konusunda yararlı bilgiler sağlar. Tüm üst düzey kuruluşlar, gelecekteki gelir ve harcama modellerini belirlemek ve tahmin etmek için finansal tahminlerden yararlanır. Bu mali eğilimler genellikle şirket hedefleri ve objektif stratejiler üzerinde doğrudan veya uzun vadeli bir etkiye sahiptir. Bu süreç, hesaplardan, satışlardan ve dış pazar veya ekonomik göstergelerden alınan geçmiş verileri ve mali kayıtları kullanır. Çoğu şirket, geçmiş ve güncel bilgileri analiz eden ve belirli bir dönem için gelecekteki mali eğilimleri ve koşulları yansıtan mali yönetim araçlarını kullanmaktadır. Tanımlanan modeller; işletme politikalarına, pragmatik karar vermeye ve nakit akışı yönetimine rehberlik etmeye yardımcı olur.

Mali tahmin süreci, tahminin varsayımlarını ve dinamiklerini tanımlamakla başlar. Bu aşamadan sonra gelecekteki sonuçları geliştirmek için en uygun metodoloji belirlenir. Finansal tahmin, bir şirketin gelecekteki mali sonuçlarının tahminlerini içeren tabloların veya raporların hazırlanmasını içerir. Tahmin süreci; ilgili kayıtları, pazar araştırması bulgularını, ekonomik anketleri ve endüstriyel ekonomik koşulları inceler. Mali kararları ve mali ortamı etkileyen mevcut politikalara, yasalara, hedeflere ve temel konulara bakar. Finansal tahmin süreci, mali niteliksel bir model geliştirmeyi amaçlamaktadır. Model, finansal durum tablosu (veya bilanço) biçimindeki tahminlerden, özkaynaklardaki veya hissedarların özkaynağındaki değişikliklerin tablolarından, gelirden ve nakit akış tablolarından oluşur. Finansal tahmin, yönetim ekiplerinin şirketin doğru yönde ilerleyip ilerlemediğini bilmelerine yardımcı olur. Bu bilgiler, gerekli ayarlamaları yapmalarına ve stratejik iş planları geliştirmelerine olanak tanır. Çoğu durumda, finansal tahminler, bütçe oluşturmada ve iş performansı yönetimini iyileştirmede yararlıdır.

Şirketler teknoloji ve müşteri tabanı açısından büyüdükçe, genellikle büyük verilere karşı tek bir zorlukla karşılaşırlar. Geleneksel finansal tahmin yöntemleri küçük verilerle çok iyi çalışır. Büyük verileri analiz etmek için geleneksel yöntemleri kullanan bir finansal tahmin süreci daha fazla zaman ve kaynak tüketir. Karmaşık teknikler, daha doğru yanıtlar üretmek için veri, uzmanlık ve çaba gerektirir.

Büyük veri setleri söz konusu olduğunda, bunlardan değer elde etmek için daha akıllı veri bilimi modellerine ihtiyaç duyulmaktadır. Ayrıca, doğru ve etkili finansal tahminler geliştirmek çoğu yönetim ekibi için zorlu bir iştir. Bu nedenle, şirketlerin finansal modeller geliştirmesine yardımcı olacak en iyi çözüm makine öğrenmesidir. Makine öğrenmesi araçları, teknikleri ve sistemleri tahmincilerin ve analistlerin büyük bir tarihsel veri kümesinden gelecekteki eğilimleri ve sonuçları tahmin etmelerine yardımcı olur. Ham verilerden daha derin içgörüler elde etmenin yanı sıra, muhakeme ve karar verme gerektiren katma değerli faaliyetlere odaklanmalarına olanak tanır.

Makine öğrenmesi, pazarlamada bilgi üretimi için çeşitli avantajlar ve yeni perspektifler sunmaktadır. Makine öğrenmesi diğer kullanımlarının yanı sıra, regresyon ve sınıflandırma problemlerini çözmek, kümeleme, görselleştirme, boyutluluk azaltma, ilişki kuralları oluşturma, pekiştirme-öğrenme aracıları geliştirme için uygulanabilir. Çeşitli uygulamalarına rağmen, makine öğrenmesi teknikleri genellikle pazarlamada yaygın olarak uygulanan geleneksel ekonometrik yöntemlerle

karşılaştırılmaktadır. İlk olarak, makine öğrenmesi teknikleri en iyi örneklem dışı tahminleri elde etmeye odaklanırken, nedensel ekonometrik yöntemler en iyi tarafsız tahmin edicileri türetmeyi amaçlamaktadır. Bu nedenle, nedensel çıkarım teknikleri, örneklem dışı tahminler yapılırken genellikle iyi performans göstermez. Bunun nedeni, en iyi tarafsız tahmincinin her zaman en iyi örneklem dışı tahminleri sağlamamasıdır. Makine öğrenmesi literatüründe, bu sorun yanlılık-varyans ödünleşimi olarak bilinir. İkincisi, ekonometrik yöntemlerin aksine, verilerde gözlenen sonuçların üretildiği süreç hakkında önsel kuramın olmadığı bir durumda çok sayıda makine öğrenmesi yaklaşımı geliştirilmiştir. Üçüncüsü, pazarlamada kullanılan birkaç ampirik yöntemin aksine, makine öğrenmesi yöntemleri son derece fazla sayıda değişkeni işleyebilir ve hangi değişkenlerin tutulması ve hangilerinin analizden çıkarılması gerektiğini seçebilir. Son olarak, makine öğrenmesi pazarlamada ölçeklenebilirlikle başa çıkmak için güçlü bir yaklaşım olabilir. Pek çok makine öğrenmesi yaklaşımı, ölçek ve verimlilik elde etmek için özellik seçimi ve optimizasyonu uygular. Bu, birçok farklı gerçek zamanlı pazarlama problemi için giderek daha önemli bir hedeftir. Pazarlamadaki makine öğrenmesi uygulamalarının çoğu, tanımlayıcı / keşifsel veya tahmine dayalı bir bakış açısı benimser. Makine öğrenmesi ve nedensel modellemenin kesişimine dayanan tahmine dayalı modelleme perspektifi, pazarlama teorilerinin geliştirilmesi ve test edilmesi için tanımlayıcı / keşifsel olanlardan daha güçlü bir perspektif sunar (Shmueli, 2010).

2.3. Tahmin ve Analiz

Tez çalışması kapsamında, bir modelin gerçek bir dünya sisteminde yaşamı boyunca davranışı tanımlanmaktadır. Gerçek dünya sisteminin kullanımı, bir modelin mevcut durumunun analiz edilmesi ve gelecekteki bir durumun tahmininde önemli olan dinamik davranışı sergilemektedir. Bu tür modelleri inşa etmek için birçok olasılık vardır. Her tahmin ve analiz sistemi aynı soyut davranış ile çalışır. Bu durum, Şekil 2.1 ile görselleştirilmiştir. Gerçek dünya sistemindeki veriler modelin işleyebileceği bir gösterime dönüştürülmeden önce toplanmalıdır. Son olarak, model verileri işledikten sonra tahmin ve analiz sistemi sonuçlarını ortaya çıkarır. Gerçek dünya sisteminin tahmin ve analizine ilişkin model yaklaşımları aşağıdaki alt bölümlerin içerisinde ele almaktadır.

Şekil 2.1. Tahmin ve analiz sisteminin iş akış şeması.

2.3.1. Fiziksel Modele Dayalı Metodoloji

Fiziksel modeller genel olarak matematiksel modeller temelinde gerçekleştirilir. Diferansiyel denklemler, ilgili bileşenlerin sağlamlığı üzerinde etkisi olan fiziksel süreçleri belirtmek için kullanılır. Modeller alanında uzman kişiler aracılığıyla ortaya çıkarılır. Alan uzmanları gerçek dünya sistemindeki fiziksel süreçlerle ilgili bilgi sahibi olmanın yanı sıra, bu tür sistemleri modelleme becerisine de sahip olmalıdırlar. Bir sistemi modelledikten sonra ortaya çıkan model, doğru davranışı kanıtlamak için geniş bir veri kümesi kullanılarak onaylanmalıdır. Model doğru uygulanırsa, modelin girdi parametrelerini toplamak için izleme sistemleri kullanılabilir. Model, çıktı değerini üretir ve verilen çıktıyla gerçek dünya sisteminin durumunu tespit etmek için eşiği tanımlamak için istatistiksel teknikler kullanılır.

Fiziksel modeller, birçok çalışma koşulunda çok dinamik gerçek dünya sistemleri için kullanışlıdır. Fiziksel bir model oluşturmak, gerçek dünya sistemini çok iyi anlamayı gerektirir. Genellikle, izlenen belirli parametreler için analizin ve tahminin doğruluğu, modelin karmaşıklığı ile artar. Yüksek maliyet ve bileşen özelliği fiziksel modellerin dezavantajlarıdır. Bu dezavantajlar, fiziksel modellerin diğer bileşen türlerine uygulanmaması anlamını taşımaktadır (Brotherton vd., 2000). Buna ilaveten, gerçek dünya sistemleri izlenmesi zor birçok çevresel faktörden etkilendiğinden, iyi bir doğrulukla fiziksel bir model inşa etmek oldukça zordur.

2.3.2. Bilgiye Dayalı Modeller

Fiziksel modele dayalı metodolojiye benzer şekilde bilgiye dayalı metodoloji, alan uzmanları tarafından oluşturulur, ancak fiziksel davranışını kapsayacak matematiksel bir model yoktur. Bilgiye dayalı bir modelin kalitesi, alan uzmanının deneyimine bağlıdır. Bu tür sistemler, alan uzmanının bilgilerini resmileştirmeye çalışır.

2.3.2.1. Uzman modeller

Uzman modeller, genellikle uzmanlar tarafından çözülen problemler için uygundur. Kurallarla tanımlanırlar ve gerçek dünya sisteminin durumlarını tanımlarlar. Kurallar; IF koşulu, THEN ve ELSE sonucu ifade eder. Bu tür kurallar mantık operatörleri ile birleştirilebilir. Koşullar, bir sonucun veya başka bir kuralın sonucu olan gerçek dünya sisteminin durumunu tanımlar. Alan bilgisini elde etmek ve onu kural tabanlı bir sisteme dönüştürmek zordur (Tu vd., 2001). Uzman bir model, kural tabanlı sistemde kapsanmayan durumların üstesinden gelemez. Dahası, uzman bir model kuralların sayısı önemli ölçüde arttığında ortaya çıkan bir kombinasyon patlamasıyla karşılaşabilir. Diğer taraftan, bir uzman modeli tanımlandıktan sonra, aşağıda anlatılan veri odaklı bir yaklaşımla model iyileştirilebilir. Bu çözüm, veriye dayalı modelden veya tek başına uzman modelden daha iyi sonuçlara ulaşır. Bu bağlamda, optimum çözüme daha hızlı ve daha yakın bir şekilde yaklaşmak için basit bir uzman modeli ile veri odaklı bir yaklaşım birlikte kullanılabilir (Zhou vd., 2013).

2.3.2.2. Bulanık mantık modelleri

Bulanık mantık; belirsiz, kesin olmayan, gürültülü veya eksik girdilere dayalı gerçek dünya sistemini tanımlamanın basit bir yoludur (Zadeh, 1965). Geleneksel Boole mantığının bir üst kümesidir. Geleneksel bir Boole mantığı ile bulanık mantık arasındaki temel fark, karar verme sürecidir. Geleneksel Boole mantığı açısından bir öğe ya bir kümenin üyesidir ya da değildir. Belirli bir öğe için, bulanık mantık, bir kümeye üyelik derecesini döndürür. Bulanık bir mantık modeliyle durumlar, gerçek bir dünya sistemindeki durum geçişleri gibi sürekli ve örtüşen bir şekilde tanımlanabilir. Bu nedenle, bulanık mantıkta bir gerçek dünya sisteminin açıklaması, sayısal veya matematiksel bir açıklamadan daha sezgisel ve daha az spesifiktir.

Bulanık mantık modeli, daha basit, daha sezgisel ve daha iyi davranan modeller yaratabilir (Peng vd., 2010). Daha önce de belirtildiği gibi, bulanık mantık durumları üst üste gelebilir. Her girdi değeri, bir üyelik derecesine sahip belirsiz bir kurala aittir. Üyelik derecesini kapsayan aralık 0 (kesinlikle üye değil) ile 1 (kesinlikle üye) arasında başlar. Geleneksel Boole mantığının aksine, bir giriş değeri bir kuralla eşleşsin veya eşleşmesin, bulanık bir mantık modeli tüm üyelik derecelerini dikkate alabilir. Bu, tüm üyelik derecelerinin sonucu ve bir durumu temsil eden kuralı

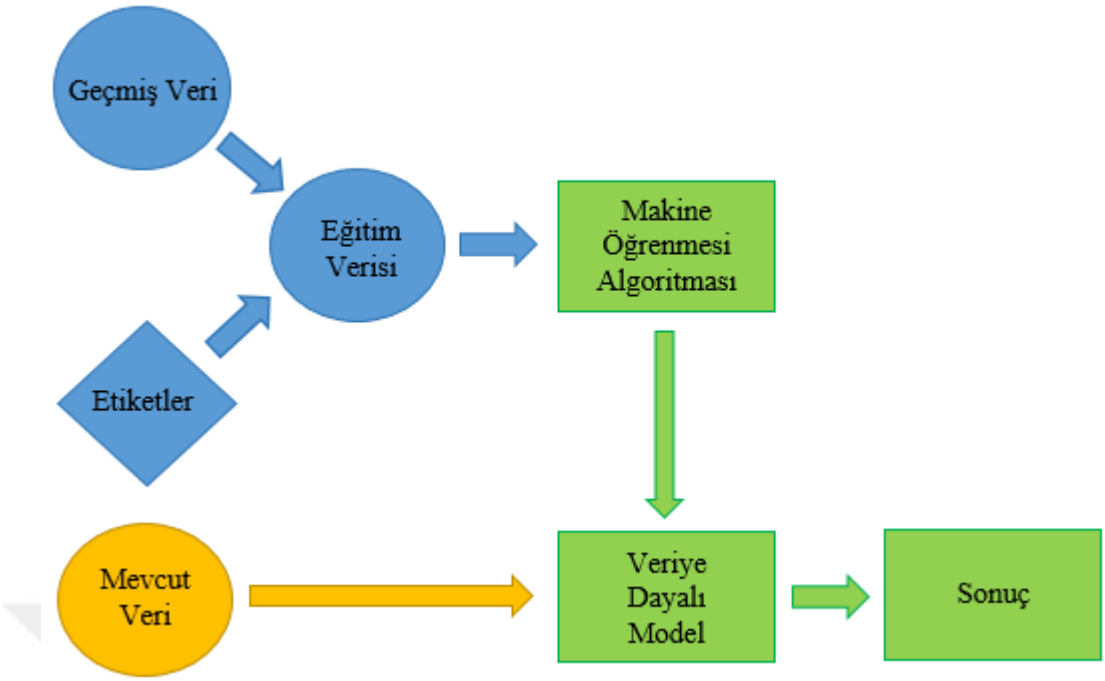
etkileyebileceği anlamına gelir; burada en yüksek sayısal sonuç otomatik olarak hesaplanan sonucun durumu değildir.

2.3.3. Veriye Dayalı Modeller

Makine öğrenmesi algoritmaları, önceden açık bir programlama görevi olmadan doğrudan verilerden öğrenebilir. Verilerden öğrenme yeteneği sayesinde makine öğrenmesi, özellikle tam olarak tanımlanamayacak kadar karmaşık veya tanımı doğru bir şekilde yapılamayan problemler için uygundur. Verilerden kaynaklanan çeşitli sorun veya problemlere birçok makine öğrenmesi algoritması uygulanabilir. Bu algoritmalar üç ana gruba ayrılabilir:

2.3.3.1. Denetimli öğrenme

Denetimli öğrenme, belirli bir eşleştirilmiş giriş-çıkış eğitim örneği kümesine dayalı olarak bir sistemin girdi-çıktı ilişkisi bilgilerini elde etmeye yönelik bir makine öğrenmesi paradigmasıdır. Çıktı, girdi verilerinin veya denetimin etiketi olarak kabul edildiğinden, bir girdi-çıktı eğitim örneğine etiketli eğitim verileri veya denetimli veriler de denir (Li ve Wu, 2012). Denetimli öğrenmede, gerçek dünya sisteminde belirli bir zaman içerisinde elde edilen verilerden yararlanılır. Toplanan veri kayıtlarının tamamı öğrenme örnekleri olarak kullanılamayacağından, verilerin öncelikli olarak filtrelenmesi gerekmektedir. Her bir eğitim verisinin kaydı, beklenilen sonuca göre etiketlenir. Etiketli eğitim verileri bir algoritma yardımıyla işlenir. Bu aşamada makine öğrenme algoritmasının görevi, girdi verileri ve etiketlenen çıktı arasındaki ilişkiyi inceler ve ardından veriye dayalı bir model oluşturmaktır. Denetimli öğrenmenin amacı, girdi ve çıktı arasındaki eşleştirmeyi öğrenebilen ve yeni girdiler verildiğinde sistemin çıktısını tahmin edebilen yapay bir sistem inşa etmektir. Denetimli makine öğrenmesine ait veriye dayalı bir modelin iş akış şeması Şekil 2.1 içerisinde sunulmuştur.



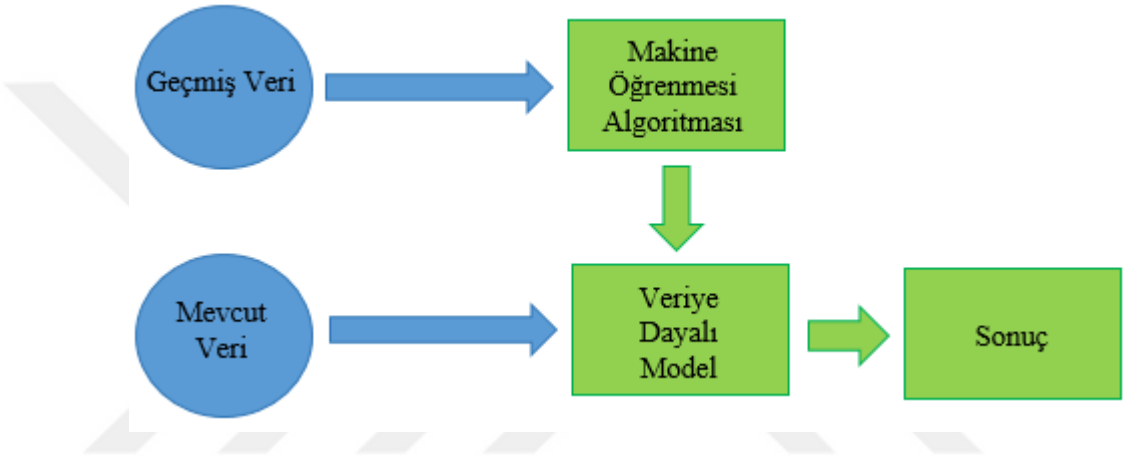
Şekil 2.2. Denetimli makine öğrenmesine ait veriye dayalı bir modelin iş akış şeması.

Denetimli öğrenme, algoritmanın girdi ve çıktı değerlerine erişebildiği her görevi içermektedir. Girdi değerleri, özellik değerleri ve meta veriler gibi algoritmanın kullanmasına izin verilen harici bilgiler olarak tanımlanırken, çıktı değerleri, sınıf özelliklerinin belirli etiketleridir. Bu durum, verilerin yapısının zaten bilindiği ve bu programların amacının yeni verileri doğru sınıflara atamak olduğu anlamına gelmektedir.

2.3.3.2. Denetimsiz öğrenme

Denetimsiz öğrenme, sınıflandırılmamış veya etiketlenmemiş veri noktaları içeren veri setlerindeki modelleri tanımlamak için makine öğrenmesi algoritmalarının kullanılması anlamına gelir. Böylece algoritmaların, bu görevi yerine getirirken herhangi bir dış kılavuza sahip olmadan veri kümeleri içinde yer alan veri noktalarını sınıflandırmasına, etiketlemesine ve gruplandırmasına izin verilir. Başka bir deyişle, denetimsiz öğrenme, sistemin kendi başına veri kümeleri içindeki kalıpları tanımlamasına izin verir. Denetimsiz öğrenmede bir makine öğrenme algoritması herhangi bir kategori sağlanmasa bile, sıralanmamış bilgileri benzerlik ve farklılıklara göre gruplayacaktır. Denetimsiz öğrenme algoritmaları, denetimli öğrenme sistemlerine göre daha karmaşık işleme görevleri gerçekleştirebilir. Denetimsiz

öğrenme veri setlerini eğitmek için makine öğrenmesi algoritmalarından geçirildiğinde başlar. Bu tür sistemleri eğitmek için kullanılan veri setlerinde yer alan hiçbir etiket veya kategori yoktur; eğitim sırasında algoritmalardan geçirilen her veri parçası, etiketlenmemiş bir girdi nesnesi veya örnektir. Denetimsiz öğrenmenin amacı, algoritmaların eğitim veri kümeleri içindeki modelleri tanımlamasını ve girdilerini sistemin kendisinin tanımladığı modelleri göre kategorize etmesini sağlamaktır. Algoritmalar, bunlardan yararlı bilgiler veya özellikler çıkararak veri setlerinin temel yapısını analiz eder. Denetimsiz makine öğrenmesine ait veriye dayalı bir modelin iş akışı Şekil 2.2'de verilmiştir.



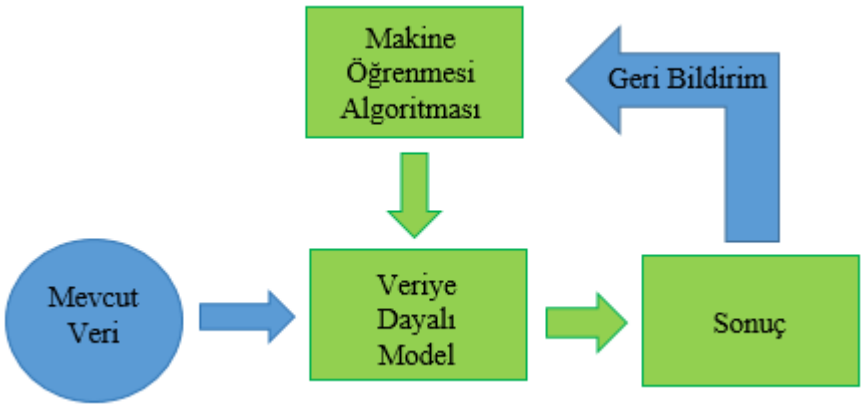
Şekil 2.3. Denetimsiz makine öğrenmesine ait veriye dayalı bir modelin iş akış şeması.

Denetimsiz öğrenme, denetimli öğrenmenin aksine çıktı değerlerine erişimi olmayan tüm görevleri içermektedir. Bu nedenle, denetimsiz öğrenmede kendi başlarına sınıflar oluşturularak verilerdeki yapılar bulunmaya çalışılır.

2.3.3.3. Pekiştirmeli öğrenme

Pekiştirmeli öğrenme, denetimli ve denetimsiz öğrenmenin aksine geçmiş verilere ait herhangi bir öğrenme işlemi bulunmamaktadır. Bu nedenle, makine öğrenmesi için kullanılan algoritmanın deneme yanılma işlemine dayanması durumu söz konusudur. Öğrenme aşamasının başında bir algoritma bazı yönlerden verilere dayalı olan modelin tanımlamasını yapar. Daha sonra, modelin tanımlandığı algoritmanın da yönleri tanımlanır. Bu sayede, tanımlanan model gerçek dünya sisteminde var olan verilere dayanan bir sonucu ortaya çıkarır. Ortaya çıkan sonuç, beklenilen sonuç ile eşleştiğinde değerlendirmeye alınır. Algoritma tarafından alınan

değerlendirme, daha başarılı stratejilerin geliştirildiği ve başarısız olan stratejilerin reddedildiği sonuçların niteliği ile ilgili bir geri bildirimdir. Pekiştirmeli makine öğrenmesine ait veriye dayalı bir modelin iş akışına Şekil 2.3 içerisinde yer verilmiştir.



Şekil 2.4. Pekiştirmeli makine öğrenmesine ait veriye dayalı bir modelin iş akış şeması.

Fiziksel model, bilgiye dayalı model ve veriye dayalı model birbirinden oldukça farklıdır. Bununla birlikte, farklı model türlerini birleştirerek avantaj elde etme olasılıkları vardır. Bu tez çalışmasında, daha genel ve farklı gerçek dünya sistemlerine uygulanabilen veriye dayalı modellere odaklanılmaktadır. Denetimli öğrenme tekniği ve bir uzman model ile veriye dayalı bir modelin kombinasyonlarını kullanan bazı uygulamalar vardır. Bu modeller, daha iyi sonuçlar elde etmek için veriye dayalı modellerin toplanan gerçek dünya bilgileriyle birlikte uzman bir modelin alan bilgisini kullanmaya çalışır. Bir sonraki kısımda, veriye dayalı bir sistemin elde edilen verileri nasıl önceden işlediği ve ondan nasıl öğrendiğini tartışır.

2.4. Özellik Mühendisliği (Feature Engineering)

Özellik, ham verilerin sayısal bir temsilidir. Ham verileri sayısal ölçümlere dönüştürmenin birçok yolu bulunduğundan özellikler birçok şeye benzeyebilir. Doğal olarak, özellikler mevcut veri türünden türetilmelidir. Özelliklerin modele bağlı oldukları gerçeği daha az açıktır. Bazı modeller, bazı özellik türleri için daha uygundur ve bunun tersi de geçerlidir. Doğru özellikler, eldeki görevle ilgilidir ve uygulanması model için kolay olmalıdır. Özellik mühendisliği, verilere, modele ve göreve verilen en uygun özelliklerin formüle edilmesi sürecidir (Zheng ve Casari, 2018). Özelliklerin sayısı önemlidir. Bu bağlamda, eğer yeterli bilgilendirici özellik yoksa, model nihai

görevini yerine getiremeyecektir. Diğer taraftan, çok fazla özellik varsa veya bunların çoğu alakasızsa, model daha pahalı ve eğitilmesi zor olacaktır. Eğitim sürecinde modelin performansını etkileyen bir şeyler ters gidebilir (Zheng ve Casari, 2018). Makine öğrenmesi algoritmalarının performansı, giriş verilerinin özellik temsiline büyük ölçüde bağlı olduğu bilinmektedir (Bengio vd., 2013). İyi bir özellik seti, hızlı ve basit modellerin seçimine olanak sağlayan esneklikleri de beraberinde getirir. Ancak, verilerin ham gösterimleri genellikle öğrenmeye uygun değildir (Domingos, 2012). Özellik mühendisliği, verilerdeki gizli modelleri keşfederek mevcut ham özelliklerden yeni özellikler oluşturma sürecidir (Zabokrtsky, 2015). Ayrıca özellik mühendisliği mevcut özellik setini zenginleştirmeyi ve buna bağlı olarak öğrenme algoritmalarının öngörü gücünü artırmayı amaçlamaktadır. Bu nedenle, özellik mühendisliği, makine öğrenmesinin uygulamadaki başarısında önemli bir rol oynamaktadır (Domingos, 2012). Tipik bir özellik mühendisliğinin önerilen süreci özellik oluşturma, özellik seçimi ve model değerlendirme aşamalarından oluşmaktadır (Zabokrtsky, 2015).

2.4.1. Özellik Oluşturma (Feature Construction)

Bu aşama, verilerdeki ham özelliklerden ekstra özellikler çıkarmayı hedefler. Özellik oluşturma, manuel (veri analizi yoluyla veya alan bilgisini hesaba katarak özelliklerin çıkarılması) ve otomatik (ham özellikleri bir araya getirerek, birleştirerek veya dönüştürerek özellikleri ayıklama) olabilir. Bu adım, toplam özellik sayısında hızlı bir artışa neden olabilir.

2.4.2. Özellik Seçimi (Feature Selection)

Çok sayıda özelliğe sahip olan öğrenme algoritmaları, boyutluluk laneti nedeniyle performansın düşmesine yol açacak şekilde aşırı uyum gösterebilir. Özellik seçimi, hesaplama maliyetini düşürmek ve doğruluğu artırmak için yaygın ve kullanışlı bir tekniktir. Özellik seçimi, bazı uygunluk değerlendirme kriterlerine göre en alakalı özelliklerin küçük bir alt kümesini seçmeyi amaçlar. Özellik seçiminin öğrenme performansını artırmaya, hesaplama maliyetini düşürmeye ve model yorumlanabilirliğini iyileştirmeye yardımcı olabileceği kanıtlanmıştır (Li vd., 2017).

Özellik seçimi, verilerin önceden işlenmesinde ve veri boyutluluğunun azaltılmasında etkilidir. Ayrıca, başarılı veri madenciliği ve makine öğrenmesi uygulamaları için çok önemlidir. İstatistik, örüntü tanıma, makine öğrenmesi ve veri

madenciliği (web, metin, görüntü ve mikro diziler dahil) gibi birçok alanda pratik önemi olan zorlu bir araştırma konusudur. Özellik seçiminde daha basit ve daha kapsamlı modeller oluşturmak, veri işleme performansını iyileştirmek, temiz ve anlaşılır verilerin hazırlanmasına yardımcı olmak hedeflenmektedir. Özellik seçiminde makine öğrenmesi algoritmalarının kullanmasının temel nedeni, bir makine öğrenmesi algoritmasını eğitmek ve eğitilen yapıyı analiz etmektir. Genel olarak, bazı özellikler algoritmaya daha uygundur ve diğerleri, girdi değeri ile belirli bir çıktı arasında bir korelasyona ulaşmakla daha az ilgilidir. Daha sık kullanılan özellikler önemli olanlardır. Daha az hesaplama maliyeti ve daha fazla doğruluğa ulaşmak için, başka bir makine öğrenmesi yaklaşımı yalnızca ilk makine öğrenmesi yaklaşımının daha sık kullanılan özelliklerini kullanmaktadır (Chen vd., 2011).

2.4.2.1. Korelasyon analizi

Korelasyon analizi, özellik seçiminde kullanılan en yaygın tekniklerden biridir. Bu yaklaşım, gereksiz davranışlara sahip özellikleri ortadan kaldırabilen bir korelasyon fonksiyonuna sahip ikili girdiler arasındaki doğrusal bir ilişkiyi değerlendirir (Hair vd., 2006). Ancak bu yaklaşımda modeldeki veri miktarı düşükse veya özellikler arasındaki ilişki doğrusal değilse başarısız olur.

2.4.2.2. Temel bileşenler analizi

Bir dizi özelliği, temel bileşenler adı verilen doğrusal ilişkisiz bir dizi özelliğe dönüştüren ortogonal bir dönüşüm tekniğidir. Bu yaklaşımın varsayımı, en büyük varyansa sahip özelliklerin en büyük bilgilendirici içeriğe sahip olmasıdır. Modeller, en ilgili özelliklere göre ortalanır ve döndürülür. Temel bileşen analizinin sonucu, birbirine ortogonal olan en büyük varyansa sahip özellikleri verir.

2.4.2.2. Entropi

Entropi, makine öğrenmesi içerisinde işlenmekte olan bilgideki rastgeleliğin bir ölçüsüdür. Entropi ne kadar yüksekse, bu bilgiden herhangi bir sonuç çıkarmak o kadar zor olur. Entropi terimi, Shannon tarafından bilgi içeriği hakkındaki iddialar ile istatistiksel temellere dayandırılmıştır. Bu bağlamda, bilgi kalitesi için yapılan tanımlamalar aşağıdaki gibidir (Shannon, 2001):

- ✓ Bir olay ne kadar nadir görülürse, bilgi içeriği o kadar yüksek olur.
- ✓ Bir olay zincirinin bilgi içeriği, meydana gelen tüm olayların toplamıdır.

x_1'den x_n'e kadar meydana gelen olaylar $I(x_1, \dots, x_n) = I(x_1) + \dots + I(x_n)$ eşitliği ile ifade edilmektedir.

- ✓ Belirli bir olayın bilgi içeriği 0'dır.

Bu bilgilerden yola çıkarak bilgi içeriğini logaritmik olarak tanımlarsak;

$$I(x) = \log_2(\frac{1}{p(x)}) \tag{2.1}$$

Denklem 2.1'deki x olaydır ve $p(x)$, x olayının meydana gelme olasılığıdır. Bir olay kümesinin entropisi H ise Denklem 2.2 ile şu şekilde tanımlanır:

$$H = \sum_{i=1}^{n} p(x_i)\ I(x_i) \tag{2.2}$$

Buradaki n, olayların sayıdır. En yüksek entropi H, tüm olayların aynı olasılığa sahip olması durumunda mevcuttur ($H = \log_2(n)$). Bu, entropi ne kadar düşükse, bilgi içeriğinin o kadar az olay üzerinde yoğunlaştığı ve tek bir olay için entropi $H(x)$ ne kadar yüksek olursa, x olayı için bilgi içeriğinin o kadar yüksek olduğu anlamına gelir. Yine, x olayı ne kadar az meydana gelirse, bilgi içeriği o kadar yüksek olacak demektir.

Shannon entropisi, olasılık dağılımlarının bilgi içeriği için bir ölçü olarak kullanılır. Shannon'un tanımına göre, verilerden bir model tahmin ederken, belirli bir veri oluşturma sürecinin varsayılması gerekir. Böyle bir modelin parametreleri, model ile gözlemlenen veriler arasındaki uyumu maksimize eden değerler olmalıdır. Bunu yaparak, tahmin edilen miktarların aslında rastgele değişkenler olduğu varsayılır. Rastgele değişkenler, sürekli bir aralıkta herhangi bir değere sahip olabildikleri (sürekli iseler) veya belirli olasılıklarla (kesikli değişkenler durumunda) belirli değerler alabildikleri için değeri belirsiz olan niceliklerdir (Gadaleta, 2019). Rastgele bir değişkenin gözlenen değeri yüksek şaşırtıcılık ile beklenmedik bir sonucu veriyorsa, bu özbilgi olarak adlandırılır (Preiswerk, 2018). Bunun arkasındaki sebep basittir: Böyle bir değer gözlemlendiğinde rastgele bir değişkenin davranışı hakkında ek bilgi edinilir (Gadaleta, 2019).

Entropi, özellik seçimi için ilişki düzeyini korur. Çünkü bir özelliğin entropisi H ne kadar düşükse, bilgi içeriği o kadar yüksek olur. Bir özelliğin karmaşıklığını ölçmek için de entropi yararlı olabilir. Buna göre, örnek entropi veya örnek entropisine dayalı çok ölçekli entropi kullanılabilir. Örnek entropide, bir özelliğin bir zaman aralığındaki değerler başka bir zaman aralığının aynı özelliğine sahip olanlarla

karşılaştırılır. Çok ölçekli entropi, örnek entropi sürecini kullanır, ancak zaman penceresi ölçeklendirilir (Jahnke, 2015). Özellik seçimi, bir sınıflandırma veya regresyon algoritmasının başarısını belirlemede çok önemli bir role sahiptir. Bazı değişkenler yüksek oranda ilişkili olabilir, bazıları ise sadece gürültü ve çok az sinyal taşıyabilir. Buna ilaveten, daha yüksek boyutlu vektörlerin boyutsal uzayı azar azar arttıkça daha fazla benzer görünür. Boyutluluk laneti olarak bilinen bu olgu, gözlemler arasındaki mesafeyi hesaplamaya dayanan algoritmalar için çok zararlıdır. Bu durumda, özellik seçimi ilgili özelliklerin tespit edilmesine ve ilgisiz olanların ayrılmasına izin verdiği için çok yardımcı olabilir. Bu genellikle bir öğrenme algoritmasının iyileştirilmiş genelleştirilmesi, verilerin anlaşılması, bellek ve CPU gereksinimlerinin azaltılması gibi çeşitli avantajlar sağlar (Gadaleta, 2019).

2.4.3. Model Değerlendirme (Model Evaluation)

Bu aşamada, seçilen özellikler kullanılarak görünmeyen veriler üzerinde model performansı tahmin edilir. Bir makine öğrenmesinde yalnızca model değil, aynı zamanda özellikler de seçilmektedir. Bu çift yönlü bir akış koldur ve birinin seçimi diğerini etkiler. İyi özellikler, sonraki modelleme adımını kolaylaştırır ve sonuçta ortaya çıkan model, istenen görevi tamamlama konusunda daha yetenekli hale getirir. Kötü özellikler, aynı performans düzeyine ulaşmak için çok daha karmaşık bir model gerektirebilir.

2.5. Veri Etiketleme

Makine öğrenmesi açısından veri etiketleme, veri örneklerinin algılanması ve etiketlemesi süreci olarak ifade edilebilir. Etiketleme işleminin manuel olarak yapılabilmesinin yanı sıra, genel olarak yapılan etiketlemeler yazılım programları tarafından desteklenmektedir. Veri etiketleme, makine öğrenmesi için veri ön işlemenin önemli bir parçasıdır. Gelecek verilerin işlenmesinde öğrenmeye bir temel sağlaması adına hem girdi hem de çıktı verilerinin sınıflandırılmasında etiketleme durumu özellikle denetimli öğrenmeye büyük katkılar sunmaktadır. Denetimli öğrenme için, etiketlenmesi gereken geçmiş verilere ihtiyaç vardır. Dolayısıyla, makine öğrenmesi yaklaşımlarında, girdi özelliği vektörü ile beklenen sonuç arasında ilişkiler kurabilir. Ayrıca makine öğrenmesinde sınıflandırma ve regresyon arasında ayrım yapabilir.

Günlük hayatta otonom araçlar üzerinden bir örnek vererek veri etiketleme kavramını pekiştirmeye çalışalım: Otonom araçlar için makine öğrenmesi algoritmaları oluşturulurken veri etiketlemeden yararlanılır. Kendi kendine giden arabalar gibi otonom araçların, dış dünyayı işleyebilmeleri ve güvenli bir şekilde yol alabilmeleri için rotalarındaki nesneler arasındaki farkı bilmeleri gerekir. Veri etiketleme, otomobilin yapay zekasının bir kişi, cadde, başka bir araba ve gökyüzü arasındaki farkı, bu nesnelerin veya veri noktalarının temel özelliklerini etiketleyerek ve aralarındaki benzerlikleri arayarak söylemesini sağlamak için kullanılır.

Daha önce belirtildiği gibi, etiketleme bir insan tarafından manuel olarak veya bir makine tarafından otomatik olarak yapılabilir. Manuel etiketleme şirket içinde yapılabileceği gibi, kitle kaynaklı olabilir ve bireylere veya şirketlere dış kaynak sağlanabilir. Bununla birlikte, çok miktarda veri göz önüne alındığında, verileri manuel olarak etiketlemek etkisiz olacağından zaman ve para kaybına neden olur. Etiketleme seçenekleri, şirket içi personel kullanımından kitle kaynak kullanımına ve veri etiketleme hizmetlerine kadar uzanmaktadır. İşletmeler, ihtiyaçlarına en uygun yöntemi veya yöntem kombinasyonunu kullanmalıdır.

Bir veri etiketleme yöntemi seçerken dikkate alınması gereken bazı kriterler şunlardır (Rouse, 2019):

- ✓ işletmenin büyüklüğü,
- ✓ etiketleme gerektiren veri setinin boyutu,
- ✓ yapılan iş üzerindeki personel beceri düzeyi,
- ✓ işletmenin mali kısıtlamaları,
- ✓ makine öğrenme modelindeki amacın etiketlenen veriler ile desteklenmesi.

Verileri etiketlemek için tek bir optimal yöntem yoktur. Bir işletme, verilerini yapılandırmak ve etiketlemek için çeşitli yöntemler kullanabilir. Veri etiketleme ile ilgili kitle kaynaklı etiketli veriler, otomatik veri etiketleme, veriye dayalı tahmin teknikleri literatürde kullanılan bazı yöntemlerdir. Kitle kaynaklı etiketli veriler; eğitim verilerinin önemli ölçüde genişletilerek görüntü tanıma için yapay zeka modellerinin ve algoritmalarının geliştirilmesi olarak ifade edilebilir (Gray ve Suri, 2019). Otomatik veri etiketleme; etiketli bir veri seti elde edildikten sonra, yeni etiketlenmemiş verilerin modele sunulabilmesi ve bu etiketlenmemiş veri parçası için olası bir etiketin tahmin edilebilmesi için verilere makine öğrenmesi modellerinin

uygulanmasıdır (Leif, 2013). Veriye dayalı yanlılık; algoritmik karar verme, programcı odaklı yanlılığa ve veriye dayalı yanlılığa tabidir. Yanlılık etiketli verilere dayanan eğitim verileri, makine öğrenmesi algoritmasının meşru olmasına rağmen, tahmine dayalı bir modelde yanlılıklara ve ihmallere neden olur. Belirli bir makine öğrenmesi algoritmasını eğitmek için kullanılan etiketli verilerin, sonuçları saptırmamak için istatistiksel olarak temsili bir örnek olması gerekir (Hu vd., 2019).

Etiketler, verilerde bulunan özellikleri tanımlamak ve belirtmek için makine öğrenmesinde kullanılmaktadır. Model tanıma, sınıflandırma ve regresyonda yüksek performanslı algoritmalar geliştirmek için etiketlemeler için bilgilendirici, ayırt edici ve bağımsız özellikler seçmek çok önemlidir. Doğru şekilde etiketlenmiş veriler, algoritmaları test etmek ve yinelemek için temel doğruluk sağlayabilir. Etiketler, insanlardan belirli bir etiketlenmemiş veri parçası hakkında yargıda bulunmalarını isteyerek elde edilebilir. Etiketli verilerin elde edilmesi, etiketlenmemiş ham verilere göre önemli ölçüde daha pahalıdır. Etiketli veriler, bir veya daha fazla etiketle etiketlenmiş bir örnek grubudur. Etiketleme genellikle bir dizi etiketlenmemiş veriyi alır ve her bir parçasını bilgilendirici etiketlerle güçlendirir. Etiketli bir veri kümesi elde edildikten sonra, yeni etiketlenmemiş verilerin modele sunulabilmesi ve bu etiketlenmemiş veri parçası için olası bir etiketin tahmin edilebilmesi veya tahmin edilebilmesi için verilere makine öğrenmesi modelleri uygulanabilir.

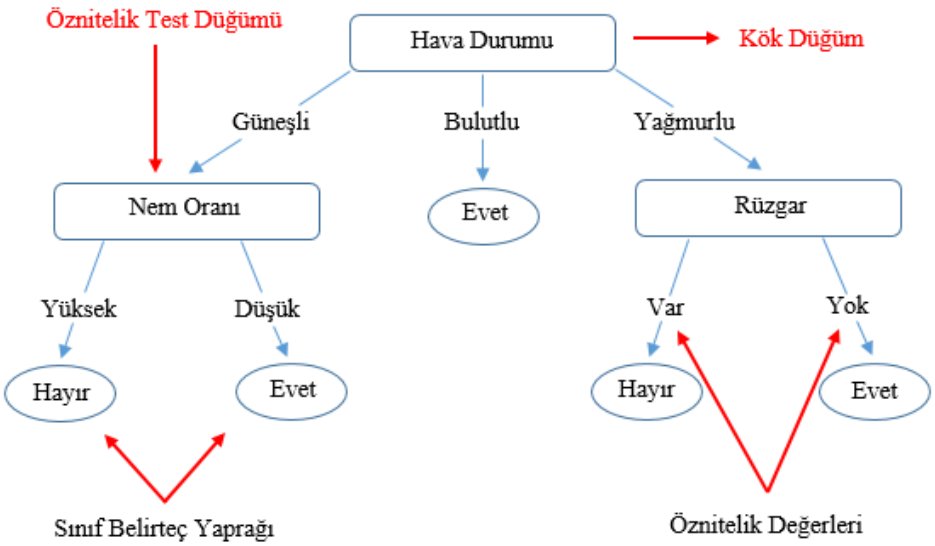
2.6. Makine Öğrenmesi Algoritmaları

Makine öğrenmesi, verilen örneklere veya alan uzmanlarının deneyimlerine göre verileri değerlendirmek için bir bilgisayar programlamasını içermektedir. Bir makine öğrenmesi algoritması, bir giriş vektörü ile bir sonuç arasındaki ilişkiyi belirlemeye çalışır. Makine öğrenmesi algoritması, uygulanan modelin parametresini optimize etmelidir. Son yıllarda, verilerin kolaylıkla toplanabilmesi ve depolanabilmesinin bir sonucu olarak makine öğrenmesi bilgisayar ve veri bilimlerinde giderek daha da önem kazanmaktadır. Elde edilen verilerin kapsamlı bilgiler içermesi, verilerin manuel olarak analiz edilmesini imkansız hale getirmektedir. Bu nedenle, makine öğrenmesi algoritmaları verilerin işlenmesi ve analizinde önemli bir role sahiptir. Makine öğrenmesi algoritmalarının popülerlik kazanmasının bir diğer nedeni ise, hesaplama maliyetlerini düşürmesidir. Son yıllarda, bilgisayar sistemindeki elektronik ve mekanik araçların gelişme göstermesi makine

öğrenmesi algoritmalarında özellikle bütçeleme konusunda hem zaman hem de maliyet açısından verimliliği artırmaktadır. Öğrenme tekniğine karar verilmesi, gerçek dünya sisteminin kullanılması ve girdi özelliklerinin kalitesi makine öğrenmesi algoritmalarından sonuçlar elde edilmesinde önem taşımaktadır. Çalışma kapsamında ele alınan makine öğrenmesi algoritmaları, pazarlama için ayrılacak bütçenin verimliliğinin arttırılması amacıyla satış ve promosyon bilgilerinin analizine yönelik tahminde bulunmaktadır. Bu bağlamda; karar ağacı, rastgele orman, yapay sinir ağları ve gradyan artırma makineleri bu çalışma içerisinde kullanılan makine öğrenmesi algoritmaları olmuştur.

2.6.1. Karar Ağacı (Decision Tree)

Karar ağacı; birçok alanda uygulanabilen tahmine dayalı modelleme yaklaşımlarından biridir. Karar ağacı, veri setini farklı koşullara göre farklı şekillerde bölebilen algoritmik bir yaklaşımla oluşturabilir. Karar ağacı algoritması, denetimli öğrenme algoritmaları kategorisinde yer almaktadır. Kararlar, denetimli algoritmalar kategorisine giren en güçlü algoritmalardır. Diğer denetimli öğrenme algoritmalarının aksine, karar ağacı algoritması sınıflandırma ve regresyon problemlerini çözmek için de kullanılabilir. Karar analizinde, kararları ve karar vermeyi görsel ve açık bir şekilde temsil etmek için bir karar ağacı kullanılabilir. Karar ağacı, belirli bir hedefe ulaşılmasında ve stratejiler üretilmesinde veri madenciliğinde ve makine öğrenmesinde yaygın olarak kullanılan bir araçtır. Karar ağacı, her bir yaprak düğümünün bir sınıf etiketine karşılık geldiği ve özniteliklerin ağacın dahili düğümünde temsil edildiği sorunu çözmek için ağaç gösterimini kullanır. Karar ağacı, önceki verilerden (eğitim verileri) çıkarılan basit karar kurallarını öğrenerek hedef değişkenin sınıfını veya değerini tahmin etmek için kullanılabilecek bir eğitim modeli oluşturmayı amaçlamaktadır. Bir ağacın iki temel varlığı, verinin bölündüğü ve sonuç alınan yerden ayrıldığı karar düğümleridir. Karar ağacı örneğine ait bir gösterime Şekil 2.5 içerisinde yer verilmiştir (Sezer, 2008).



Şekil 2.5. Bir karar ağacı gösterimi örneği.

Bir öğe hakkındaki gözlemlerden (dallarda temsil edilen), öğenin hedef değeri (yapraklarda temsil edilen) ile ilgili sonuçlara gitmek için bir karar ağacı (tahmin modeli olarak) kullanır. Hedef değişkenin ayrı bir değer kümesi alabildiği ağaç modellerine sınıflandırma ağaçları denir; bu ağaç yapılarında, yapraklar sınıf etiketlerini temsil eder ve dallar, bu sınıf etiketlerine götüren özelliklerin birleşimlerini temsil eder. Hedef değişkenin sürekli değerler alabildiği (tipik olarak gerçek sayılar) karar ağaçlarına regresyon ağaçları denir. Karar ağaçları, anlaşılırlıkları ve basitlikleri göz önüne alındığında en popüler makine öğrenmesi algoritmaları arasındadır (Wu vd., 2008). Karar ağacı, denetimli öğrenme için hiyerarşik bir modeldir; bu sayede yerel bölge, daha az sayıda adımda bir dizi yinelemeli bölmelerde tanımlanır. Bir karar ağacı, iç karar düğümleri ve uç (terminal) veya yaprak düğümlerinden oluşur. Her karar düğümü, dalları etiketleyen farklı sonuçlara sahip bir test uygular. Bir girdi verildiğinde, her düğümde bir test uygulanır ve sonuca bağlı olarak dallardan biri alınır. Bu süreç kökte başlar ve bir uç düğümüne ulaşılana kadar yinelemeli olarak tekrarlanır, bu noktada yaprakta yazılan değer çıktıyı oluşturur.

ID3, CART, CHAID, C4.5, C5.0, MARS, QUEST, SLIQ, SPRINT başlıca karar ağacı algoritmalarıdır. Birçok veri madenciliği ve makine öğrenmesi yazılım paketi, bir veya daha fazla karar ağacı algoritmasının uygulanmasına olanak sağlar. Salford Systems CART, IBM SPSS Modeler, RapidMiner, SAS Enterprise Miner, Matlab, R, WEKA, Orange, KNIME, Microsoft SQL Server ve scikit-learn gibi çok sayıda yazılım programı karar ağacı algoritması içerir.

Makine öğrenmesinde karar ağacı algoritmalarının kullanılmasının çeşitli avantajları bulunmaktadır:

- ✓ Anlaması ve yorumlaması basittir. İnsanlar kısa bir açıklamadan sonra karar ağacı modellerini anlayabilirler. Ağaçlar, uzman olmayan kişilerin kolay bir şekilde yorumlayabilmesi için grafik olarak da görüntülenebilir (Gareth vd., 2015).
- ✓ Hem sayısal hem de kategorik verileri işleyebilir (Gareth vd., 2015). Diğer teknikler genellikle yalnızca bir tür değişkeni olan veri setlerini analiz etmede kullanılır. İlk karar ağaçları yalnızca kategorik değişkenleri işleyebiliyorken, C4.5 gibi daha yeni sürümler bu sınırlamaya dahil değildir (Piyonesi ve El-Diraby, 2020a).
- ✓ Çok az veri hazırlığı gerektirir. Diğer teknikler genellikle veri normalleştirme gerektirir. Ağaçlar niteliksel öngörücülerle başa çıkabildiğinden, kukla değişkenler oluşturmaya ihtiyaç duymaz (Gareth vd., 2015).
- ✓ Beyaz kutu veya açık kutu modeli kullanır (Piyonesi ve El-Diraby, 2020a). Karar ağaçları herhangi bir Boole işlevine yaklaşabilir (Mehtaa ve Raghavan, 2002). Bir modelde belirli bir durum gözlemlenebilir ise, koşulun açıklaması Boole mantığı ile kolayca açıklanabilir. Buna karşılık, yapay bir sinir ağı gibi bir kara kutu modelinde sonuçların açıklamasını anlamak genellikle zordur.
- ✓ İstatistiksel testler kullanarak bir modelin doğrulanmasını mümkündür. Bu, modelin güvenilirliğini açıklamayı olanaklı hale getirir.
- ✓ Eğitim verileri veya tahmin kalıntıları hakkında hiçbir varsayımda bulunmayan dağılım, bağımsızlık veya sabit değişken varsayımları gibi istatistiksel olmayan yaklaşımları yoktur.
- ✓ Büyük veri setleri ile iyi performans gösterir. Standart hesaplama kaynakları kullanılarak makul sürede büyük miktardaki verileri analiz edebilir.
- ✓ İnsanın karar verme sürecini diğer yaklaşımlardan daha yakından yansıtır (Piyonesi ve El-Diraby, 2020a). Bu, insan kararlarını ve davranışlarını modellerken faydalı olabilir.
- ✓ Eş doğrusallığa karşı destekleyicidir.

- ✓ Yerleşik özellik seçiminde kullanılır. Buna ilaveten, alakasız özellik daha az kullanılır, böylece sonraki çalıştırmalarda kaldırılabilirler. Bir karar ağacındaki özniteliklerin hiyerarşisi özniteliklerin önemini yansıtır (Provost, 2013). Bu üstteki özelliklerin en bilgilendirici olduğu anlamına gelir (Piyonesi ve El-Diraby, 2020b).

Diğer taraftan, karar ağacı algoritmalarının kullanılmasında çeşitli dezavantajlar veya sınırlamalar bulunmaktadır:

- ✓ Ağaçlar çok sağlam olmayabilir. Eğitim verilerindeki küçük bir değişiklik, ağaçta büyük bir değişikliğe ve dolayısıyla nihai tahminlere neden olabilir (Witten ve Frank, 2016).
- ✓ Sınıf sayısının çok ve öğrenme setinin örneklerinin sayısının az olması durumunda oluşturulan model yeterli olmayabilir (Sezer, 2008).
- ✓ Optimal bir karar ağacı öğrenme probleminin, eniyiliğin çeşitli yönleri altında ve hatta basit kavramlar için NP-tamlığı olarak bilinmektedir (Hyafil ve Rivest, 1976). Sonuç olarak, uygulamada karar ağacı öğrenme algoritmaları, her düğümde yerel olarak en uygun kararların alındığı açgözlü algoritma gibi buluşsal yöntemlere dayanır. Bu tür algoritmalar, küresel olarak en uygun karar ağacını döndürmeyi garanti edemez. Yerel optimalliğin açgözlü etkisini azaltmak için ikili bilgi mesafesi (DID) ağacı gibi bazı yöntemler önerilir (Ben-Gal vd., 2014).
- ✓ Karar ağacı, eğitim verilerinden iyi bir şekilde genelleme yapmayan aşırı karmaşık ağaçlar oluşturabilirler. Bu, ağaçların budamasında da karmaşıklığın olduğu anlamını taşır (Sezer, 2008).

Karar ağaçlarının oluşturulmasında kullanılan algoritma önemli bir role sahiptir. Zira kullanılan algoritma türü oluşturulmuş olan ağacın şeklinde değişiklik gösterebilir. Ağaç yapılarının değişik olması da sınıflandırma için farklı sonuçlar ortaya koyar. Kök düğümünü meydana getiren ilk düğüm farklı olduğunda, en uç düğüme giderken izlenecek olan yolda değişime sebep olacağından sınıflandırmayı da değiştiririr. Kök düğümü ile diğer düğümlerin belirlenmesinde, bir nokta için düğümün dallanmasının veri tabanının kalan kısımlarında da aynı sayıda parçaya bölünüp bölünmediği durumu önemlidir. Yani veri tabanına verilen cevabın değişken

sayısı ile aynı parçaya bölünmesi istenir. Bu bağlamda, istenilen cevap veya sınıfa mümkün olan en kısa şekilde ulaşılması amaçlanmaktadır.

Karar ağaçları, belirli bir veri kümesinin tanımlanmasına, sınıflandırılmasına ve genelleştirilmesine yardımcı olmak için matematiksel ve hesaplama tekniklerinin kombinasyonu olarak da tanımlanabilir. Veriler, $(x, Y) = (x_1, x_2, x_3, \dots, x_k, Y)$ formunun kayıtlarında gelir. Bu ifadedeki bağımlı değişken Y; anlamaya, sınıflandırmaya veya genellemeye çalışılan hedef değişkendir. x vektörü, bu görev için kullanılan x_1, x_2, x_3 gibi özelliklerden oluşur. Her bir öznitelik, eğitim örnekleri üzerinde yapılan sınıflandırmaların karar analizinde istatistiksel testlerin kullanımıyla değerlendirmeye alınır.

Seçilen en iyi öznitelik, karar ağacının kök düğümünün test edilmesi amacıyla kullanılır. Kök düğümündeki dalların sayısı, seçilen özniteliklerin alabilecekleri değer sayısıyla değişim gösterir. Karar ağaçları oluşturmaya yönelik algoritmalar genellikle her adımda veri setini en iyi şekilde bölen bir değişken seçerek yukarıdan aşağıya çalışır (Rokach ve Maimon, 2005). Farklı algoritmalar en iyi özniteliği ölçmek için farklı ölçütler kullanır. Bunlar genellikle veri seti içindeki hedef değişkenin homojenliğini ölçer. Gini safsızlığı, bilgi kazancı ve budama bu ölçütlerden bazılarıdır. Bu ölçütler, her aday veri setine uygulanır ve elde edilen değerler, bölünmenin kalitesinin bir ölçüsünü sağlamak için birleştirilir.

2.6.1.1. Gini safsızlığı (Gini impurity)

Sınıflandırma ağaçları için CART (sınıflandırma ve regresyon ağacı) algoritması tarafından kullanılan Gini safsızlığı, kümeden rastgele seçilen bir öğenin, alt kümedeki etiketlerin dağılımına göre rastgele etiketlendiğinde ne sıklıkla yanlış etiketleneceğinin bir ölçüsüdür. Gini safsızlığı, seçilen i etiketli bir öğenin p_i olasılığının, o öğenin kategorize edilirken olasılığının çarpımının toplanmasıyla Denklem 2.3 ile hesaplanabilir.

$$\sum_{k \neq i} p_k = 1 - p_i \tag{2.3}$$

Düğümdeki tüm durumlar tek bir hedef kategoriye girdiğinde minimuma (sıfır) ulaşır. Gini safsızlığı aynı zamanda bilginin teorik ölçüsü olup, $q = 2$ için deformasyon katsayısına sahip Tsallis Entropisi'ne karşılık gelir. Bu; fizikte denge dışı, kapsamlı olmayan, enerji tüketen ve kuantum sistemlerinde bilgi eksikliği ile

ilişkilendirilir. q'nun limiti 1'e giderken, klasik Boltzmann-Gibbs veya Shannon entropisini kurtarır. Bu anlamda, Gini safsızlığı, karar ağaçları için olağan entropi ölçüsünün bir varyasyonudur.

2.6.1.2. Bilgi kazancı (Information gain)

ID3, C4.5 ve C5.0 ağaç oluşturma algoritmaları tarafından kullanılır. Bilgi kazancı, entropi kavramına ve bilgi teorisindeki bilgi içeriğine dayanır. Bilgi kazancı, ağacı oluşturmanın her adımında hangi özelliğin bölüneceğine karar vermek için kullanılır (Witten ve Frank, 2016). Bilgi kazancında en yüksek ve en iyi dallara ayıran değişken bölünmeye başlanılan değişken olarak seçilir. Bir D veri seti, n adet alt bölüm için Y değişkeni üzerinden bölünürse, X değişkenine ait bilgi kazancı Denklem 2.4 ile elde edilir.

$$Bilgi\ Kazancı\ (D,Y) = E(D) - \sum_{k=1}^{n} p(D_i)\ E(D_i) \qquad (2.4)$$

Bu denklemde yer alan $E(D)$, D veri setinin Y değişkeni üzerinden bölünme öncesine ait entropisi; $p(D_i)$, i alt bölümü için Y değişkeni üzerinden bölünme sonrasına ait olasılığı ve $E(D_i)$, i alt bölümü için Y değişkeni üzerinden bölünme sonrasına ait entropisi olarak ifade edilir. Bilgi kazancının hesaplanmasında ilk olarak alt bölümlere ayrılmadan önce veri setinin entropisi elde edilir. Daha sonraki aşamada, her bir alt bölüme ait entropi değeri bulunur. İki entropi arası fark, hangi değişkende en yüksek çıkarsa o değişken dalları ayırmak için en iyi kriteridir.

2.6.1.3. Budama (Pruning)

Makine öğrenmesinde budama, örnekleri sınıflandırmak için ağacın kritik olmayan ve gereksiz bölümlerini kaldırarak bir karar ağacını basitleştirmek, sıkıştırmak ve optimize etmek için kullanılır (Klimek, 2020). Budama, makine öğrenmesi aracılığıyla oluşturulan ağaçlara aşırı öğrenmeyi (overfitting) önleme girişiminden kaynaklanır. Aşırı öğrenme, bir ağaçta istenmeyen gürültü indüksiyonunu tanımlar (Bramer, 2007). Gürültü ise veri setlerini tahrif eden ve dolayısıyla karar ağaçlarını gereksiz yere genişleten yanlış öznitelik değerlerini veya sınıf üyeliğini tarif eder. Bu bağlamda, ağaçlar budanarak gereksiz alt ağaçlar tekrar kısaltılır.

Karar ağacı algoritmasında ortaya çıkan sorulardan biri, son ağacın optimum boyutudur. Çok büyük bir ağaç, eğitim verisini aşırı öğrenme ve yeni örneklere yetersiz bir şekilde genelleme riski taşır. Küçük bir ağaç, örnek alanıyla ilgili önemli yapısal bilgileri yakalayamayabilir. Bununla birlikte, bir ağaç algoritmasının ne zaman durması gerektiğini söylemek zordur. Çünkü tek bir ekstra düğümün eklenmesinin hatayı önemli ölçüde azaltıp azaltmayacağını söylemek imkansızdır. Bu sorun ufuk etkisi olarak bilinir. Yaygın olan strateji, her düğümde az sayıda örnek içerecek şekilde ağacı büyütmek ve ardından ek bilgi sağlamayan düğümleri kaldırmak için budamayı kullanmaktır (Hastie vd., 2001). Budama, çapraz geçerlilik seti ile ölçülen tahmin doğruluğunu azaltmadan öğrenme ağacının boyutunu küçültmelidir.

2.6.2. Rastgele Orman Algoritması (Random Forest)

Rastgele ormanlar (RO) veya rastgele karar ormanları, eğitim zamanında çok sayıda karar ağacı oluşturarak ve sınıfların modu veya bireysel ağaçların ortalama tahmini olan sınıfı ortaya çıkararak sınıflandırma, regresyon ve diğer görevler için toplu bir öğrenme yöntemidir (Ho, 1995). Rastgele ormanlar, karar ağaçlarının eğitim setlerine aşırı öğrenme alışkanlığını düzeltir (Hastie vd., 2008). Rastgele karar ormanları için ilk algoritma, Ho'nun formülasyonunda Eugene Kleinberg tarafından önerilen sınıflandırmaya "stokastik ayrımcılık" yaklaşımını uygulamanın bir yolu olan rastgele alt uzay yöntemi kullanılarak oluşturulmuştur (Ho, 1998; Kleinberg, 1990). Karar ormanları algoritması, Leo Breiman tarafından geliştirilmiştir. Algoritma, Breiman'ın "torbalama" fikrini ve kontrollü değişkene sahip bir karar ağaçları koleksiyonu oluşturmak için önce Ho tarafından (Ho, 1995) ve daha sonra bağımsız olarak Amit ve Geman tarafından sunulan özelliklerin (Amit ve Geman, 1997) rastgele seçimini birleştirmiştir (Breiman, 2001). Rastgele ormanların gerçek manada ilk tanımını yapan Leo Breiman, rastgele düğüm optimizasyonu ve torbalama ile birlikte CART benzeri bir prosedür kullanarak ilişkisiz ağaçlardan oluşan bir orman inşa etmenin bir yöntemini açıklamıştır. Buna ilaveten, Breiman modern rastgele orman uygulamasının temelini oluşturan, bazıları önceden bilinen ve bazıları yeni olan torba dışı hatayı genelleme hatasının bir tahmini olarak kullanmak ve değişkenin önemini permütasyon yoluyla ölçmek gibi çeşitli bileşenleri bir araya getirmiştir. Ayrıca rastgele ormanlar için elde edilen ilk teorik sonuç ormandaki ağaçların gücüne ve korelasyonlarına bağlı olan genelleme hatasına bağlı olarak gösterilmiştir.

Karar ağaçları, çeşitli makine öğrenmesi görevleri için popüler bir yöntemdir. Özellikle, çok derin büyüyen ağaçlar oldukça düzensiz modeller öğrenme eğilimindedir. Eğitim setlerinden aşırı öğrenirler, yani düşük yanlılığa, ancak çok yüksek varyansa sahiptirler. Rastgele ormanlar, varyansı azaltmak amacıyla aynı eğitim setinin farklı bölümlerinde eğitilmiş birden çok derin karar ağacının ortalamasını almanın bir yoludur (Hastie vd., 2008). Bu, yanlılıkta küçük bir artış ve biraz yorumlanabilirlik kaybı olmasına rağmen, genellikle nihai modeldeki performansı büyük ölçüde artırır. Ormanlar, karar ağacı algoritmalarının bir araya getirilmesi gibidir. Birçok ağacın takım çalışmasını alarak tek bir rastgele ağacın performansını arttırır. Oldukça benzer olmasa da, ormanlar k-kat çapraz geçerlemedeki etkileri verir.

2.6.2.1. Torbalama (Bagging)

Rastgele ormanlar için eğitim algoritması, ağaç öğrenmesinde önyükleme (bootstrap) yığma veya torbalama tekniğini uygular. $Y = y_1, \dots, y_n$ yanıtlı $X = x_1, \dots, x_n$ eğitim seti, tekrar tekrar torbalanarak (B kez), eğitim setinin değiştirilmesiyle rastgele bir örnek seçer ve ağaçları bu örneklere uydurur. $b = 1, \dots, B$ için, n tane eğitim örneğindeki X ve Y değiştirilerek, X_b ve Y_b olarak çağırılır. Daha sonra, X_b ve Y_b üzerinde bir sınıflandırma veya regresyon ağacı f_b eğitilir. Eğitim sonrası, x' üzerindeki tüm bağımsız regresyon ağaçlarından tahminlerin ortalaması alınarak görünmeyen örnekler x' için tahminler yapılabilir. Bu tahmin, Denklem 2.5 ile ifade edilir.

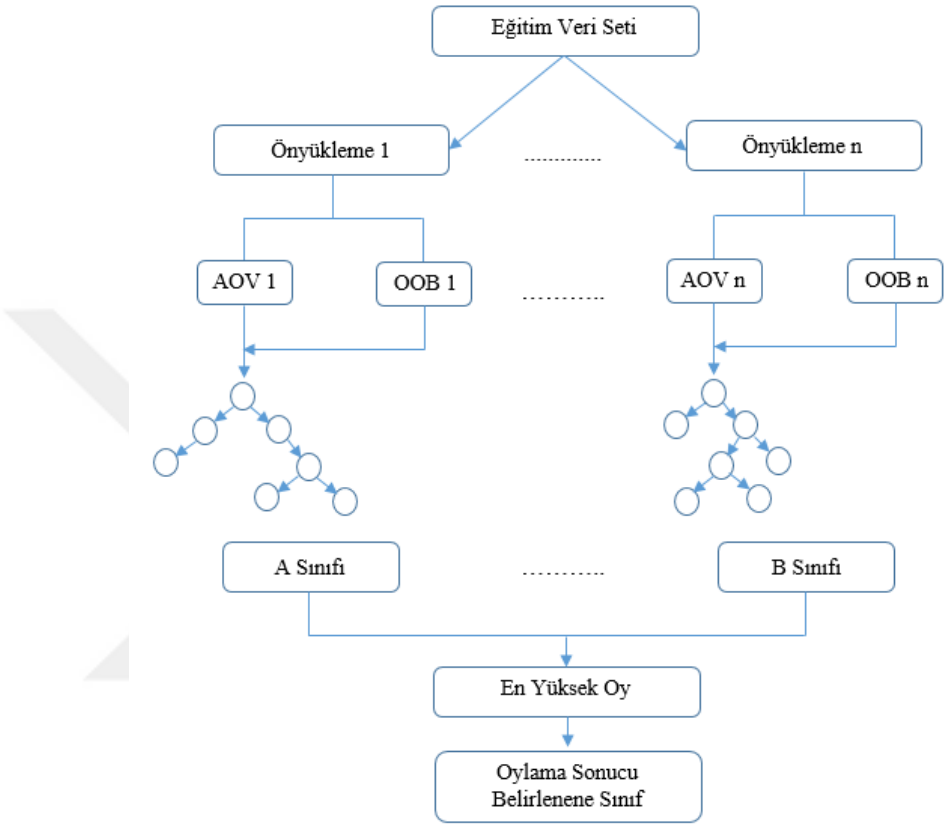
$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{2.5}$$

Önyükleme tekniği, yanlılığı artırmadan modelin varyansını azalttığı için daha iyi model performansına yol açar. Bu, eğitim setindeki tek bir ağacın yanlılığı gürültüye karşı oldukça hassas olsa da, ağaçlar birbirleriyle ilişkili olmadığı sürece birçok ağacın ortalamasının hassas olmadığı anlamına gelir. Tek bir eğitim setinde çok sayıda ağacın eğitilmesi, güçlü bir şekilde ilişkili ağaçları verir. Önyükleme örneklemesi, ağaçlara farklı eğitim setleri göstererek birbirleriyle olan ilişkilerini gidermenin bir yoludur. Önyükleme tekniği, ağaçlar için orijinal torbalama algoritmasını açıklamaktadır. Rastgele ormanlar, bu genel şemadan yalnızca bir şekilde farklılık gösterir. Rastgele ormanlar, öğrenme sürecindeki her aday

bölünmesinde, özelliklerin rastgele bir alt kümesini seçen değiştirilmiş bir ağaç öğrenme algoritması kullanırlar. Bu sürece bazen "özellik torbalama" adı verilir. Bunun yapılmasının nedeni, sıradan bir önyükleme örneğindeki ağaçların ilişkisidir. Bir veya birkaç özellik, yanıt değişkeni (hedef çıktı) için çok güçlü öngörücülerse, bu özellikler birçok B ağacında seçilecek ve ilişkili hale gelmelerine neden olacaktır (Ho, 2002). Genel olarak p özellikli bir sınıflandırma problemi için her bölünmede $\sqrt{p}$ özellikleri kullanılır. Regresyon problemleri için varsayılan olarak minimum düğüm boyutu 5 olan $p/3$ özellik önerilir. Uygulamada bu parametreler için en iyi değerler probleme bağlı olacağından bunlar ayar parametreleri olarak ele alınmalıdır (Hastie, 2008).

Özetle, orman içerisindeki her bir karar ağacı için veri seti üzerinde önyükleme (bootstrap) tekniği uygulanarak farklı örneklem seçimi gerçekleştirilir ve torbalama ile seçilmiş olan özellik kümesi veri setini eğitir (Breiman, 2001). Bu aşamadan sonra, birbirlerinden farklı olan çok sayıdaki bireysel ağacın verdiği kararlar oylamaya tabi tutularak, oylamada en fazla oy alan sınıf topluluk sınıf tahminini gösterir. Rastgele ormanlarda ağaçların oluşturulmasında sınıflandırma ve regresyon algoritmalarının yanı sıra, önyükleme ve torbalamanın birleşimi olan bir yöntem kullanılır (Breiman, 2001). Veri setleri, "eğitim" ve "test" veri setleri olarak iki gruba ayrılır. Eğitim veri seti üzerinden önyükleme tekniğiyle ağaç oluşturacak (inbag/AOV) ve oluşturulmayan (out-of-bag/OOB) veriler şeklinde örneklem seçimi yapılır. Ağaç oluşturacak ve oluşturulmayan veriler eğitim setinden sırasıyla 2/3 ve 1/3 oranında ayrılır. Tüm değişkenlerin içerisinden ağaç oluşturulacak veriler aracılığıyla her düğüm için n adet değişken seçilerek bilgi kazancı veya Gini safsızlığı yardımıyla en iyi ayrılma gerçekleşir. Ağaç oluşturulmayan verilerle oluşturulan model üzerinde tahminler yapılarak hatalar bulunur. Her bir ağaç için yapılan ağaç oluşturmayan verilerin kestirimleri bir araya getirilerek ağacın hatası hesaplanır. Her bir ağaç için ağaç oluşturmayan verilerin model üzerindeki tahminlerin hata oranı ölçüsünde bir ağırlık verilir. Ağaçların hata oranı ve ağırlıkları birbiriyle ters orantılıdır (Ho, 1995; Breiman, 2001). Sınıflandırılan her bir ağaç bireysel olarak oy alır ve sürecin sonunda en yüksek oya sahip olan karar ağacı tarafından yapılan sınıflama kullanılır. Karar ağaçları eğitilmiş olduğu veri gruplarından farklı gruplar ile kıyaslandığında farklı performans gösterir. Bu bağlamda, rastgele orman algoritması pek çok karar ağacını birleştirerek, doğru sınıflandırma oranı ve sınıflandırmanın performansı artırılır.

Rastgele orman algoritmasında ağaçlar, bir düğüm ile başlar ve tüm örnekler aynı sınıfa dahil ise, düğümler uç veya yaprak şeklinde sonlanır ve sınıf etiketi verilir. Eğer örnekler aynı sınıf içerisinde yer almıyorsa, örneklerin sınıflara ayrılmasında en iyi özellikler seçilir. Rastgele orman algoritmasına ait bir gösterime Şekil 2.6 içerisinde yer verilmiştir (Yu vd., 2018).



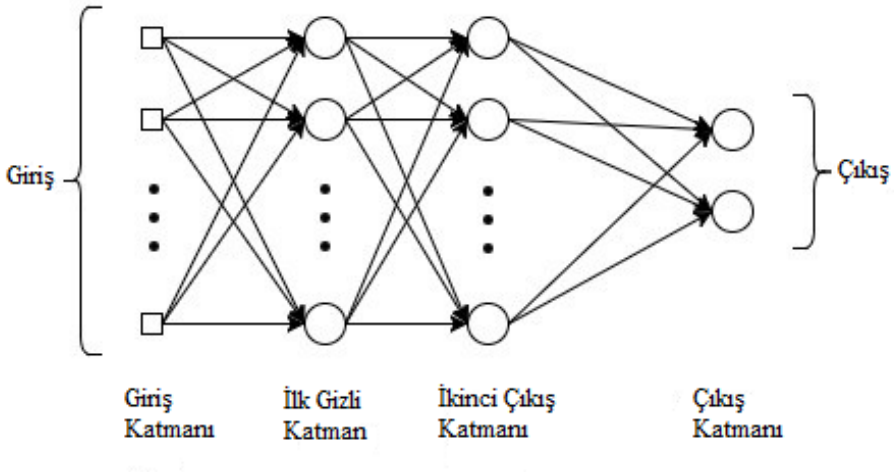
Şekil 2.6. Rastgele orman gösterimi.

2.6.3. Çok Katmanlı Algılayıcılar (Multilayer Perception)

Algılayıcılar yalnızca doğrusal olarak ayrılabilen örnek veri setlerini sınıflandırabilir. Girdi örneklerini doğru kategorilere ayırmak için düz bir çizgi veya düzlem çizilebilirse, girdi örnekleri doğrusal olarak ayrılabilir ve algılayıcı bu probleme çözüm bulabilir. Örnekler doğrusal olarak ayrılabilir değilse, öğrenme hiçbir zaman tüm örneklerin uygun şekilde sınıflandırıldığı bir noktaya ulaşmayacaktır. Bu problemi çözmek için çok atmanlı algılayıcılar (multilayer perception) oluşturulmuştur (Rumelhart vd., 1986).

Çok katmanlı algılayıcılar (MLP), ileri beslemeli yapay sinir ağının (YSA) bir sınıfıdır. Çok katmanlı bir sinir ağı, bir bağlantı modelinde birbirine bağlanmış çok

sayıda birimden (nöron) oluşur. Genellikle çok katmanlı bir sinir ağı; işlenecek bilgileri alan giriş katmanı, işlemin sonuçlarının elde edildiği çıktı katmanı, giriş ve çıkış katmanları arasındaki gizli katman şeklindeki üç düğüm katmandan oluşur. Şekil 2.7'de çok katmalı algılayıcılara ait bir gösterime yer verilmiştir. Çok katmanlı algılayıcılar, özellikle tek bir gizli katmana sahip olduklarında, bazen "vanilya" sinir ağları olarak adlandırılır (Hastie vd., 2009). Giriş düğümleri dışında, her düğüm doğrusal olmayan bir aktivasyon işlevi kullanan bir nörondur. Çok katmanlı algılayıcılar, verilerin girişten çıkışa yalnızca tek yönde gitmesine izin verir. MLP, eğitim için geri yayılım adı verilen denetimli bir öğrenme tekniğini kullanır (Rosenblatt, 1961; Rumelhart vd., 1986). Çoklu katmanları ve doğrusal olmayan aktivasyonu çok katmanlı algılayıcıyı doğrusal bir algılayıcıdan ayırır. Çünkü doğrusal olarak ayrılamayan verileri ayırt edebilir (Cybenko, 1989).



Şekil 2.7. Çok katmanlı algılayıcılara ait gösterim.

Yapay sinir ağı öncelikle giriş-çıkış eşlemesini belirlemek için bir dizi eşleştirilmiş veri üzerinde eğitilir. Nöronlar arasındaki bağlantıların ağırlıkları daha sonra sabitlenir ve ağ, yeni bir veri kümesinin sınıflandırmalarını belirlemek için kullanılır. Sınıflandırma sırasında, giriş nöronlarındaki veri, tüm çıkış nöronlarındaki etkinleştirme değerlerini belirlemek için ağ boyunca tüm yol boyunca yayılır. Her giriş nöronu, ağın dışındaki bazı özellikleri temsil eden bir etkinleştirme değerine sahiptir. Daha sonra her giriş nöronu, bağlı olduğu gizli nöronların her birine aktivasyon değerini gönderir. Bu gizli nöronların her biri kendi aktivasyon değerini hesaplar ve bu veri daha sonra çıkış nöronlarına iletilir. Her alıcı nöronun aktivasyon değeri, basit bir aktivasyon fonksiyonuna göre hesaplanır. Fonksiyon, tüm gönderen nöronların

katkılarını toplar; burada bir nöronun katkısı, gönderen ve alan nöronlar arasındaki bağlantının ağırlığı ile gönderen nöronun etkinleştirme değerinin çarpımı olarak tanımlanır. Bu toplam aktivasyon toplamının 0 ile 1 arasındaki bir değere ayarlanması ve/veya bu toplam için bir eşik seviyesine ulaşılmadıkça aktivasyon değerinin sıfıra ayarlanmasıyla genellikle daha sonra değiştirilir. Genel olarak, gizli katmanın boyutunun doğru bir şekilde belirlenmesi bir sorundur, çünkü nöron sayısının eksik tahmin edilmesi, zayıf yaklaşım ve genelleme yeteneklerine yol açabilirken, aşırı düğümler aşırı öğrenmeye neden olabilir ve nihayetinde küresel optimum aramayı daha zor hale getirebilir (Camargo ve Yoneyama, 2001).

YSA, nöronun giriş ve aktivasyon fonksiyonları, ağ mimarisi ve her giriş bağlantısının ağırlığına bağlı olarak üç temel özelliğe bağlıdır. İlk iki yönün sabit olduğu göz önüne alındığında, YSA'nın davranışı ağırlıkların mevcut değerleri ile tanımlanır. Eğitilecek ağın ağırlıkları başlangıçta rastgele değerlere ayarlanır ve ardından eğitim setinin örnekleri tekrar tekrar ağa maruz bırakılır. Bir örneğin girdisinin değerleri giriş katmanlarına yerleştirilir ve ağın çıktısı bu örnek için istenen çıktıyla karşılaştırılır. Daha sonra, ağdaki tüm ağırlıklar, ağın çıktı değerlerini istenen çıktı değerlerine yaklaştıracak yönde hafifçe ayarlanır. Bir ağın eğitilebileceği birkaç algoritma vardır (Neocleous ve Schizas, 2002). Bununla birlikte, ağırlıkların değerlerini tahmin etmek için en iyi bilinen ve yaygın olarak kullanılan öğrenme algoritması, geri yayılım (backpropagation-BP) algoritmasıdır. Genel olarak, BP algoritması aşağıdaki altı adımı içerir (Kotsiantis, 2007):

1. Sinir ağına bir eğitim örneği sunulur.

2. Ağın çıktısı o örnekten istenen çıktıyla karşılaştırılır. Her çıkış nöronundaki hata hesaplanır.

3. Her bir nöron için, çıktının ne olması gerektiği ve bir ölçeklendirme faktörü hesaplanır, çıktının istenen çıktıya uyması için çıktının ne kadar düşük veya yüksek olması gerektiğini belirlenir. Bu yerel hatadır.

4. Yerel hatayı azaltmak için her bir nöronun ağırlıkları ayarlanır.

5. Daha güçlü ağırlıklarla bağlanan nöronlara daha fazla sorumluluk vererek, önceki seviyedeki nöronlara yerel hata için "kusur" atanır.

6. Yukarıdaki adımların her biri "kusur" hatası olarak kullanarak, önceki seviyedeki nöronlar üzerinde tekrarlanır.

Daha fazla ayrıntıyla, ağırlıkları güncellemenin genel kuralı şudur: $\Delta W_{ji} = \eta \delta_j O_i$. Bu denklemdeki η ifadesi, gradyan iniş aramasında adım boyutunu belirleyen pozitif bir sayıdır (öğrenme hızı olarak adlandırılır). Büyük bir değer, geri yayılımın hedef ağırlık konfigürasyonuna daha hızlı hareket etmesini sağlar, ancak aynı zamanda bu hedefe ulaşma şansını da artırır. O_i ifadesi, i nöronu tarafından hesaplanan çıktıdır. Çıktı nöronları için $\delta_j = O_j(1 - O_j)(T_j - O_j)$, burada T_j ifadesi j nöron için istenen çıktıdır. İç (gizli) nöronlar için $\delta_j = O_j(1 - O_j) \sum_k \delta_k W_{kj}$. Geri yayılım algoritması, iyi bir ağırlık konfigürasyonuna ulaşmadan önce bir dizi ağırlık modifikasyonu gerçekleştirmelidir. n eğitim örnekleri ve W ağırlıkları için, öğrenme sürecindeki her yineleme $O(nW)$ süresi alır; ancak en kötü durumda, yinelemelerin sayısı girişlerin sayısına göre üssel olabilir. Bu nedenle, sinir ağları, eğitimin ne zaman biteceğini kontrol etmek için bir dizi farklı durdurma kuralı kullanır.

İleri beslemeli sinir ağları genellikle orijinal geri yayılım algoritması veya bazı varyantlar tarafından eğitilir. Bir sinir ağı eğitildikten sonra tahminlerde bulunmak için kullanılabilir. Modelin görünmeyen veriler üzerindeki becerisini tahmin etmek için test veya doğrulama verileri üzerinde tahminler yapılabilir. Ayrıca onu operasyonel olarak dağıtabilir ve sürekli tahminler yapmak için kullanılabilir. Ağ topolojisi ve son ağırlık seti, modelden kaydedilmesi gereken tek şeydir. Tahminler, ağa girdi sağlayarak ve tahmin olarak kullanabileceğiniz bir çıktı oluşturmasına izin veren bir ileri-geçiş gerçekleştirerek yapılır (Brownlee, 2016). MLP'lerin en büyük problemi, çoğu uygulama için çok yavaş olmalarıdır. Eğitim oranını hızlandırmaya yönelik yaklaşımlardan biri, optimum başlangıç ağırlıklarını tahmin etmektir (Yam ve Chow, 2001). Çok katmanlı ileri beslemeli YSA'ları eğitmek için başka bir yöntem, uygun topolojiyi otomatik olarak türeten ve dolayısıyla aşırı öğrenme ile ilgili sorunları da ortadan kaldıran ağırlık eleme algoritmasıdır (Weigend vd., 1991). Sinir ağlarının ağırlıklarını eğitmek (Siddique ve Tokhi, 2001) ve sinir ağlarının mimarisini bulmak için genetik algoritmalar kullanılmıştır (Yen ve Lu, 2000). Sinir ağlarını eğitmeye çalışan Bayes yöntemleri de var. Son zamanlarda, eğitim ilerledikçe ağların mimarisini değiştirerek YSA eğitim algoritmalarını iyileştirmeye çalışan bir dizi başka teknik ortaya çıkmıştır. Bu teknikler, gereksiz düğümleri veya ağırlıkları budamayı

(Castellano vd., 1997) ve gerektiğinde fazladan düğümlerin eklendiği yapıcı algoritmaları içerir (Parekh vd., 2000). MLP'ler, problemleri stokastik olarak çözme yetenekleri için araştırmalar için faydalılıdır ve bu da genellikle uygunluk tahmini gibi son derece karmaşık problemler için yaklaşık çözümlere izin verir. MLP'ler, Cybenko teoreminde gösterildiği gibi evrensel fonksiyon yaklaşımlarıdır, bu nedenle regresyon analizi ile matematiksel modeller oluşturmak için kullanılabilirler (Cybenko, 1989). Sınıflandırma, yanıt değişkeni kategorik olduğunda belirli bir regresyon durumu olduğundan, MLP'ler iyi sınıflandırıcı algoritmalar yapar.

2.6.4. Gradyan Artırma Makineleri (Gradient Boosted Machine)

Gradyan artırma makineleri (GBM), tipik olarak karar ağaçları gibi zayıf tahmin modelleri topluluğu şeklinde bir tahmin modeli üreten, regresyon ve sınıflandırma problemleri için bir makine öğrenme tekniğidir. Modeli, diğer hızlandırma yöntemlerinin yaptığı gibi aşama bazında oluşturur ve keyfi bir türevlenebilir kayıp fonksiyonunun optimizasyonuna izin vererek genelleştirir.

Gradyan artırma fikri, Leo Breiman'ın, artırmanın uygun bir maliyet fonksiyonu üzerinde bir optimizasyon algoritması olarak yorumlanabileceği gözleminden kaynaklanmıştır (Breiman, 1997). Açık regresyon gradyan yükseltme algoritmaları daha sonra Jerome H. Friedman tarafından Llew Mason ve arkadaşlarının daha genel fonksiyonel gradyan artırma perspektifiyle eş zamanlı olarak geliştirildi (Friedman, 1999; Mason vd., 1999a). Yine Mason ve arkadaşları yinelemeli fonksiyonel gradyan iniş algoritmaları olarak artırma algoritmaları görüşünü tanıttı (Mason vd., 1999a; 1999b). Yani, negatif gradyan yönünü işaret eden bir işlevi yinelemeli olarak seçerek (zayıf hipotez) bir maliyet işlevini işlev alanı üzerinden optimize eden algoritmalardır. Artırmanın bu işlevsel gradyan görünümü, regresyon ve sınıflandırmanın ötesinde makine öğrenmesinin ve istatistiklerin birçok alanında artırma algoritmalarının geliştirilmesine yol açmıştır.

Rastgele orman algoritmaları gibi, gradyan artırma makineleri (GBM) de sınıflandırma ve regresyon gibi denetimli makine öğrenmesi görevlerini gerçekleştirmek için kullanılan başka bir tekniktir. Rastgele ormanlara benzer şekilde, gradyan artırma toplu bir öğrenicidir. Bu, gradyan artırma algoritmalarının bireysel modellerden oluşan bir koleksiyona dayalı nihai bir model oluşturacağı anlamına gelir. Çünkü bireysel modellerin öngörü gücü zayıftır ve aşırı öğrenmeye eğilimlidir, ancak

bu tür birçok zayıf modeli bir grupta birleştirmek, genel olarak çok daha gelişmiş bir sonuca yol açacaktır. Gradyan artırmada kullanılan en yaygın zayıf model türü karar ağaçlarıdır.

Gradyan artırma, "zayıf öğrenenler" denilen bireylerden yinelemeli bir şekilde modeller oluşturur. Gradyan artırmada, bireysel modeller tamamen rastgele veri ve özellik alt kümeleri üzerine değil, yanlış tahminlere ve yüksek hatalara sahip örneklere daha fazla ağırlık vererek sırayla oluşturulmaktadır. Bunun arkasındaki genel fikir, doğru tahmin edilmesi zor olan modelin geçmiş hatalardan ders alması için öğrenme sırasında odaklanılmasıdır. Her bir topluluk eğitim setinin bir alt kümesi üzerinde eğitildiğinde, modelin genelleştirilebilirliğinin geliştirilmesine yardımcı olunmasına stokastik gradyan artırma adı verilir.

Gradyan, sinir ağlarının ağırlıkları optimize etmek için gradyan inişini kullanmasına benzer şekilde, kayıp bir fonksiyonu en aza indirmek için kullanılır. Her bir eğitim döngüsünde, zayıf öğrenici oluşturulur ve öngörüler beklenilen doğru sonuçla karşılaştırılır. Tahmin ve gerçek arasındaki açıklık, modelin hata oranını temsil eder. Bu hatalar artık gradyanı hesaplamak için kullanılabilir. Gradyan hayali değildir, temelde kayıp fonksiyonun kısmi türevidir. Gradyan, bir sonraki eğitim döngüsünde "gradyan inişi" hatayı azaltmak için (maksimum olarak) model parametrelerinin değiştirileceği yönü bulmak için kullanılabilir.

Sinir ağlarında, gradyan inişi kayıp fonksiyonunun minimumunu aramak için kullanılır, yani tek bir modelde tahmin hatasının en düşük olduğu model parametrelerini (örneğin ağırlıklar) öğrenmede kullanılır. Gradyan artırmada birden fazla modelin tahminleri birleştirilir. Bu nedenle, doğrudan model parametreleri değil de, artırılmış model tahminleri optimize edilir. Bu bağlamda, gradyanlar bir sonraki ağacı da bu değerlere uydurarak eğitim sürecine eklenir.

Gradyan inişi uygulandığından, gradyan artırma modelleri de tıpkı sinir ağlarında olduğu gibi öğrenme hızı (gradyanın indiği "adım boyutu"), küçülme (öğrenme hızının azaltılması) ve hiperparametreler olarak kayıp fonksiyonu bulunur. Gradyan artırmanın yineleme sayısı (yani bir araya getirilecek ağaç sayısı), her bir yapraktaki gözlem sayısı, ağaç karmaşıklığı ve derinliği, örneklerin oranı ve eğitim alınacak özelliklerin oranı gibi hiperparametreleri rastgele ormanlara benzer.

2.7. Model Performansının Değerlendirilmesi

Bir makine öğrenmesi algoritması verilere uygulandıktan sonra, bilgisayar programı sonuçlarından hangi modellerin uygun olduğu ve hangilerinin hata içerdiği bilinmelidir. Eldeki veri setine bağlı olarak, farklı makine öğrenmesi algoritmalarını değerlendirmek ve bu algoritmaların verimliliğini ölçmek için farklı performans ölçütleri kullanılmaktadır. Yapılan çalışma model performansının değerlendirilmesinde tabakalı örnekleme yöntemi kullanılmıştır.

Bir modelin performansının ölçümünde kullanılan algoritmanın yanı sıra, veri setinin boyutu, sınıflandırmada ortaya çıkan hatalar veya sınıf dağılımları önemli bir role sahiptir (Zhang, 2014). Tabakalı örnekleme, araştırmacılar tarafından farklı alt gruplardan veya katmanlardan sonuçlar çıkarmaya çalışırken kullanılan yaygın bir örnekleme tekniğidir. Katmanlar veya alt gruplar farklı olmalı ve veriler örtüşmemelidir. Tabakalı örneklemeyi kullanırken, basit olasılık örneklemesi kullanmalıdır. Nüfus, yaş, cinsiyet, ülke, iş profili, eğitim düzeyi gibi çeşitli alt gruplara ayrılır. İki grup arasındaki mevcut ilişkinin anlaşılması istediğinildiğinde tabakalı örnekleme yöntemi kullanılır. Yapılan çalışmada tabakalı örnekleme yöntemi, "eğitim" ve "test" veri setleri olarak iki gruba ayrılmıştır. Bu veri setleri, sırasıyla sınıflandırma modelinin eğitilmesinde (uygun parametre değerlerinin belirlenmesinde) ve genel performansın ölçülmesinde (test edilmesinde) kullanılmıştır.

Sınıflandırma performansının değerlendirilmesinde ve sınıflayıcı modellemesinin yönlendirilmesinde değerlendirme yöntemi önemli bir etkendir. Sınıflandırma süreci eğitim, geçerlilik ve test aşamaları olmak üzere üç temel aşamadan oluşur. Girdiler ile modelin eğitilmesi eğitim aşamasıdır. Bu aşamada kullanılan veriler, eğitim verileridir. Yine model parametrelerinin ayarlanması bu aşamada gerçekleşir. Eğitim hatası ise, eğitilen model için eğitim modelinin ne ölçüde uygun olduğunu tespit eder. Bu hata, her zaman için test ve doğrulama hatalarından daha küçük çıkar. Bunun nedeni, eğitilen modelin eğitim aşaması sürecinde yararlanılan aynı verilere uymasıdır. Makine öğrenmesi algoritmaları, test aşamasında bilinmeyen verileri eğitim verilerinden öğrenerek sınıf etiketleri için tahminler yapar. Sınıf etiketlerinin ya da çıktılarının bilinmemesinden dolayı, testteki ya da örneğin dışındaki hatalar tahmin edilememektedir. Bu nedenle, eğitilen modelin

performansının değerlendirilmesinde test aşaması kullanılmaktadır. Bu aşamada, model hiperparametrelerinin ayarlanmasında eğitilen modelin objektif bir şekilde değerlendirme yapması test verileri aracılığıyla sağlanır.

Sınıflandırma modeli, verilen girdi verileri için sınıf etiketlerini tahmin eder. İkili ve çok sınıflı sınıflandırmada, sırasıyla iki çıktı sınıfı ve ikiden fazla çıktı sınıfı vardır. Pozitif ve negatif sınıflar P ve N şeklinde sınıflanır. Bilinmeyen örneklere gerçek sınıf tahminleri eğitilen sınıflandırma modeli sayesinde yapılır. Model, sürekli ya da ayrık çıktı üretebilir. Sürekli çıktılar, örneklerin sınıf üyeliğiyle ilgili olasılığı tahmin ederken; ayrık çıktılar, bilinmeyen örneklerin tahmin edilen sınıf etiketleri olarak ifade edilir.

Sınıflandırma modelinin performansını ölçmek için çeşitli yöntemler kullanılabilir. Log-kaybı, hassas geri çağırma ve karmaşıklık matrisi bu yöntemlerden bazılarıdır. Log-kaybı, sınıflandırıcı modelin olasılıksal güvenini hesaplayan doğruluğun yumuşak bir ölçümüdür. Yani veri noktasının belirli bir sınıfa ait olma olasılığının ne kadar güvenilir olduğunu gösteren sayısal bir olasılıktır. Hassas geri çağırma, öncelikli olarak arama motorları tarafından kullanılan algoritmaları sıralamak için kullanılır. Bu çalışma kapsamında, sınıflandırma modellerinin performanslarının ölçümünde karmaşıklık matrisi yöntemi ele alınmıştır. Nitekim, karmaşıklık matrisi sınıflandırılmış olan verilerin gerçek ve tahmin sınıfını gösterir. Sınıflandırmadan elde edilen sonuçlar matris satırlarında bulunurken, gerçek değerler matris sütunlarında bulunmaktadır. Tablo 2.1 içerisinde ikili sınıflandırma durumunun karmaşıklık matrisine ait bir örneğe yer verilmiştir (Witten ve Frank, 2016).

Tablo 2.1. İkili sınıflandırma durumunun karmaşıklık matrisine ait bir örnek.

Karmaşıklık Matrisi		**Verinin Tahmin Edilen Sınıfı**	
		Sınıf = Evet	**Sınıf = Hayır**
Verinin Gerçek Sınıfı	**Sınıf = Evet**	**Doğru Pozitif (DP)**	**Yanlış Negatif (YN)**
	Sınıf = Hayır	**Yanlış Pozitif (YP)**	**Doğru Negatif (DN)**

Makine öğrenmesi algoritmalarının performansını göstermek için en iyi yaklaşımlardan biri, doğru pozitif, yanlış pozitif, doğru negatif ve yanlış negatif tahminleri birbirinden ayıran, aynı zamanda olasılık tablosu olarak da adlandırılan karmaşıklık matrisidir (Powers, 2007). Tablo 2.1'e göre, yeşil renkte gösterilen doğru pozitif ve negatif, doğru şekilde tahmin edilen gözlemlerdir. Kırmızı renkte gösterilen yanlış pozitif ve negatifin en aza indirilmesi istenmektedir. Bu terimleri tek tek ele alalım ve tam olarak ne anlama geldiğini ifade edelim.

Doğru pozitif (DP), doğru tahmin edilen pozitif değerlerdir, bu da verinin gerçek sınıf değerinin evet olduğu ve tahmin edilen sınıf değerinin evet olduğu anlamını taşır. *Doğru Negatifler (DN),* doğru tahmin edilen negatif değerlerdir, bu da verinin gerçek sınıf değerinin hayır olduğu ve tahmin edilen sınıf değerinin hayır olduğu anlamını taşır. Hem doğru pozitif hem de doğru negatifte verilerin gerçek sınıfı ile tahmin edilen sınıfı aynı şeyi söyler. *Yanlış pozitif (YP)*, gerçek sınıf hayır iken ve tahmin edilen sınıf evet olduğunda ortaya çıkar. *Yanlış negatif (YN),* gerçek sınıf evet iken, ancak tahmin edilen sınıf hayır olduğunda ortaya çıkar. Yani, yanlış pozitif ve yanlış negatif gerçek sınıf ile tahmin edilen sınıfın birbirleriyle çeliştiği durumlarda gözlemlenir.

Çok sayıdaki sınıflandırma ölçütünün hesaplanmasında yararlanılan karmaşıklık matrisinin bu dört parametresini anladıktan sonra, tez kapsamında model performansının ölçümünde hesaplanan bazı ölçüler aşağıda yer almaktadır:

- ✓ Doğruluk (accuracy),
- ✓ Duyarlılık (recall or sensitivity)
- ✓ Kesinlik (precision),
- ✓ Seçicilik (specificity)
- ✓ F-Ölçütü (F-Measure),
- ✓ ROC Eğrisinin Altındaki Alan (AUC).

2.7.1. Doğruluk

Sınıflandırma problemlerinde doğruluk, yapılan her türlü tahmin üzerinden model tarafından yapılan doğru tahminlerin sayısıdır (Witten ve Frank, 2016). Doğruluk, verilerdeki hedef değişken sınıfları neredeyse dengeli olduğunda iyi bir ölçüdür. Doğruluğun oranı, 0 ile 1 arası bir değerde değişim göstermektedir. Bu bağlamda, yüksek doğruluğa sahip olan bir modelin başarısı için doğruluk oranının 1

değerine yakın olmasına bağlıdır. Doğruluk oranının hesaplanmasında kullanılan formül, Denklem 2.6 ile aşağıda ifade edilir:

$$Doğruluk = \frac{Doğru\ tahmin\ sayısı}{Yapılan\ toplam\ tahmin\ sayısı} = \frac{DP+DN}{DP+YP+DN+YN} \quad (2.6)$$

Denklem 2.6'dan da görüleceği üzere, doğruluk sınıflandırıcının genel veri noktalarına kıyasla doğru tahmininin ölçüsüdür. Basitçe ifade etmek gerekirse, sınıflandırıcılar tarafından yapılan doğru tahmin birimlerinin ve toplam tahmin sayısının oranıdır. Yanlış pozitif ve yanlış negatiflerin değerlerinin neredeyse aynı olduğu simetrik veri setleri için doğruluk büyük bir ölçüdür. Bu nedenle, model performansını değerlendirmek için diğer parametrelere bakmak gerekir.

2.7.2. Duyarlılık

Sınıflandırma performansı sonuçları için geri çağırma olarak da adlandırılan duyarlılık, doğru bir şekilde pozitif tahmin edilen doğru pozitif değerlerin oranıdır. Bir başka deyişle, doğru pozitif sonuçların sayısının, sınıflandırıcı tarafından tahmin edilen pozitif sonuçların sayısına bölünmesiyle elde edilen orandır (Han vd., 2012). Denklem 2.7 aracılığıyla kolaylıkla hesaplanabilir:

$$Duyarlılık = \frac{DP}{DP+YN} \quad (2.7)$$

Duyarlılık oranı, 0 ile 1 arası bir değerde değişim göstermektedir ve yüksek değere sahip bir duyarlılık değeri sınıflandırma açısından yüksek bir performansa sahiptir.

2.7.3. Kesinlik

Pozitif tahmin değeri olarak da bilinen kesinlik, gerçek sınıflandırılmış tüm örnekler arasındaki doğru sınıflandırılmış örneklerin bağıl miktarı şeklinde ifade edilir (Witten ve Frank, 2016). Yani doğru pozitif değerlerin toplam tahmin edilen pozitif değerlere oranıdır. Yüksek kesinlik, düşük yanlış pozitif oranıyla ilgilidir. Denklem 2.8 ile kesinlik değeri elde edilir:

$$Kesinlik = \frac{DP}{DP+YP} \quad (2.8)$$

Kesinlik oranı, 0 ile 1 arası bir değerde değişim göstermektedir ve yüksek değere sahip bir kesinlik değeri sınıflandırma açısından yüksek bir performansa sahiptir.

2.7.4. Seçicilik

Seçicilik, doğru negatif sonuçların sayısının, sınıflandırıcı tarafından tahmin edilen negatif sonuçların sayısına bölünmesiyle elde edilen orandır (Han vd., 2012). Denklem 2.9 ile seçicilik değeri bulunur:

$$Seçicilik = \frac{DN}{DN+YP} \tag{2.9}$$

Seçicilik oranı, 0 ile 1 arası bir değerde değişim göstermektedir ve yüksek değere sahip bir seçicilik değeri sınıflandırma açısından yüksek bir performansa sahiptir.

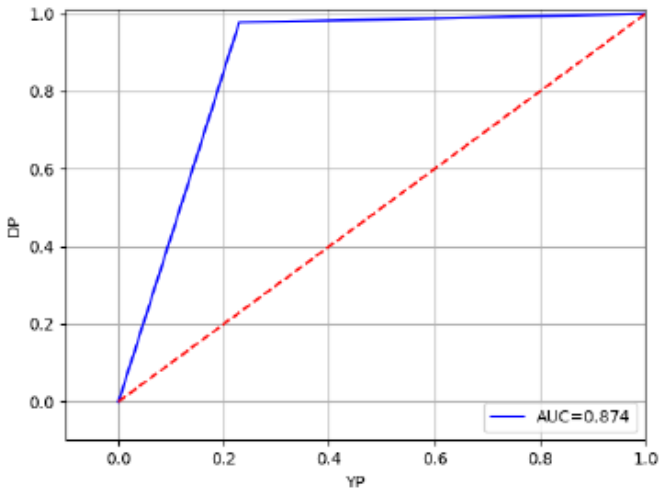
2.7.5. F-Ölçütü

Kesinlik ve duyarlılığın harmonik ortalamasıdır. F-Ölçütü hem yanlış pozitif hem de yanlış negatif değerleri hesaplar. Sezgisel olarak doğruluk ölçütü gibi anlaşılması zor olmasına karşın, eşit olmayan bir sınıf dağılımının olduğu durumlarda F-Ölçütü genellikle doğruluktan daha kullanışlıdır. Örneğin; doğruluk değeri için yanlış pozitif ve yanlış negatif benzer değerlere sahipse en iyi sonucu verir. Yanlış pozitif ve yanlış negatif değerleri birbirinden çok farklıysa, hem kesinliğe hem de duyarlılığa bakmak daha yararlıdır. Bu bağlamda, F-Skoru şeklinde de bilinen F-Ölçütü kesinlik ve duyarlılık model performans değerlendirme ölçütlerinin hesaplanmasında tek başına yetersiz olabileceği durumlar gözetilerek hesaplanır (Powers, 2007). Yine F-Ölçütü, 0 ile 1 arası bir değerde değişim göstermektedir ve yüksek değere sahip bir F-Ölçütü değeri sınıflandırma açısından yüksek bir performansa sahiptir. F-Ölçütü hesabında kullanılan formül, Denklem 2.10 ile aşağıda sunulmuştur.

$$F\text{–Ölçütü} = 2\,x\,\frac{Duyarlılık\ x\ Kesinlik}{Duyarlılık+Kesinlik} = \frac{\frac{DP}{DP+YN}x\frac{DP}{DP+YP}}{\frac{DP}{DP+YN}+\frac{DP}{DP+YP}} \tag{2.10}$$

2.7.6. ROC Eğrisinin Altındaki Alan (AUC)

Herhangi bir sınıflandırma modelinin performansını kontrol etmek için kullanılan en önemli değerlendirme ölçütlerinden biridir. İngilizcede "Receiver Operating Characteristics" ifadesinin baş harflerinin kısaltması olan ROC, işlem karakteristik eğrisi olarak bilinir. ROC eğrisi, x ekseni yanlış pozitif oranından ve y ekseni doğru pozitif oranından oluşan iki boyuta sahip bir grafiktir. ROC analizi, tanısal karar vermede maliyet veya fayda analiziyle doğrudan ve doğal bir şekilde ilişkilidir. ROC analizi; tıpta, radyolojide, biyometride, doğal tehlikelerin tahmin edilmesinde, meteorolojide, model performans değerlendirmesinde ve diğer alanlarda uzun yıllardan beri kullanılmaktadır. Makine öğrenmesi ve veri madenciliği araştırmalarında son yıllarda önemi daha fazla artmıştır (Peres vd., 2015). ROC eğrisinin altındaki alan Area Under Curve ifadesinin kısaltması olarak AUC şeklinde adlandırılır. Kullanılan modellerden hangisinin sınıfları en iyi tahmin ettiğini belirlemek için sınıflandırma analizinde AUC kullanılır. AUC uygulamasının bir örneği de ROC eğrileridir. AUC değeri, 0 ile 1 arası bir değerde değişim göstermektedir ve yüksek değere sahip bir AUC değeri sınıflandırma açısından yüksek bir performansa sahiptir. AUC, ölçekle değişmez. Kesin değerlerinden ziyade tahminlerin ne kadar iyi sıralandığını ölçer. Yine AUC değeri sınıflandırma eşiği ile değişmez. Hangi sınıflandırma eşiğinin seçildiğine bakılmaksızın modelin tahminlerinin kalitesini ölçer. Şekil 2.8 içerisinde AUC değerlerine ait bir görsel sunulmuştur (Yu vd., 2018).



Şekil 2.8. ROC eğrisine ait bir gösterim.

3. MATERYAL VE YÖNTEM

Bu bölümde pazarlama için ayrılacak bütçenin verimliliğinin arttırılmasına yönelik sınıflandırma modellerinin analizinde kullanılan yazılım programı, verilerin elde edilmesinde yararlanılan veri toplama araçları ve veri seti ile ilgili bilgiler sunulmuştur.

3.1. KNIME Analytics Platform

Simülasyonlar sırasında genellikle büyük hacimli veriler oluşturulur ve modüler veri analizi ortamlarına duyulan ihtiyaç son yıllarda önemli ölçüde artmıştır. Etrafındaki çok çeşitli veri analizi yöntemlerinden yararlanmak için, böyle bir ortamın kullanımı kolay ve sezgisel olması, analizde hızlı ve etkileşimli değişikliklere izin vermesi ve kullanıcının sonuçları görsel olarak keşfetmesini sağlaması önemlidir. Bu zorlukların üstesinden gelmek için bir veri hattı oluşturma ortamı uygun bir modeldir. Kullanıcının standartlaştırılmış yapı taşlarından analiz akışını görsel olarak birleştirmesine ve uyarlamasına olanak tanır ve aynı zamanda yapılanları belgelemek için sezgisel ve grafiksel bir yol sunar.

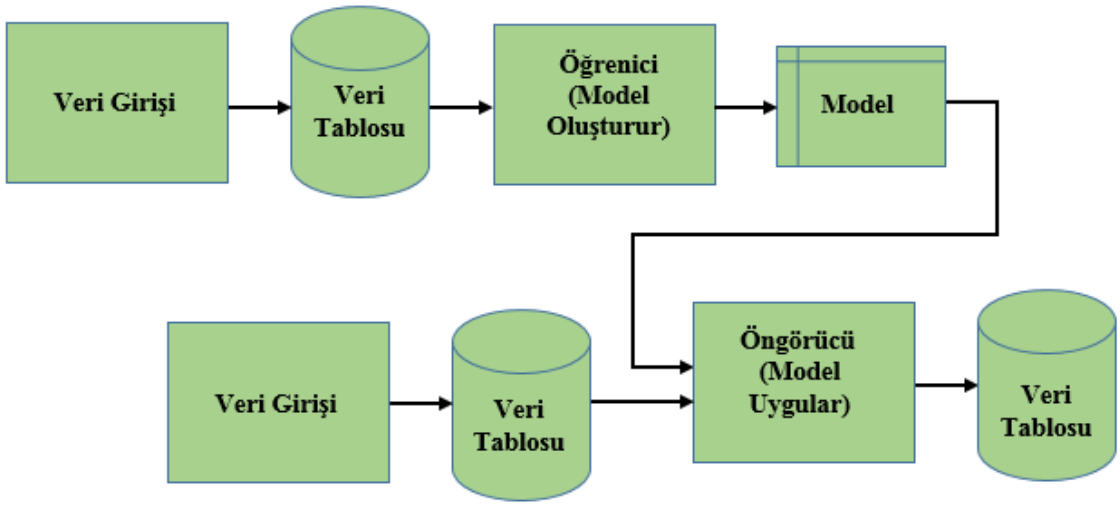
Konstanz Information Miner (KNIME) Analytics Platform, bir veri hattının kolay görsel montajını ve etkileşimli yürütülmesini sağlayan modüler bir ortamdır. Yeni algoritmaların, veri işleme veya görselleştirme yöntemlerinin yeni modüller veya düğümleri (node) olarak kolay entegrasyonunu sağlayan bir öğretim, araştırma ve işbirliği platformu olarak tasarlanmıştır. KNIME mimarisi üç ana prensip göz önünde bulundurularak oluşturulmuştur (Berthold vd., 2006):

Görsel, Etkileşimli Çerçeve: Veri akışları, çeşitli işleme birimlerinden basit bir sürükle ve bırak yöntemiyle birleştirilmelidir. Özelleştirilmiş uygulamalar, bireysel veri hatları aracılığıyla modellenebilir.

Modülerlik: İşlem birimleri ve veri taşıyıcıları, hesaplamanın kolay dağıtımını sağlamak ve farklı algoritmaların bağımsız olarak geliştirilmesine olanak sağlamak için birbirine bağlı olmamalıdır. Veri türleri kapsüllenir, yani hiçbir tür önceden tanımlanmaz, türe özgü işleyiciler ve karşılaştırıcılar getirilerek kolayca yeni türler eklenebilir. Yeni türler, mevcut türlerle uyumlu olarak tanımlanabilir.

Kolay Genişletilebilirlik: Yeni işlem düğümleri veya görünümleri eklemek ve bunları karmaşık yükleme / kaldırma prosedürlerine gerek kalmadan basit bir aç kapa prensibiyle sınıflandırmak kolay olmalıdır.

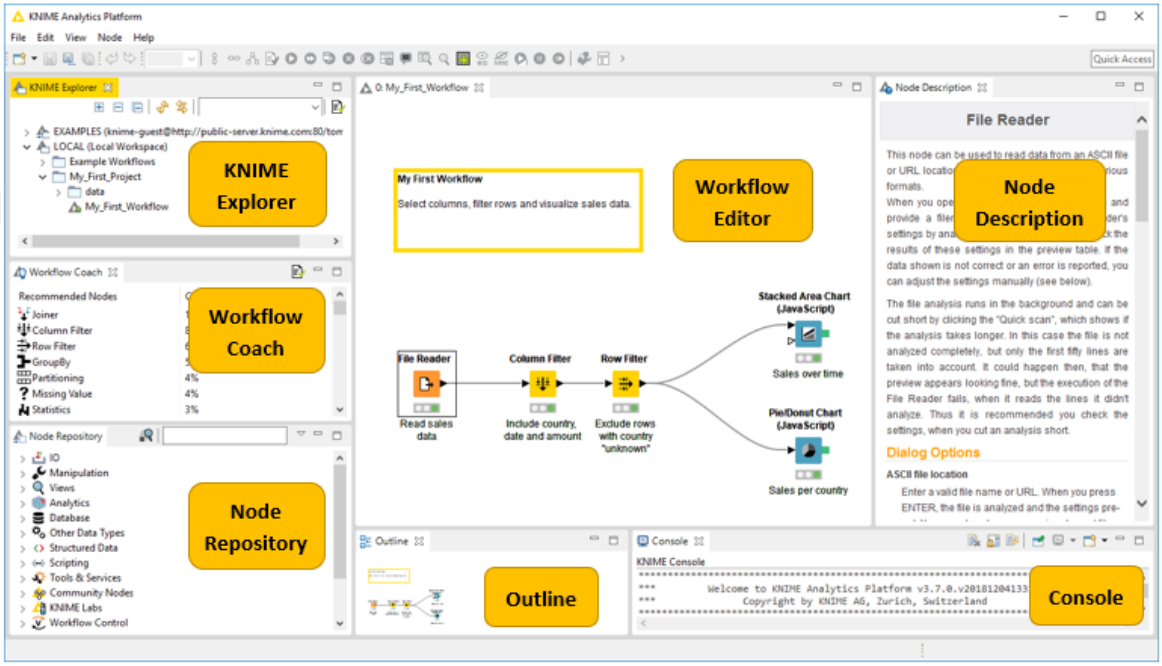
Bir veri analizi süreci, verileri veya modelleri taşıyan çizgiler ile birbirine bağlanan düğümlerden oluşan bir veri hattından meydana gelir. Her düğüm, gelen verileri veya modelleri işleyerek, bunların çıktıları üzerinde sonuçlar üretir. Şekil 3.1'de veri analiz süreci şematik olarak gösterilmiştir. İşlem türü; filtreleme veya birleştirme gibi basit veri işlemlerinden, ortalama hesaplamaları, standart sapma veya doğrusal regresyon katsayıları gibi daha karmaşık istatistiksel işlevlere, yoğun hesaplama kullanan veri modelleme operatörlerine (kümeleme, karar ağaçları, sinir ağları) kadar değişir. Buna ilaveten, modelleme düğümlerinin çoğu, eşlik eden görünümler aracılığıyla sonuçlarını etkileşimli olarak keşfetmeye izin verir.



Şekil 3.1. KNIME iş akışında veri ve modellerin şematik gösterimi.

KNIME Analytics Platform, verilerin oluşturulmasında ve işlemesinde kullanılan açık kaynak yapısına sahip bir yazılımdır. Sürekli bir şekilde gelişim ve yenilikleri yapısına entegre eden KNIME, verilerin ve iş akışlarının anlaşılması ve yeniden kullanılabilen bileşenlerin tasarlanmasına imkan sağlamaktadır. KNIME Analytics Platform ile ortalama, nicelikler ve standart sapma dahil olmak üzere istatistikler türetilebilir veya bir hipotezi doğrulamak için istatistiksel testler uygulanabilir. Boyut azaltma, korelasyon analizi ve daha fazlasını iş akışlarına entegre edilebilir. Yerel makinede, veritabanı içinde veya dağıtılmış büyük veri ortamlarında verilerin toplanması, sıralanması, filtrelenmesi ve birleştirilmesi olanağı sağlar.

Normalleştirme, veri türü dönüştürme ve eksik değer işleme yoluyla verileri temizleyebilir. Platform; veri setlerini genetik algoritmalar, rastgele arama veya geriye ve ileriye dönük özelliklerin ortadan kaldırılmasıyla makine öğrenmesine hazırlamak için özellikleri ayıklar, seçer veya yenilerini oluşturur. Metinleri işleyebilir, sayısal verilere formüller uygulayabilir, örnekleri filtrelemek veya işaretlemek için kurallar uygulanabilir. Şekil 3.2 içerisinde kurulumu kolay ve ücretsiz olan KNIME Analytics Platform'a ait kullanıcı arayüzü gösterilmiştir.



Şekil 3.2. KNIME Analytics Platform kullanıcı arayüzü.

KNIME; derin öğrenme, ağaç tabanlı yöntemler ve lojistik regresyon dahil olmak üzere gelişmiş algoritmaları kullanarak sınıflandırma, regresyon, boyut küçültme veya kümeleme için makine öğrenmesi modelleri oluşturur. Hiperparametre optimizasyonu, artırma, torbalama, istifleme veya karmaşık topluluklar oluşturma ile model performansını optimize eder. Doğruluk, duyarlılık, kesinlik ve F-ölçütü gibi performans ölçümlerini uygulayarak modelleri doğrular. Model kararlılığını garantilemek için çapraz doğrulama gerçekleştirir. Çubuk ve dağılım grafiği gibi klasik grafiklerin yanı sıra paralel koordinatlar, güneş patlaması, ağ grafiği, ısı haritası gelişmiş grafiklerle verileri görselleştirir ve kullanıcıların ihtiyaçlarına göre özelleştirir. Bir KNIME tablosundaki sütunlarla ilgili özet istatistikler görüntülenebilir ve alakasız olan her şey filtrelenebilir. Veri sonuçlarının paydaşlara sunulmasında oluşturulan raporlar pdf, ppt, doc, xls gibi formatlarda dışa aktarılabilir. İşlenmiş

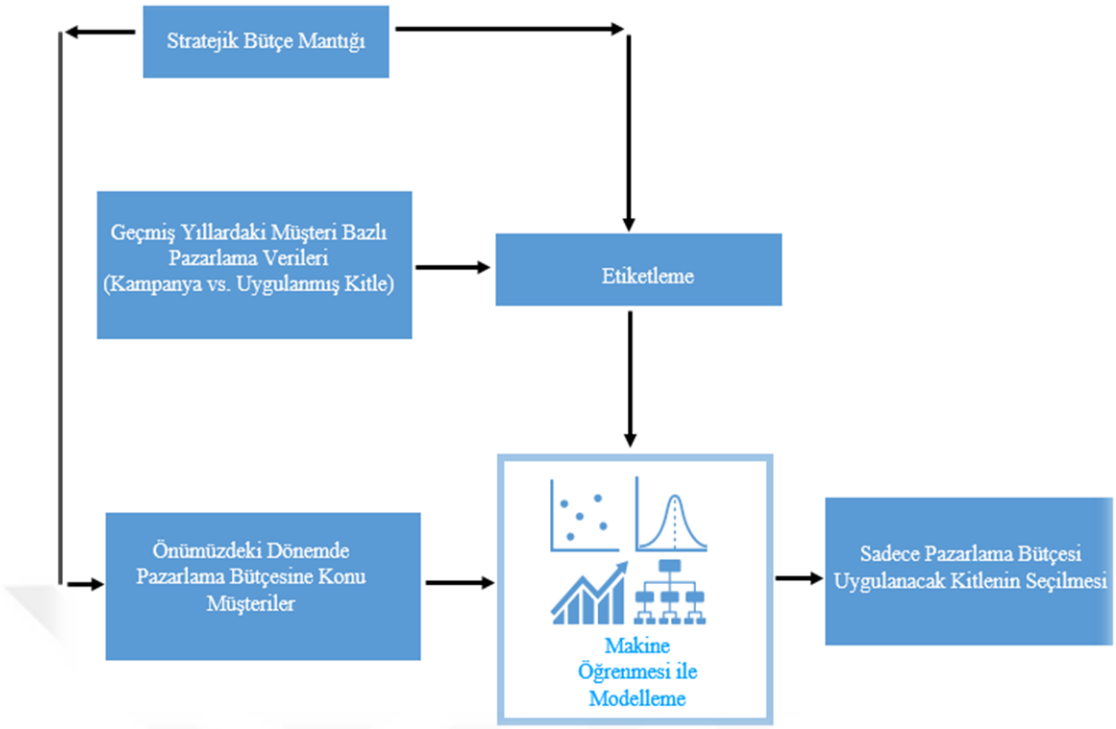
veriler veya analiz sonuçları birçok yaygın dosya biçiminde veya veritabanında depolanabilir.

Çalışma kapsamındaki sınıflandırma modellerinin analizinde ve çıktıların elde edilmesinde KNIME 4.2.1 sürümü kulllanılmıştır. KNIME Analytics; Windows, Linux ve Mac OSX ile kolay bir şekilde çalıştırılabilmektedir. Java diliyle yazılmış olan ve Eclipse tabanında kurulan KNIME Analytics Platform, ek işlevselikler sağlamak için eklentileri eklemede uzantı mekanizmalarını kullanmaktadır. Ayrıca KNIME Analytics Java'da oluşturulmasına rağmen, JavaScript, Python ve diğer kod parçalarını çalıştırmaya izin veren düğümleri desteklemesinin yanı sıra başka kodu çağıran fonksiyon çeviricilerin kullanılmasına da olanak sağlamaktadır.

3.2. Verilerin Elde Edilmesi ve Veri Seti Detayları

Bu çalışmada, pazarlama için ayrılacak bütçenin verimliliğinin arttırılması amacıyla farklı özelliklere sahip makine öğrenmesi tekniklerinden yararlanılmıştır. Kullanılan her bir tekniğin matematiksel farklılıkları araştırma problemine farklı katkılar sunacağı öngörülerek, veri seti için karar ağacı, rastgele orman, sinir ağları ve gradyan artırma algoritmaları kullanılmıştır. Bütçe verimliliğini artırmak için uygulanan genel çerçeve Şekil 3.3 içerisinde gösterilmiştir.

Bu çerçeveye göre, pazarlamanın etkin olabileceği kitlenin çeşitli makine öğrenmesi teknikleri ile seçilmesi ve skorlanması sağlanacaktır. Bu bağlamda, belirlenen genel çerçevede geçmiş yıllarda elde edilen müşteri bazlı pazarlama verileri üzerinden daha önce pazarlama bütçesi ayrılmış (kampanyaya katılan müşteriler, çağrı merkezinin aradığı ve ürün satabildiği müşteriler vb.) ve başarılı olunmuş kitle işaretlenerek farklı makine öğrenmesi teknikleri ile modeller ortaya konulur. Bu modellerden en başarılı olan seçilerek, bu modele önümüzdeki dönem pazarlama bütçesine konu olabilecek tüm müşteriler girdi olarak verilir. Model üzerinden geçen veri etiketlenir ve skorlanır. Böylece en yüksek skoru ve başarılı etiketi alan müşteriler hedeflenir.



Şekil 3.3. Pazarlamaya ayrılacak bütçenin verimliliğini artırmak için uygulanan genel çerçeve.

Çalışma kapsamında perakende sektöründe müşterilerin modern verilerine dayalı olarak hizmet veren "Dunnhumby" şirketi tarafından akademik amaçlar ile kullanıma sunulan ve "Notre Dame Üniversitesi Erkek Öğrenci Birliği (Frat)"ne ait kahvaltı ürünlerinin haftalık satışları veri seti olarak kullanılmıştır (Dunnhumby, 2019). "Frat'ta kahvaltı: Bir zaman serisi analizi" başlığı altında sunulan veri seti, dört seçilmiş kategoride (ağız bakım suyu, çubuk kraker, dondurulmuş pizza ve kahvaltılık gevrek) en çok satış yapan üç markanın her birinden en iyi beş ürün hakkında satış ve promosyon bilgilerini içermektedir.

"Frat'ta kahvaltı: Bir zaman serisi analizi" kaynak dosyasında yer alan satış ve promosyon verilerinin değişkenleri aşağıdaki gibidir:

- ✓ Ürün, mağaza ve haftaya göre birim satışlar, satın alan kişi sayısı, ziyaretler ve harcama verileri,
- ✓ Bir ürünün indirimini belirlemek için taban fiyatı ve gerçek raf fiyatı,

- ✓ Verilen ürün / mağaza / hafta için geçerliyse promosyon desteği ayrıntıları (örneğin; satış etiketi, mağaza içi reklam),
- ✓ Boyut ve konum içeren mağaza bilgileri, ve bir fiyat kademesi tayini (örneğin; pahalı vs. değerleri),
- ✓ Evrensel ürün kodu (UPC), boyut ve açıklama dahil ürün bilgileri.

Dunnhumby şirketi tarafından elde edilen veriler, Aralık 2011 tarihinden Ocak 2019 tarihine kadar geçen süre içerisindeki 156 hafta boyunca toplanmıştır. Şekil 3.3'te sunulan genel çerçevenin uygulanacağı örnek veri setindeki değişkenlerin adları ve tanımlamaları ait bilgiler Tablo 3.1 içerisinde verilmiştir. Şirket tarafından toplanılan verilerin, stok tutma birimi (SKU) ve ürün kategorilerinin sayısı sınırlı olmasına rağmen, veri seti dünyada önde gelen bir perakende veri sağlayıcısı olduğundan, yüksek kaliteli ve halka açık veriler içerdiğinden bilimsel açıdan benzersiz olup, birçok yayında doğrudan kullanılabilir bir özelliktedir (Antipov ve Pokryshevskaya, 2020; Tareq vd., 2020; Li, 2018).

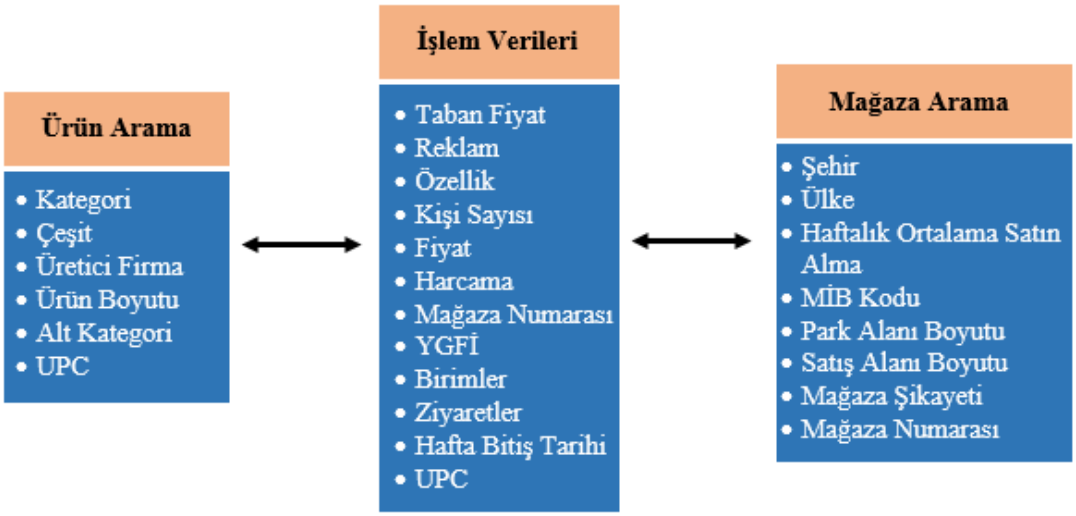
Bu veri setinin kullanılmasının en önemli amaçlarından biri; daha önce "reklam" değişkeni ile pazarlama için müşteriye ürünün gösterilip gösterilmediği hakkındaki verilere yer verilmesidir. Bu değişken hedef olarak seçilerek, kurulan modellerle verideki örüntüler bulunur. Böylece, yeni gelecek bir müşteri-ürün verisinde bu değişkenin durumu tahmin edilerek skorlanır.

Bu veri seti, bütçe verimliliğinde ve vaka çalışmalarında başarıyla kullanılabilir ve çalışmanın kullanım amacı için idealdir. Araştırmacıların "gerçek dünya" verileriyle çalışmalar yapmalarına ve tahminlerde bulunabilmelerine olanak sağlamaktadır. Bu verilerin zenginliği ve sağladığı potansiyel analizler, onu değerli bir akademik araştırma aracı haline getirmektedir. Diğer taraftan, kullanılan veri seti aynı zamanda araştırmacılara veri madenciliği yapmak için gereken süreci anlama fırsatı sağlar. Kullanılan kaynak dosya ilişkisel bir veritabanı olduğundan, araştırmacıların birden çok veri tablosunu bir araya getirmesi ve tahminlerde bulunabilmesi için de verileri bir araya getirmesi gerekmektedir.

Tablo 3.1. FRAT'a ait kahvaltı ürünlerinin haftalık satışlarına ilişkin veri seti detayları.

DEĞİŞKEN ADI	TANIMI
Taban Fiyat (Base_Price)	*Ürünün taban fiyatı*
Reklam (Display)	*Ürün, mağaza içi tanıtımın bir parçasıdır.*
Özellik (Feature)	*Ürün, mağaza içi sirkülerdedir.*
Kişi Sayısı (HHS)	*Ürün satın alan kişi sayısı*
Fiyat (Price)	*Raftaki ürün için tahsil edilen gerçek miktar*
Harcama (Spend)	*Toplam harcama (örneğin, $ satışı)*
Mağaza Numarası (Store_Num)	*Mağaza numarası*
YGFİ (TPR_Only)	*Yalnızca geçici fiyat indirimi (yani, yalnızca raf etiketi, ürünün fiyatı düşürülür ancak ürün reklamda değildir.)*
Birimler (Units)	*Satılan birimler*
Ziyaretler (Visits)	*Ürünü kapsayan benzersiz satın alma (sepet) sayısı*
Hafta Bitiş Tarihi (Week_End_Date)	*Hafta bitiş tarihi*
Evrensel Ürün Kodu (UPC)	*Ürüne özel tanımlayıcı kod*

Şekil 3.3'te gösterilen genel çerçeve, Şekil 3.4 içerisindeki veriler kullanılarak açık kaynak kodlu bir veri analiz platformu olan KNIME Analytics Platform üzerinde kurulmuştur.



Şekil 3.4. KNIME Analytics Platform'da kurulan veri yapısı.

İşlem verilerindeki her UPC için ayrıntılı ürün bilgisi sağlayan "Ürün Arama" ve her mağaza için ayrıntılı mağaza bilgileri sağlayan "Mağaza Arama" verilerinde bulunan değişkenlerin tanımlamalarına Tablo 3.2 ve Tablo 3.3 içerisinde yer verilmiştir. Veri setinin analiz edilmesi neticesinde elde edilen bulgulara bir sonraki bölümde yer verilmiştir.

Tablo 3.2. İşlem verilerindeki her UPC için ayrıntılı ürün bilgisi sağlayan "Ürün Arama" verilerindeki değişkenler ve tanımları.

DEĞİŞKEN ADI	TANIMI
Kategori	*Ürünün kategorisi*
Çeşit	*Ürünün çeşidi*
Üretici Firma	*Ürünü imal eden firma*
Ürün Boyutu	*Paket boyutu veya ürün miktarı*
Alt Kategori	*Ürünün alt kategorileri*
UPC	*Ürüne özel tanımlayıcı kod*

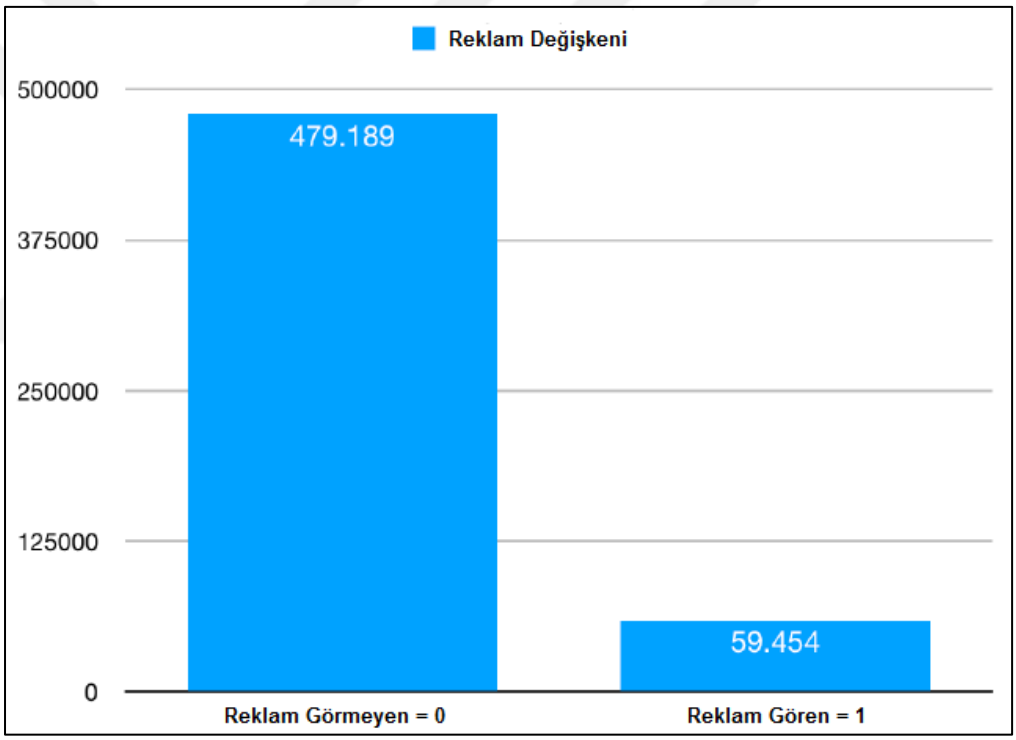
Tablo 3.3. İşlem verilerindeki her mağaza için ayrıntılı mağaza bilgileri sağlayan "Mağaza Arama" verilerindeki değişkenler ve tanımları.

DEĞİŞKEN ADI	TANIMI
Şehir	*Şehir*
Ülke	*Ülke / Eyalet*
Haftalık Ortalama Satın Alma	*Mağazada satılan ortalama haftalık sepetler*
MİB Kodu	*(Metropolitan İstatistik Bölgesi) yüksek bir çekirdek nüfus yoğunluğuna sahip coğrafi bölge ve çevredeki alanlar boyunca yakın ekonomik bağlar*
Park Alanı Boyutu	*Kroger otoparkındaki park yeri sayısı*
Satış Alanı Boyutu	*Kroger mağazasının alan ölçüsü*
Mağaza Şikayeti	*Kroger mağazasına yapılan şikayet başvuruları*
Mağaza Numarası	*Mağaza numarası*

4. BULGULAR

Bu bölümde pazarlama için ayrılacak bütçenin verimliliğinin arttırılmasına yönelik karar ağacı, rastgele orman, sinir ağları ve gradyan artırma algoritmalarının deney veri setine uygulanması ve yapılan analizler sonucu elde edilen bulgulara ait detaylara yer verilmiştir.

Bu çalışmada kullanılan deney veri setinde 538.643 adet pazarlama verisi bulunmaktadır. Deney veri setindeki reklam (display) değişkenine göre reklam görmeyen 479.189 müşteri "0" olarak ve reklam gören 59.454 müşteri "1" olarak etiketlenmiştir. Reklam değişkenine göre deney veri setinin sınıf dağılımı Şekil 4.1. içerisinde gösterilmiştir.



Şekil 4.1. Deney veri setindeki reklam değişkeninin sınıf dağılımı histogramı.

4.1. Özellik Seçimi

Özellik seçiminde sarmalayıcı yöntemlerden (wrapper methods) ileri özellik seçimi ve geri özellik eleme teknikleri kullanılmıştır. Sarmalayıcı yöntemlerde, bir özellik alt kümesi kullanmaya ve bunları kullanarak bir model eğitmeye çalışılır. Önceki modelden yapılan çıkarımlara dayanarak, alt kümeden özellikler eklemeye

veya kaldırmaya karar verilir. Sarmalayıcı yöntemlerin ileri özellik seçimi, geri özellik eleme, özyinelemeli özellik eliminasyonu gibi bazı yaygın örnekleri bulunmaktadır. İleri özellik seçiminde, modelde hiçbir özellik olmadan başlanılan yinelemeli bir yöntemdir. Her yinelemede, yeni bir değişken eklenmesi ile modelin performansı artmadığı sürece modeli en iyi şekilde geliştiren özelliği eklemeye devam edilir. Geri özellik elemede tüm özelliklerle başlanır ve her yinelemede modelin performansını artıran en önemsiz özellik kaldırılır. Özelliklerin kaldırılmasında herhangi bir gelişme görülmeyene kadar bu durum tekrarlanır. Hem ileri özellik seçimi hem de geri özellik eleme ile 24 farklı özellikten ekleme ve elemeler yapılarak verilerin karmaşıklık durumu azaltılmış ve daha az özellik ile daha yüksek doğruluk elde etmek için kullanılmıştır. Şekil 4.2. içerisinde ileri özellik seçiminde ve Şekil 4.3. içerisinde geriye doğru özellik eliminasyonunda seçilen özellikler sunulmuştur. Seçilen özellikler mavi renk ile gösterilmiştir. İleri özellik seçimi tekniği, deney setindeki verilerin en fazla 0.91 doğruluk ile 19 farklı özelliğin yer aldığı kombinasyonun seçimine olanak sağlamıştır. Geri özellik eleme tekniğinde ise, elde edilen sonuçlar ileri özellik seçimi tekniğine çok yakın olmakla birlikte, maksimum doğruluk 0.909 değerine sahiptir.

Accuracy	Nr. of features	
0.91	19	WEEK_END_DATE, STORE_NUM
0.909	18	UPC
0.909	20	UNITS
0.909	11	VISITS
0.909	10	HHS
0.909	9	SPEND
0.909	8	PRICE
0.909	7	BASE_PRICE
0.909	6	FEATURE
0.908	17	DISPLAY
0.908	14	TPR_ONLY
0.908	13	STORE_NAME
0.908	12	ADDRESS_CITY_NAME
0.908	5	ADDRESS_STATE_PROV_CODE
0.908	4	MSA_CODE
0.908	3	SEG_VALUE_NAME
0.908	2	PARKING_SPACE_QTY
0.908	1	SALES_AREA_SIZE_NUM
0.908	16	AVG_WEEKLY_BASKETS
0.907	15	DESCRIPTION
0.907	21	MANUFACTURER
0.901	22	CATEGORY
0.895	23	SUB_CATEGORY
0.894	24	PRODUCT_SIZE

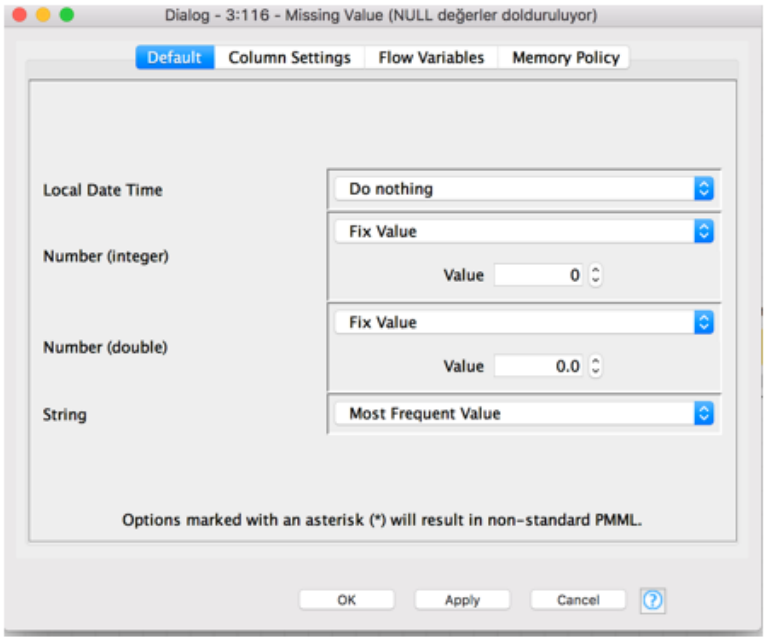
Şekil 4.2. İleri özellik seçimi tekniğindeki özelliklerin sayısı ve doğruluk değişimi.

Accuracy	Nr. of features	
0.909	21	WEEK_END_DATE
0.909	20	STORE_NUM
0.909	19	UPC
0.909	18	UNITS
0.909	17	VISITS
0.909	16	HHS
0.909	15	SPEND
0.909	14	PRICE
0.909	13	BASE_PRICE
0.909	12	FEATURE
0.909	11	DISPLAY
0.909	10	TPR_ONLY
0.909	9	STORE_NAME
0.909	8	ADDRESS_CITY_NAME
0.909	7	ADDRESS_STATE_PROV_CODE
0.908	22	MSA_CODE
0.908	6	SEG_VALUE_NAME
0.908	5	PARKING_SPACE_QTY
0.908	4	SALES_AREA_SIZE_NUM
0.908	3	AVG_WEEKLY_BASKETS
0.908	2	DESCRIPTION
0.908	1	MANUFACTURER
0.899	23	CATEGORY
0.894	24	SUB_CATEGORY
		PRODUCT_SIZE

Şekil 4.3. Geri özellik eleme tekniğindeki özelliklerin sayısı ve doğruluk değişimi.

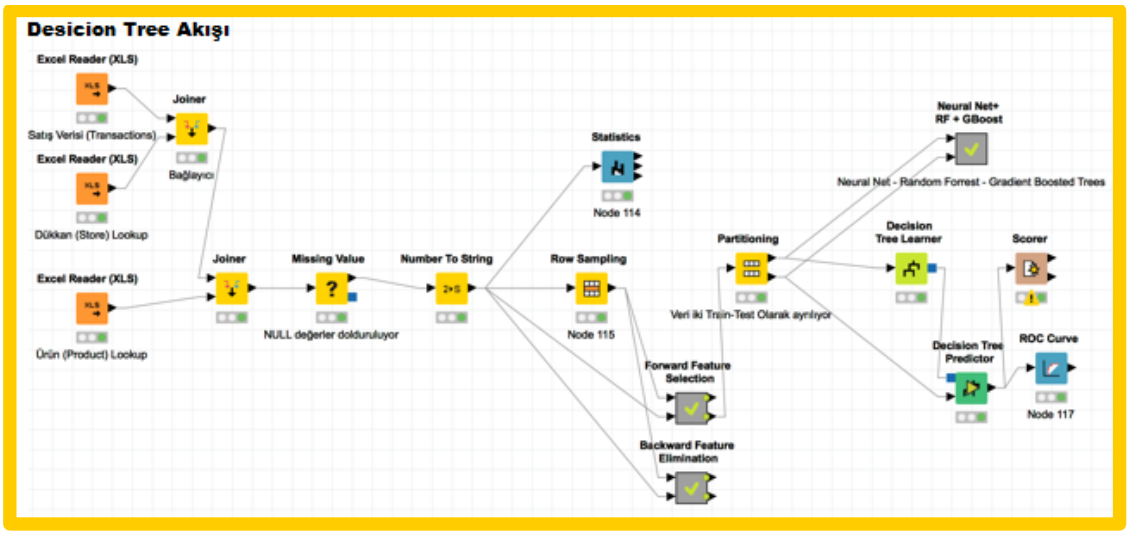
4.2. Sistem Mimarisi ve Akışı

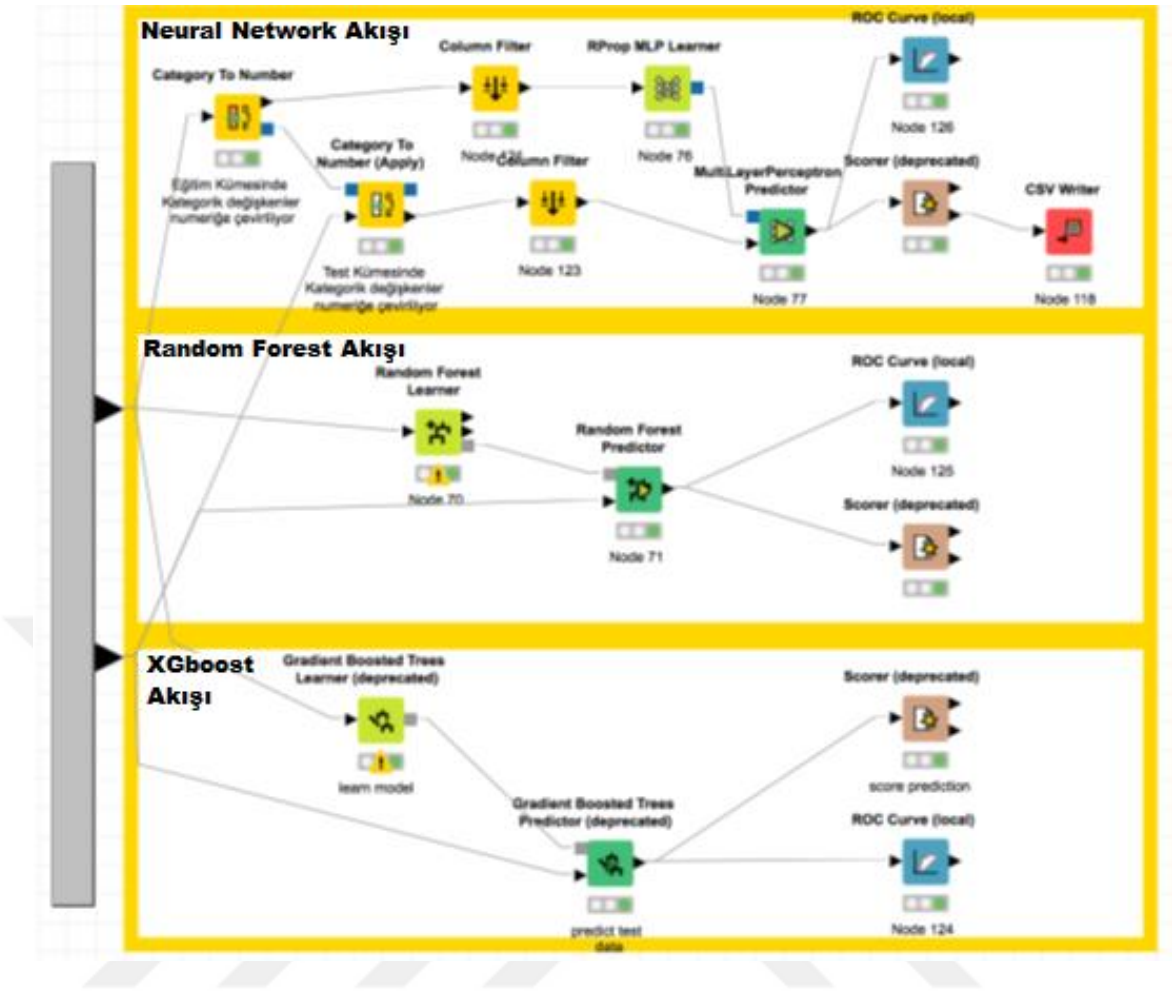
Deney veri seti kullanılmadan önce, MS Excel dosyası halinde indirilen üç farklı mağaza, ürün ve satış verileri birleştirilmiştir. Dosyadaki boş sütunlar uygun değerler ile doldurulmuştur. Numerik boş değerler, 0 ile kategorik değişkenler en sık geçen değişken ile doldurulur. Şekil 4.4. içerisinde bu değerlere yer verilmiştir.



Şekil 4.4. Boş sütunların uygun değerler ile doldurulması.

Bu işlemden sonra, sistemin performansının ölçülmesi amacıyla rastgele 50.000 kayıt seçilerek, özellik seçiminde ileri özellik seçimi ve geri özellik eleme teknikleri ile sistem çalıştırılmıştır. İleri özellik seçimi yöntemi daha iyi sonucu verdiğinden önerilen özellikler seçilerek tüm veriler iki gruba ayrılmıştır. Bir makine öğrenmesi algoritması performansının değerlendirilmesinde hangi veri setinin kullanılması gerektiği önemlidir. Eğitim verileri, bir makine öğrenmesi algoritması veya modelini eğitmek için kullanılan zenginleştirilmiş verilerdir. Orijinal verileri eğitim ve teste bölmek son derece önemlidir. Bu, herhangi bir makine öğrenmesi modelinin doğruluğunu ve güvenilirliğini belirlemektedir. Makine öğrenmesi modelini eğitme süreci önceden tahmin edilen sonucun eğitim veri setine ayarlanır. Bu nedenle, genelleme hatasını tahmin etmek için modelin henüz görmediği bir veri setinin test edilmesi gerekir. Bu bağlamda, modeli test etmek için etiketlenmiş bir veri setine ihtiyaç duyulur. Yapılan çalışmada veri setinde bulunan 538.643 adet pazarlama verisinin %70'i (377.050 adet) modeli eğitmek için geri kalan %30'u (161.593 adet) ise modeli test etmek için kullanılmıştır. Sınıflandırma modellerinin eğitilmesinde eğitim veri setinden yararlanılmıştır. Model başarısının ölçümü, %30 oranında ayrılmış olan test veri setiyle gerçekleştirilmiştir. Her bir algoritma için sistem mimarisinin akış şeması Şekil 4.5. içerisinde sunulmuştur.





Şekil 4.5. Sistem mimarisinin akış şeması.

4.3. Analiz Sonuçları

Çalışmanın bu kısmında karar ağacı, rastgele orman, çok katmanlı algılayıcılar ve gradyan artırma algoritmalarının pazarlama bütçesi üzerine sınıflandırma uygulamaları gerçekleştirilmiş ve yapılan analizler sonucu elde edilen bulgulara detaylı şekilde yer verilmiştir.

Çalışma kapsamında, test veri setinin analizinde kullanılan algoritmalar sınıflandırma modelleri için sıklıkla tercih edilen algoritmalardır. En iyi model başarısının test edilmesinde kullanılan bu sınıflandırma algoritmaları KNIME 4.2.1 programı ile analiz edilmiştir.

Bu çalışma kapsamında, sınıflandırma modellerinin performanslarının ölçümünde çok sayıda sınıflandırma ölçütünün hesaplanmasında kullanılan karmaşıklık matrisi yönteminden yararlanılmıştır. Karmaşıklık matrisi,

sınıflandırılmış olan verilerin gerçek ve tahmin sınıfları gösterilir. Sınıflandırmadan elde edilen sonuçlar matris satırlarında bulunurken, gerçek değerler matris sütunlarında bulunmaktadır. Çalışmada kullanılan veri setinde "reklam (display)" değişkeninin pazarlama için müşteriye ürünün gösterilip gösterilmediği durumu önemlidir. Reklam gören ve reklam görmeyen şeklinde ikiye ayrılan ikili sınıflandırma durumunun karmaşıklık matrisine ait gösterim Tablo 4.1. içerisinde verilmiştir.

Tablo 4.1. Reklam değişkenine göre ikili sınıflandırma durumunun karmaşıklık matrisine ait gösterimi.

Karmaşıklık Matrisi		Verinin Tahmin Edilen Sınıfı	
		Reklam Gören	Reklam Görmeyen
Verinin Gerçek Sınıfı	Reklam Gören	Doğru Pozitif (DP)	Yanlış Negatif (YN)
	Reklam Görmeyen	Yanlış Pozitif (YP)	Doğru Negatif (DN)

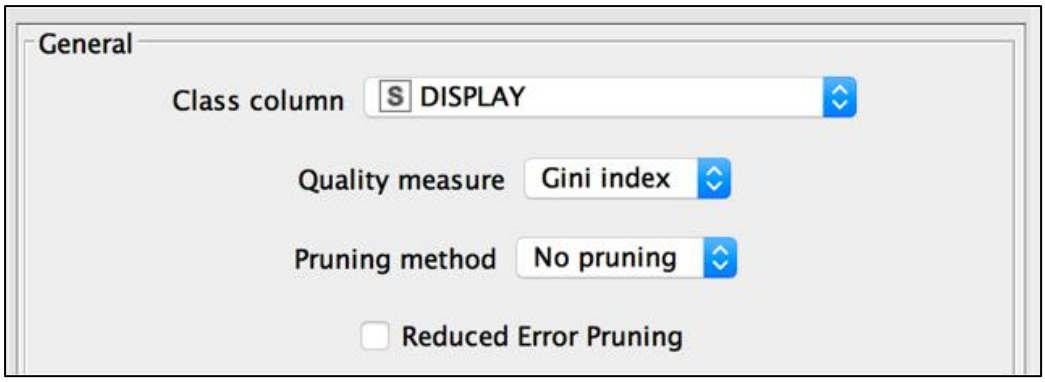
Karmaşıklık matrisi tablosundaki *doğru pozitif (DP)* gerçekte reklam görenleri ve sınıflandırma sonucu reklam görenler şeklinde etiketlenen örneğin sayısını, *doğru negatif (DN),* gerçekte reklam görmeyenleri ve sınıflandırma sonucu reklam görmeyenler şeklinde etiketlenen örneğin sayısını temsil eder. Hem doğru pozitif hem de doğru negatif verileri gerçek sınıfı ile tahmin edilen sınıfı aynı şeyi söyler. *Yanlış pozitif (YP)* gerçekte reklam görmeyenleri ve sınıflandırma sonucu reklam görenler şeklinde etiketlenen örneğin sayısını göstermektedir. *Yanlış negatif (YN)* gerçekte reklam görenleri ve sınıflandırma sonucu reklam görmeyenler şeklinde etiketlenen örneğin sayısını gösterir. Yanlış pozitif ile yanlış negatif verileri gerçek sınıf ile tahmin edilen sınıfın birbirleriyle çeliştiği durumlarda gözlemlenir.

Tez çalışması kapsamında model performansının ölçümünde doğruluk (accuracy), duyarlılık (recall or sensitivity), kesinlik (precision), seçicilik (specificity), F-Ölçütü (F-Measure) metrikleri hesaplanmıştır. Bir makine öğrenmesi sınıflandırma algoritmasının doğruluğu, algoritmanın bir veri noktasını ne sıklıkla doğru şekilde sınıflandırdığını ölçmenin bir yoludur. Doğruluk, tüm veri noktalarından doğru tahmin

edilen veri noktalarının sayısıdır. Çoğunlukla doğruluk, çeşitli doğru/yanlış pozitif/negatif oranlarını kullanan diğer ölçümler olan duyarlılık ve geri çağırmayla birlikte kullanılır. Bu ölçümlerle birlikte, algoritmanın veri noktalarını nasıl sınıflandırdığına ayrıntılı bir bakış açısı sağlar. Bu bağlamda, yalnızca doğruluk oranının en yüksek değeri en iyi model performansını ortaya koymaktadır. Doğruluk ölçütü, modelin bir bütün halinde doğru olarak sınıflandırma oranıdır. Reklam gören müşterilerin doğru olarak tahmin edilmesinin önemli olduğu kadar, reklam görebilecek müşterilerin tespiti de o ölçüde önemlidir. Bu bağlamda, duyarlılık ölçütü dikkate alınması gereken önemli bir ölçüttür. Duyarlılık, gerçekte reklam gören müşterilerin ne kadarının sınıflandırıcı tarafından reklam gören müşteriler olarak tahmin edildiğine ilişkin oranı verir. Kesinlik, doğru pozitif sonuçların sayısının sınıflandırıcı tarafından tahmin edilen pozitif sonuçların sayısına bölünmesiyle elde edilen orandır. Bu durumda, kesinlik oranı gerçekte reklam gören müşterilerin sayısının sınıflandırıcı tarafından reklam gören olarak tahmin edilen müşteri sayısına bölünmesidir. Seçicilik, gerçekte reklam görmeyen müşterilerin ne kadarının sınıflandırıcı tarafından reklam görmeyen müşteriler olarak tahmin edildiğine ilişkin oranı verir. Kesinlik ve duyarlılığın harmonik ortalaması olan F-Ölçütü hem yanlış pozitif hem de yanlış negatif değerleri hesaplar. F-Ölçütü, kesinlik ve duyarlılık ölçütlerinin hesaplanmasında tek başına eksik veya yetersiz olabileceği durumlar gözetilerek hesaplanır (Powers, 2007).

4.3.1. Karar Ağacı Algoritması

Karar ağacı algoritmasının uygulanmasında, reklam değişkeni en iyi öznitelik olarak seçilmiş ve karar ağacının kök düğümünün test edilmesi amacıyla kullanılmıştır. En iyi özniteliği, veri seti içindeki hedef değişkenin homojenliğini ölçmek ve karar ağacını oluşturmak için ölçüt olarak Gini index kullanılmıştır. Deney veri setinde karar ağacı algoritmasının uygulanmasında budama ölçütleri kullanılmamıştır. Şekil 4.6. içerisinde karar ağacı algoritmasının deney veri setine uygulanmasında kullanılan parametrelere yer verilmiştir.



Şekil 4.6. Karar ağacı algoritmasının deney veri setine uygulanmasında kullanılan parametreler.

Deney veri setinde kullanılan karar ağacı algoritmasına yönelik performans ölçümü için kullanılan karmaşıklık matrisine ait değerler Tablo 4.2. içerisinde sunulmuştur.

Tablo 4.2. Deney veri setinde kullanılan karar ağacı algoritmasının performans ölçümünden elde edilen karmaşıklık matrisine ait değerler.

Algoritma Türü	Karmaşıklık Matrisi		Verinin Tahmin Edilen Sınıfı		Toplam Veri Sayısı
			Reklam Gören	Reklam Görmeyen	
Karar Ağacı	Verinin Gerçek Sınıfı	Reklam Gören (Display = 1)	10.689	7.147	17.836
		Reklam Görmeyen (Display = 0)	3.583	140.174	143.757
	Toplam Veri Sayısı		14.272	147.321	161.593

Karar ağacı algoritmasının deney veri setine uygulanması sonucu performans ölçümü için kullanılan karmaşıklık matrisi tablosundaki değerler incelendiğinde, DP değerinin 10.689, YP değerinin 3.583, YN değerinin 7.147 ve DN değerinin 140.174 olduğu görülmektedir. Doğru olarak sınıflandırılmış olan reklam gören ve reklam görmeyen sayısı 150.863'tür. 7.147 pazarlama verisinin gerçekte reklam gören olup sınıflandırma sonucu reklam görmeyen, 3.583 pazarlama verisinin gerçekte reklam görmeyen olup sınıflandırma sonucu reklam gören şeklinde etiketlendiği görülmektedir.

Karar ağacı algoritması için model performansı ölçümünden elde edilen değerlere ilişkin doğruluk, duyarlılık, kesinlik, seçicilik ve F-Ölçütü oranları Tablo 4.3. içerisinde sunulmuştur.

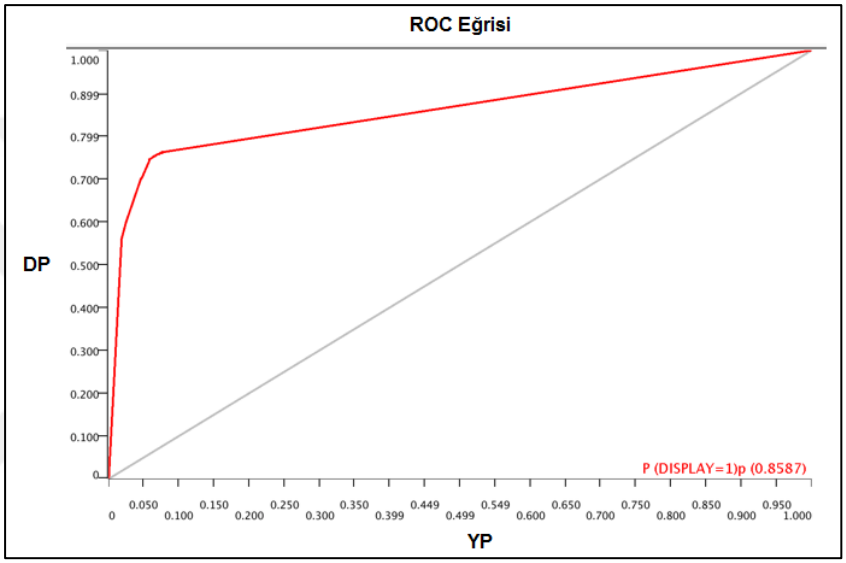
Tablo 4.3. Karar ağacı algoritmasının model performansı ölçümünden elde edilen değerler.

Algoritma Türü	**Model Performansının Ölçümüne Ait Hesaplamalar**					
		Duyarlılık	**Kesinlik**	**Seçicilik**	**F-Ölçütü**	**Doğruluk**
Karar Ağacı	**Reklam Gören**	**0.975**	**0.952**	0.604	**0.964**	
	Reklam Görmeyen	0.604	0.749	**0.975**	0.669	
	Genel	0.790	0.851	0.790	0.816	**0.934**

Karar ağacı algoritması için model performansının ölçümünden elde edilen doğruluk değeri 0.934'tür. Makine öğrenmesinde sınıflandırma algoritmasının doğruluğu, algoritmanın bir veri noktasını ne sıklıkla doğru şekilde sınıflandırdığını ölçmektedir. Elde edilen 0.934 doğruluk değeri, karar ağacı algoritmasının deney veri setindeki tüm veriler için doğru tahmin edilen veri sayısının yüksek olduğunu göstermektedir. Bu bağlamda, doğruluk oranı karar ağacı algoritması için iyi bir model performansını ortaya koyduğu söylenebilir. Gerçekte reklam gören müşterilerin ne kadarının sınıflandırıcı tarafından reklam gören müşteriler olarak tahmin edildiğine ilişkin oranı veren duyarlılık değeri 0.790'dır. Yine gerçekte reklam gören müşterilerin sayısının sınıflandırıcı tarafından reklam görenler şeklinde tahmin edilen müşteri sayısına bölünerek elde edilen kesinlik değeri 0.851'dir. Gerçekte reklam görmeyen müşterilerin ne kadarının sınıflandırıcı tarafından reklam görmeyen müşteriler olarak tahmin edildiğine ilişkin oranı veren seçicilik değeri 0.790'dır. Kesinlik ve duyarlılığın harmonik ortalaması olan F-Ölçütü değeri 0.816'dır. Model performans ölçümünden elde edilen değerler reklam gören ve reklam görmeyenler şeklinde ayrı ayrı incelendiğinde, reklam görenler için duyarlılık, kesinlik, seçicilik ve F-Ölçütü değerleri sırasıyla 0.975, 0.952, 0.604 ve 0.964 iken, reklam görmeyenler için duyarlılık, kesinlik, seçicilik ve F-Ölçütü değerleri sırasıyla 0.604, 0.851, 0.790 ve

0.816'dır. Reklam görenler için duyarlılık, kesinlik ve F-ölçütü metriklerinin 1'e çok yakın yüksek değerlerde olduğu, reklam görmeyenler için ise sadece seçicilik metriğinin 0.975 oranı ile 1'e çok yakın yüksek değerde olduğu görülmektedir.

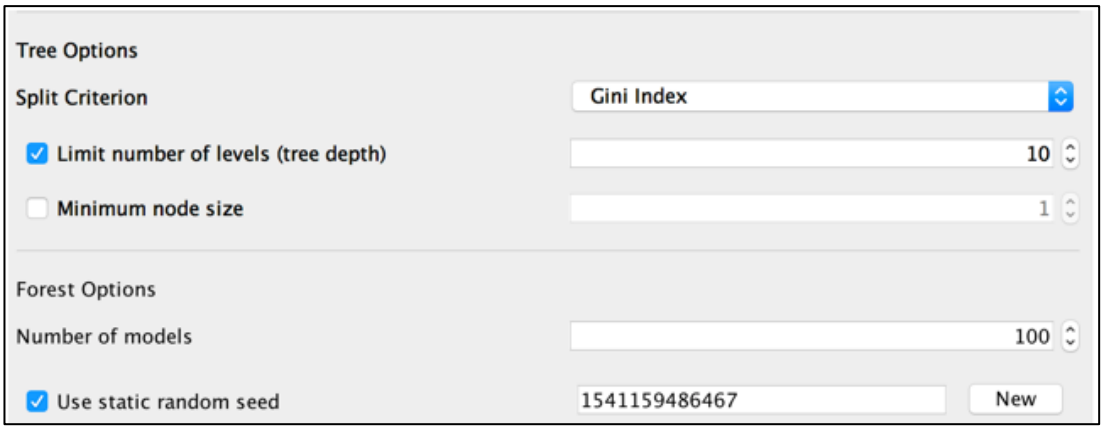
Çalışmada kullanılan algoritmalardan hangisinin sınıfları en iyi tahmin ettiğini belirlemek için her bir algoritmanın sınıflandırma analizinde AUC değerleri elde edilmiştir. Karar ağacı algoritmasının uygulanması sonucu model performansının değerlendirilmesine ilişkin AUC kriterine ait sonuç Şekil 4.7.'de sunulmuştur. Karar ağacı algoritması için AUC değerinin 0.8587 olduğu görülmektedir.



Şekil 4.7. Karar ağacı algoritması için ROC eğrisi.

4.3.2. Rastgele Orman Algoritması

Rastgele orman algoritması, karar ağacı algoritmalarının bir araya getirilmesi gibidir. Rastgele ormanlar, değişkenlerin azaltılması amacıyla aynı eğitim setinin farklı bölümleri için eğitilmiş çok sayıdaki derin karar ağacının ortalamasını almanın bir yoludur. Rastgele orman algoritmasının uygulanmasında 100 model kullanılmış olup bu ağacın 100 kez tekrarlanması anlamına gelmektedir. Ağaç derinliği sayısı 10 ile sınırlı tutulmuştur. Rastgele orman algoritmasının uygulanmasında Gini index kriteri kullanılmıştır. Şekil 4.8. içerisinde rastgele orman algoritmasının deney veri setine uygulanmasında kullanılan parametrelere yer verilmiştir.



Şekil 4.8. Rastgele orman algoritmasının deney veri setine uygulanmasında kullanılan parametreler.

Deney veri setinde kullanılan rastgele orman algoritmasına yönelik performans ölçümü için kullanılan karmaşıklık matrisine ait değerler Tablo 4.4. içerisinde sunulmuştur.

Tablo 4.4. Deney veri setinde kullanılan karar ağacı algoritmasının performans ölçümünden elde edilen karmaşıklık matrisine ait değerler.

Algoritma Türü	Karmaşıklık Matrisi		Verinin Tahmin Edilen Sınıfı		Toplam Veri Sayısı
			Reklam Gören	Reklam Görmeyen	
Rastgele Orman	Verinin Gerçek Sınıfı	Reklam Gören (Display = 1)	8.161	9.675	17.836
		Reklam Görmeyen (Display = 0)	2.654	141.103	143.757
	Toplam Veri Sayısı		10.815	150.778	161.593

Rastgele orman algoritmasının deney veri setine uygulanması sonucu performans ölçümü için kullanılan karmaşıklık matrisi tablosundaki değerler incelendiğinde, DP değerinin 8.161, YP değerinin 2.654, YN değerinin 9.675 ve DN değerinin 141.103 olduğu görülmektedir. Doğru olarak sınıflandırılmış olan reklam gören ve reklam görmeyen sayısı 149.264'tür. 9.675 pazarlama verisi gerçekte reklam gören olup sınıflandırma sonucu reklam görmeyen, 2.654 pazarlama verisi gerçekte reklam görmeyen olup sınıflandırma sonucu reklam gören şeklinde etiketlenmektedir.

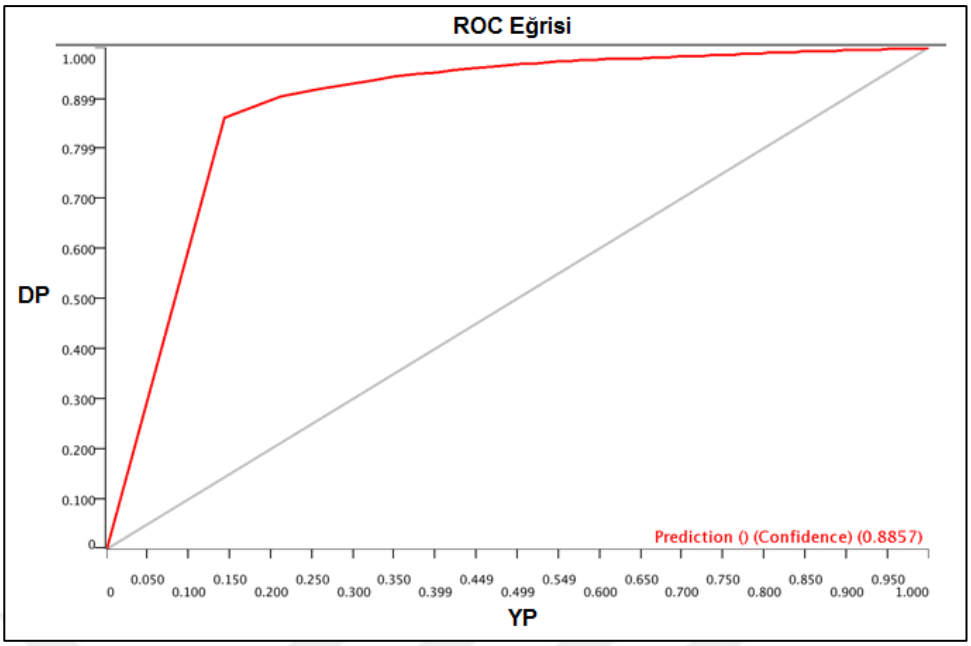
Model performansı ölçümünden rastgele orman algoritması için elde edilen değerlere ilişkin doğruluk, duyarlılık, kesinlik, seçicilik ve F-Ölçütü oranlarına Tablo 4.5. içerisinde yer verilmiştir.

Tablo 4.5. Rastgele orman algoritmasının model performansı ölçümünden elde edilen değerler.

Algoritma Türü	**Model Performansının Ölçümüne Ait Hesaplamalar**					
		Duyarlılık	**Kesinlik**	**Seçicilik**	**F-Ölçütü**	**Doğruluk**
Rastgele Orman	**Reklam Gören**	**0.982**	**0.936**	0.458	**0.958**	
	Reklam Görmeyen	0.458	0.755	**0.982**	0.570	
	Genel	0.720	0.845	0.720	0.764	**0.924**

Rastgele orman algoritmasının doğruluğuna ilişkin model performans ölçümü sonucu 0.924'tür. Rastgele orman algoritması için elde edilen bu doğruluk değeri doğru tahmin edilen veri sayısının yüksek olduğunu göstermektedir. Doğruluk oranı bu algoritma için iyi bir değere sahiptir. Model performansının ölçümüne ait duyarlılık, kesinlik, seçicilik ve F-Ölçütü değerleri sırasıyla 0.720, 0.845, 0.720 ve 0.764'tür. Model performans ölçümünden elde edilen değerler reklam gören ve reklam görmeyenler şeklinde ayrı ayrı incelendiğinde, reklam görenler için duyarlılık, kesinlik, seçicilik ve F-Ölçütü değerleri sırasıyla 0.982, 0.936, 0.458 ve 0.958 iken, reklam görmeyenler için duyarlılık, kesinlik, seçicilik ve F-Ölçütü değerleri sırasıyla 0.458, 0.755, 0.982 ve 0.570'tir. Reklam görenler için duyarlılık, kesinlik ve F-ölçütü metriklerinin 1'e çok yakın yüksek değerlerde olduğu, reklam görmeyenler için ise sadece seçicilik metriğinin 0.982 oranı ile 1'e çok yakın yüksek bir değerde olduğu görülmektedir.

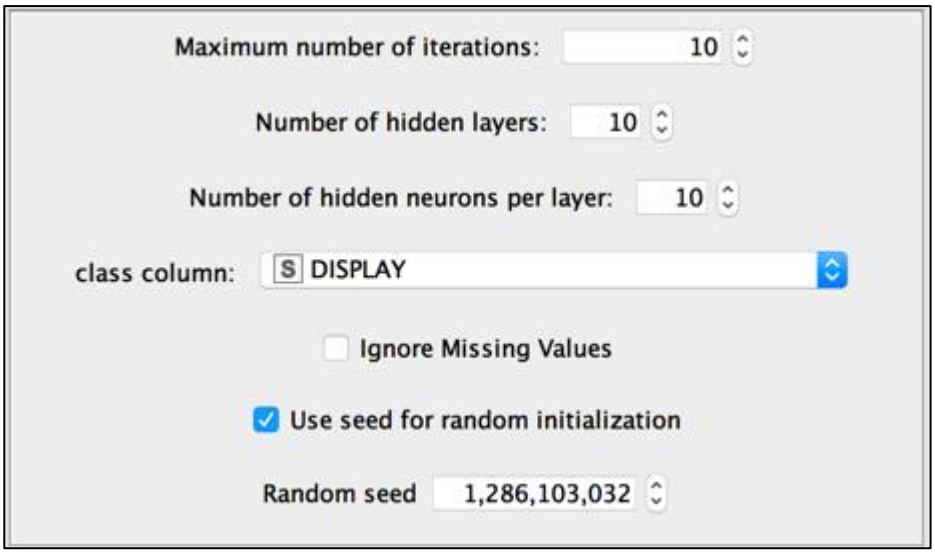
Rastgele orman algoritması için AUC kriterine ilişkin sonuç Şekil 4.9.'da sunulmuştur. Rastgele orman algoritması için AUC değeri 0.8857'dir.



Şekil 4.9. Rastgele orman algoritması için ROC eğrisi.

4.3.3. Çok Katmanlı Algılayıcı Algoritması

Bu tez çalışması kapsamında, ileri beslemeli yapay sinir ağının (YSA) bir sınıfı olan ve bir bağlantı modelinde birbirine bağlanmış çok sayıda nörondan oluşan çok katmanlı algılayıcı (MLP) algoritması kullanılmıştır. Veri setinin eğitilmesinde ise geri yayılım algoritmasından yararlanılmıştır. Deney veri setinde kullanılan çok katmanlı sinir ağının yapısı; işlenecek bilgilerin alındığı giriş katmanı, işlem sonuçlarının elde edildiği çıktı katmanı, giriş ve çıkış katmanları arasındaki gizli katman şeklindeki üç düğüm katmandan oluşmaktadır. Giriş ve çıkış katmanındaki nöron sayısı 10'dur. Çıkış katmanındaki optimum nöron sayısının bulunmasında deneme-yanılma yöntemi ile gizli katmandaki farklı sayıdaki nöronlar kullanılarak belirlenmiştir. Maksimum tekrar sayısı ise denemeler yapılarak 10 olarak belirlenmiştir. Şekil 4.10. içerisinde deney veri setine uygulanan çok katmanlı algılayıcı algoritması için kullanılan parametrelere yer verilmiştir.



Şekil 4.10. Çok katmanlı algılayıcı algoritmasının deney veri setine uygulanmasında kullanılan parametreler.

Deney veri setinde kullanılan çok katmanlı algılayıcı algoritmasına yönelik performans ölçümü için kullanılan karmaşıklık matrisine ait değerler Tablo 4.6. içerisinde sunulmuştur.

Tablo 4.6. Deney veri setinde kullanılan çok katmanlı algılayıcı algoritmasının performans ölçümünden elde edilen karmaşıklık matrisine ait değerler.

Algoritma Türü	Karmaşıklık Matrisi		Verinin Tahmin Edilen Sınıfı		Toplam Veri Sayısı
			Reklam Gören	Reklam Görmeyen	
Çok Katmanlı Algılayıcılar	Verinin Gerçek Sınıfı	Reklam Gören (Display = 1)	0	17.836	17.836
		Reklam Görmeyen (Display = 0)	0	143.757	143.757
	Toplam Veri Sayısı		0	161.593	161.593

Çok katmanlı algılayıcı algoritmasının deney veri setine uygulanması sonucu performans ölçümü için kullanılan karmaşıklık matrisi tablosundaki değerler incelendiğinde, DP değerinin 0, YP değerinin 0, YN değerinin 17.836 ve DN değerinin 143.757 olduğu görülmektedir. Doğru olarak sınıflandırılmış olan reklam görmeyen sayısı 143.757'dir. 17.836 pazarlama verisi gerçekte reklam gören olup sınıflandırma

sonucu reklam görmeyen, 0 pazarlama verisi gerçekte reklam görmeyen olup sınıflandırma sonucu reklam gören şeklinde etiketlenmektedir.

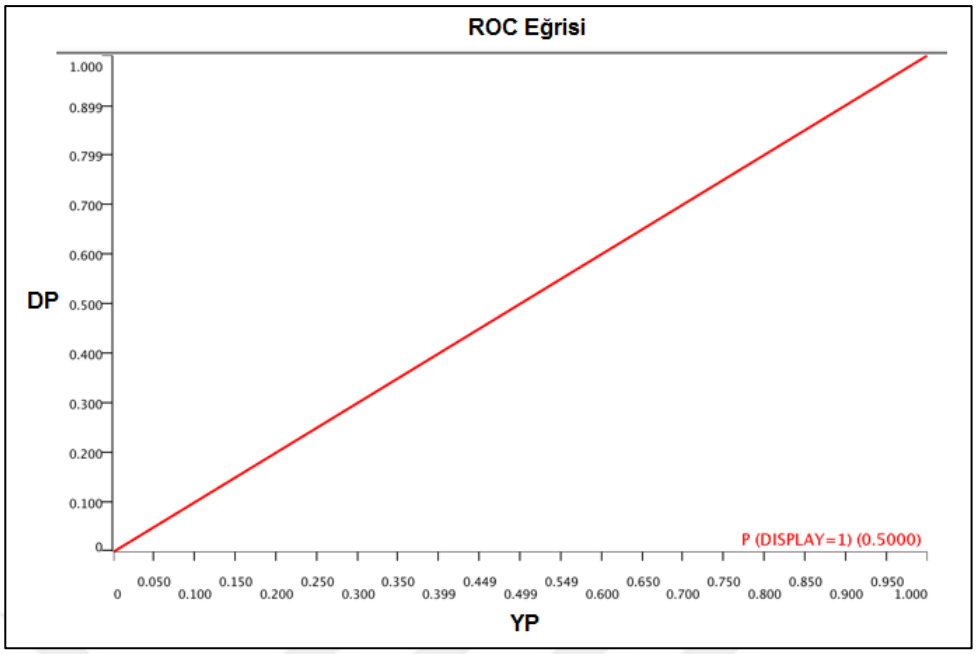
Model performansı ölçümünden çok katmanlı algılayıcı algoritması için elde edilen değerlere ilişkin doğruluk, duyarlılık, kesinlik, seçicilik ve F-Ölçütü oranlarına Tablo 4.7. içerisinde yer verilmiştir.

Tablo 4.7. Çok katmanlı algılayıcı algoritmasının model performansı ölçümünden elde edilen değerler.

Algoritma Türü	**Model Performansının Ölçümüne Ait Hesaplamalar**					
		Duyarlılık	**Kesinlik**	**Seçicilik**	**F-Ölçütü**	**Doğruluk**
Çok Katmanlı Algılayıcılar	**Reklam Gören**	**1**	**0.890**	0	**0.942**	
	Reklam Görmeyen	0	0	**1**	0	
	Genel	0.500	0.445	0.500	0.471	**0.890**

Model performansının ölçümüne ilişkin çok katmanlı algılayıcı algoritmasının doğruluk değeri 0.890'tür. Elde edilen bu doğruluk değeri için doğru tahmin edilen veri sayısının iyi bir düzeyde olduğu söylenebilir. Model performansının ölçümüne ait duyarlılık, kesinlik, seçicilik ve F-Ölçütü değerleri sırasıyla 0.500, 0.445, 0.500 ve 0.471'dir. Model performans ölçümünden elde edilen değerler reklam gören ve reklam görmeyenler şeklinde ayrı ayrı incelendiğinde, reklam görenler için duyarlılık, kesinlik, seçicilik ve F-Ölçütü değerleri sırasıyla 1, 0.890, 0 ve 0.942 iken, reklam görmeyenler için seçicilik değeri 1 ve duyarlılık, kesinlik, F-Ölçütü değerlerinin tamamı 0'dır. Reklam görenler için duyarlılık oranının 1 olması, bu metriğin mükemmel bir değere, kesinlik ve F-Ölçütü değerlerinin 1'e çok yakın yüksek bir değere sahip olduğunu göstermektedir. Reklam görmeyenler için seçicilik metriği haricindeki diğer metriklerin 0 değerine sahip olması, çok katmanlı algılayıcı algoritması için sadece reklam gören müşteri verilerinin değerlendirmeye alınmasından kaynaklanmaktadır.

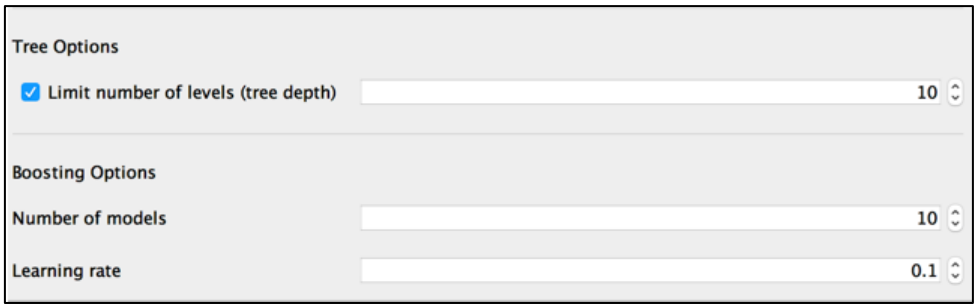
Çok katmanlı algılayıcı algoritması için AUC kriterine ilişkin sonuç Şekil 4.11.'de sunulmuştur. Çok katmanlı algılayıcı algoritması için AUC değeri 0.50'dir.



Şekil 4.11. Çok katmanlı algılayıcı algoritması için ROC eğrisi.

4.3.4. Gradyan Artırma Makineleri Algoritması

Gradyan artırma makineleri zayıf öğrenen bireylerden yinelemeli modeller oluşturur. Gradyan artırmada, bireysel modelleri tamamen rastgele veri ve özellik alt kümeleri ile göz önüne almaktan ziyade, yanlış tahmin ve yüksek hatalara sahip örneklere öncelik vererek sırayla oluşturmaktadır. Bunun nedeni, doğru tahmin edilmesi zor olan modelin geçmiş hatalardan ders alması için öğrenme sırasında odaklanmasıdır. Bireysel modellerin öngörü gücünün zayıf ve aşırı öğrenmeye eğilimli olmasından dolayı birçok zayıf modelin bir grupta birleştirilmesi daha çok gelişmiş sonucu doğuracaktır. Gradyan artırma makinelerinde en yaygın kullanılan zayıf model türü karar ağaçlarıdır. Çalışma kapsamında, gradyan artırma algoritmasının uygulanmasında 10 model kullanılmış olup bu ağacın 10 kez yinelenmesi anlamına gelmektedir. Ağaç derinliği sayısı 10 ile sınırlı tutulmuştur. Gradyan artırma algoritmasının deney veri setine uygulanmasında kullanılan öğrenme oranı 0.1'dir. Şekil 4.12. içerisinde gradyan artırma algoritmasının deney veri setine uygulanmasında kullanılan parametrelere yer verilmiştir.



Şekil 4.12. Gradyan artırma algoritmasının deney veri setine uygulanmasında kullanılan parametreler.

Deney veri setinde kullanılan gradyan artırma algoritmasına yönelik performans ölçümü için kullanılan karmaşıklık matrisine ait değerler Tablo 4.8. içerisinde sunulmuştur.

Tablo 4.8. Deney veri setinde kullanılan gradyan artırma algoritmasının performans ölçümünden elde edilen karmaşıklık matrisine ait değerler.

Algoritma Türü	**Karmaşıklık Matrisi**		**Verinin Tahmin Edilen Sınıfı**		**Toplam Veri Sayısı**
			Reklam Gören	**Reklam Görmeyen**	
Gradyan Artırma Makineleri	**Verinin Gerçek Sınıfı**	**Reklam Gören (Display = 1)**	11.322	6.514	17.836
		Reklam Görmeyen (Display = 0)	2.743	141.014	143.757
	Toplam Veri Sayısı		14.065	147.528	161.593

Gradyan artırma algoritmasının deney veri setine uygulanması sonucu performans ölçümü için kullanılan karmaşıklık matrisi tablosundaki değerler incelendiğinde, DP değerinin 11.322, YP değerinin 2.743, YN değerinin 6.514 ve DN değerinin 141.014 olduğu görülmektedir. Doğru olarak sınıflandırılmış olan reklam gören ve reklam görmeyen sayısı 152.336'dır. 6.514 pazarlama verisi gerçekte reklam gören olup sınıflandırma sonucu reklam görmeyen, 2.743 pazarlama verisi gerçekte reklam görmeyen olup sınıflandırma sonucu reklam gören şeklinde etiketlenmektedir.

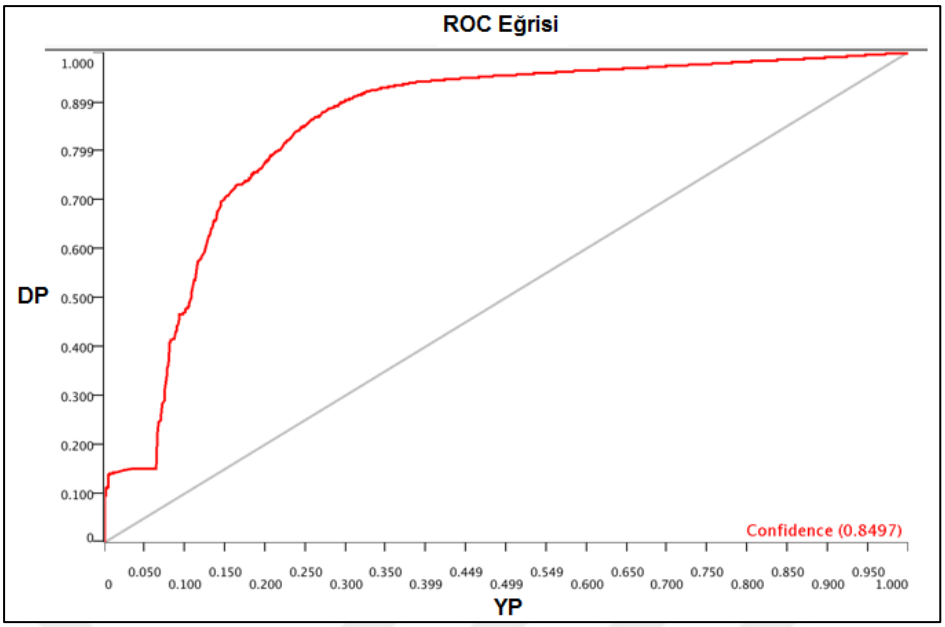
Model performansı ölçümünden gradyan artırma algoritması için elde edilen değerlere ilişkin doğruluk, duyarlılık, kesinlik, seçicilik ve F-Ölçütü oranlarına Tablo 4.9. içerisinde yer verilmiştir.

Tablo 4.9. Gradyan artırma algoritmasının model performansı ölçümünden elde edilen değerler.

Algoritma Türü	**Model Performansının Ölçümüne Ait Hesaplamalar**					
		Duyarlılık	**Kesinlik**	**Seçicilik**	**F-Ölçütü**	**Doğruluk**
Gradyan Artırma Makineleri	**Reklam Gören**	**0.981**	**0.956**	0.635	**0.968**	
	Reklam Görmeyen	0.635	0.805	**0.981**	0.710	
	Genel	0.808	0.880	0.808	0.839	**0.943**

Model performansının ölçümüne ilişkin gradyan artırma algoritmasının doğruluk değeri 0.943'tür. Gradyan artırma algoritması için elde edilen bu doğruluk değeri doğru tahmin edilen veri sayısının yüksek olduğunu göstermektedir. Model performansının ölçümüne ait duyarlılık, kesinlik, seçicilik ve F-Ölçütü değerleri sırasıyla 0.808, 0.880, 0.808 ve 0.808'dir. Model performans ölçümünden elde edilen değerler reklam gören ve reklam görmeyenler şeklinde ayrı ayrı incelendiğinde, reklam görenler için duyarlılık, kesinlik, seçicilik ve F-Ölçütü değerleri sırasıyla 0.981, 0.956, 0.635 ve 0.968 iken, reklam görmeyenler için duyarlılık, kesinlik, seçicilik ve F-Ölçütü değerleri 0.635, 0.805, 0.981 ve 0.710'dir. Reklam görenler için duyarlılık, kesinlik ve F-ölçütü metriklerinin 1'e çok yakın yüksek değerlerde olduğu, reklam görmeyenler için ise sadece seçicilik metriğinin 0.982 oranı ile 1'e çok yakın yüksek bir değerde olduğu görülmektedir.

Çok katmanlı algılayıcı algoritması için AUC kriterine ilişkin sonuç Şekil 4.13.'te sunulmuştur. Çok katmanlı algılayıcı algoritması için AUC değeri 0.8497'dir.



Şekil 4.13. Gradyan artırma algoritması için ROC eğrisi.

4.4. Model Performanslarının Karşılaştırılması

Çalışma kapsamında kullanılan dört algoritma doğruluk oranları benzer olmakla birlikte, dört algoritmanın da doğruluk değerleri 1'e çok yakın bir değerdedir. Kullanılan dört algoritma arasında en yüksek doğruluk değerine (0.943) sahip olan gradyan artırma algoritmasının en iyi model performansını sergilediği görülmektedir. Gradyan artırma algoritmasının en iyi performans değerini sırasıyla karar ağacı (0.934), rastgele orman (0.924) ve çok katmanlı algılayıcı (0.890) algoritmaları takip etmektedir.

Kullanılan dört algoritma içerisinde duyarlılık, kesinlik, seçicilik ve F-Ölçütü oranı (0.808, 0.880, 0.808, 0.839) en başarılı olan algoritma türü gradyan artırma makineleridir. Duyarlılık, kesinlik, seçicilik ve F-Ölçütü oranı gradyan artırma makinelerinden sonra en başarılı olan algoritmalar sırasıyla karar ağacı (0.790, 0.851, 0.790, 0.816), rastgele orman (0.720, 0.845, 0.720, 0.764) ve çok katmanlı algılayıcılardır (0.5, 0.445, 0.5, 0.471). Model performans ölçümünden elde edilen değerler reklam gören ve reklam görmeyenler şeklinde ayrı ayrı incelenirse, reklam görenler için duyarlılık metriği en başarılı algoritma 1 değerine sahip olan çok katmanlı algılayıcı algoritmasıdır. Duyarlılık metriği için en başarılı olan diğer algoritmalar sırasıyla rastgele orman (0.982), gradyan artırma (0.981) ve karar ağacı (0.975) algoritmasıdır. Reklam görenler için kesinlik metriği en başarılı algoritma

0.956 değerine sahip olan gradyan artırma algoritmasıdır. Kesinlik metriği için en başarılı olan diğer algoritmalar sırasıyla karar ağacı (0.952), rastgele orman (0.936) ve çok katmanlı algılayıcı (0.890) algoritmasıdır. Reklam görenler için seçicilik metriği en başarılı algoritma 0.635 değerine sahip olan gradyan artırma algoritmasıdır. Seçicilik metriği için en başarılı olan diğer algoritmalar sırasıyla karar ağacı (0.604), rastgele orman (0.458) ve çok katmanlı algılayıcı (0) algoritmasıdır. Reklam görenler için F-Ölçütü metriği en başarılı algoritma 0.968 değerine sahip olan gradyan artırma algoritmasıdır. F-Ölçütü için en başarılı olan diğer algoritmalar sırasıyla karar ağacı (0.964), rastgele orman (0.958) ve çok katmanlı algılayıcı (0.942) algoritmasıdır.

Reklam görmeyenler için duyarlılık metriği en başarılı algoritma 0.635 değerine sahip olan gradyan artırma algoritmasıdır. Duyarlılık metriği için en başarılı olan diğer algoritmalar sırasıyla karar ağacı (0.604), rastgele orman (0.458) ve çok katmanlı algılayıcı (0) algoritmasıdır. Reklam görmeyenler için kesinlik metriği en başarılı algoritma 0.805 değerine sahip olan gradyan artırma algoritmasıdır. Duyarlılık metriği için en başarılı olan diğer algoritmalar sırasıyla rastgele orman (0.755), karar ağacı (0.749) ve çok katmanlı algılayıcı (0) algoritmasıdır. Reklam görmeyenler için seçicilik metriği en başarılı algoritma 1 değerine sahip olan çok katmanlı algılayıcı algoritmasıdır. Seçicilik metriği için en başarılı olan diğer algoritmalar sırasıyla gradyan artırma makineleri (0.635), karar ağacı (0.604) ve rastgele orman (0.458) algoritmasıdır. Reklam görmeyenler için F-Ölçütü metriği en başarılı algoritma 0.710 değerine sahip olan gradyan artırma algoritmasıdır. F-Ölçütü için en başarılı olan diğer algoritmalar sırasıyla karar ağacı (0.669), rastgele orman (0.570) ve çok katmanlı algılayıcı (0) algoritmasıdır.

AUC kriterine göre, çalışmada kullanılan dört algoritma arasında (0.8857) Rastgele orman algoritmasının en iyi model performansını sergilediği görülmektedir. Rastgele orman algoritmasının en iyi performans değerini sırasıyla karar ağacı (0.8587), gradyan artırma (0.8497) ve çok katmanlı algılayıcı (0.5) algoritmaları izlemektedir.

5. SONUÇ VE ÖNERİLER

Makine öğrenimi, özellikle gelişmiş matematiksel tekniklerle birleştirildiğinde finansal planlamayı, bütçelemeyi ve öngörmeyi geliştirmek için önemli bir potansiyele sahiptir. İşletmelerin gelişmiş planlama, bütçeleme ve tahmin araçları doğru ve uygun stratejik kararlar almalarını destekleyecektir. Bir planlama, bütçeleme ve tahmin aracı türü, çok sayıda büyük veri bilgi girdisinin işlenmesine dayanan bir dizi karar için en uygun çözümü belirlemeye yönelik bir metodoloji içerir. Bir işletmenin finansal planları, bütçeleri ve tahminleri için stratejik olarak makine öğrenmesi algoritmalarını kullanarak özellikle çok sayıda belirsiz karar içeren finansal varlıkların yönetimini yaparak ekonomik faydalar elde edebilir.

Bu tez çalışmasında, Dunnhumby şirketi tarafından akademik amaçlar için 156 hafta boyunca kahvaltı ürünlerinin haftalık satışlarından oluşturulan "Frat" veri seti üzerinde yapılan satışlarda reklam değişkenin etkisi kullanılarak pazarlamanın etkin olabileceği kitlenin çeşitli makine öğrenmesi algoritmaları aracılığıyla seçilerek skorlanmıştır. Frat veri seti, dört seçilmiş kategoride (ağız bakım suyu, çubuk kraker, dondurulmuş pizza ve kahvaltılık gevrek) en çok satış yapan üç markanın her birinden en iyi beş ürün hakkında satış ve promosyon bilgilerini içermektedir. Bu veri seti, araştırmacıların "gerçek dünya" verileriyle çalışmalar yapmalarına ve tahminlerde bulunabilmelerine olanak sağladığından dolayı, pazarlamanın etkin olabileceği müşteri kitlesinin tespit edilerek bütçe verimliliğine katkı vermesi açısından ideal ve güvenilir bir kaynaktır. Uzun süreli olarak elde edilen verilerin zenginliği ve sağlamış olduğu potansiyel analizler, veri setinin akademik araştırmalar için elverişli bir araç hâline getirmektedir. Çalışma açısından bu veri setinin kullanılmasındaki en önemli etken; pazarlama için müşterilere ürünlerin gösterilip gösterilmediği hakkındaki bilgilerin toplandığı "reklam (display)" değişkenine yer verilmesidir. Nitekim, çalışma içerisinde bu değişken ana hedef olarak seçilmiş ve kurulan modeller ile veri içerisindeki örüntüler bulunmuştur. Bu sayede, gelecekteki yeni müşteri-ürün verisi için bu değişkenin durumu tahmin edilerek skorlanabilmiştir.

Bütçe verimliliğine veya planlamasına ilişkin büyük veriler ile yapılmış olan çalışma sayısı oldukça sınırlıdır. Büyük verilerin bütçe verimliliğinde önemli olduğuna vurgu yapan birkaç çalışma Ma ve arkadaşları tarafından gerçekleştirilmiştir (Ma vd., 2016). Bu çalışmaların birinde, Ma ve arkadaşları özelliklerin seçiminde dört

aşamalı bir yaklaşım önererek, çapraz ürün ve dönemler arasındaki etkileri hesaplamıştır. Bir diğer çalışmada ise, özellikler ile ilgili yaklaşımlarını çok dönemli kâr maksimizasyonu içeren kategoriler için bir optimizasyon algoritmasıyla birleştirmiştir (Ma ve Fildes, 2017). Büyük veriler ile yapılan çalışmalarda özelliklerin uygun şekilde eklenmesi veya elenmesi ile model performansı açısından yüksek doğruluk oranları elde edilebilir. Nitekim yapılan çalışmada, özelliklerin seçiminde sarmalayıcı yöntemlerden hem ileri özellik seçimi hem de geri özellik eleme ile 24 farklı özellikten ekleme ve elemeler yapılarak verilerin karmaşıklık durumu azaltılmış ve daha az özellik ile daha yüksek doğruluk elde etmek için kullanılmıştır. Diğer taraftan, makine öğrenmesi algoritması performansının değerlendirilmesinde hangi veri setinin kullanılması gerektiği durumu önemlidir. Bu bağlamda, herhangi bir makine öğrenmesi algoritmasının doğruluğunun ve güvenilirliğinin belirlenmesinde orijinal verilerin eğitim ve test veri seti şeklinde ikiye bölünmesi son derece önemlidir. Çalışmada kullanılan veri setinde yer alan 538.643 adet pazarlama verisinin %70'i (377.050 adet) modelin eğitilmesinde, geri kalan %30'u (161.593 adet) ise modelin test edilmesinde kullanılmıştır.

Çalışma için öncelikli olarak genel bir çerçeve belirlenmiştir. Belirlenen bu genel çerçeve doğrultusunda, geçmiş yıllardan toplanılan müşteri bazlı pazarlama verileri üzerinden daha önce pazarlama bütçesi için ayrılmış (kampanyaya katılan müşteriler, çağrı merkezinin aradığı ve ürün satabildiği müşteriler vb.) ve başarılı olunmuş kitle işaretlenerek, deney veri seti farklı makine öğrenmesi algoritmaları KNIME 4.2.1 programı ile analiz edilmiştir. Çalışmada algoritmaların matematiksel ve veri işleme farklılıkları göz önünde bulundurularak, veri setinde karar ağacı, rastgele orman, çok katmanlı algılayıcı ve gradyan artırma algoritmaları kullanılarak performans analizleri gerçekleştirilmiştir. Bu algoritmaların sonuçları doğruluk, duyarlılık, kesinlik, seçicilik ve F-Ölçütü oranları belirlenerek birbirleriyle karşılaştırılmış ve en iyi performansı sergileyen algoritmanın gradyan artırma makineleri olduğu görülmüş ve bu algoritmanın pazarlamanın etkin olabileceği kitlenin tespit edilmesinde kullanımı tavsiye edilmektedir.

Pazarlama bütçesinin tüm müşterilere harcanması yerine sadece harcanması gereken potansiyel müşterilerin bir yapay zeka modeli kullanılarak pazarlama bütçesinin daha verimli kullanılmasını öngören bu çalışma kapsamında karar ağacı, rastgele orman, çok katmanlı algılayıcı ve gradyan artırma algoritmalarının üst

yönetimin veya yöneticilerin pazarlama bütçesi kararlarına katkı sağlaması için geliştirilmiştir.

Gerçekleştirilen tez çalışmasının sağladığı katkılar aşağıda sunulmuştur:

- ✓ Pazarlama için ayrılacak bütçenin verimliliğinin arttırılmasına yönelik pazarlamanın etkin olabileceği müşteri kitlesinin belirlenmesi için kullanılan karar ağacı, rastgele orman, çok katmanlı algılayıcı ve gradyan artırma algoritmalarına ait model performans ölçümündeki doğruluk oranlarının birbirlerine benzer olmasının yanı sıra, 1 değerine çok yakın yüksek değerde oldukları görülmektedir. Pazarlamanın etkin olabileceği müşteri kitlesinin belirlenmesinde dört algoritma da kullanışlıdır. Ancak en iyi sınıflandırmayı yapan algoritmanın gradyan artırma makinesi algoritmasına ait olduğu tespit edilmiştir. Gradyan artırma makinelerinin ağaç temelli diğer algoritmalardan (karar ağacı ve rastgele orman) daha iyi performans göstermesi müşteri satış tepkisinin modellemesinde bu algoritmayı daha önemli ve kullanışlı bir hâle getirmektedir.

- ✓ Çalışma kapsamında pazarlama için müşterilere ürünlerin gösterilip gösterilmediği hakkındaki "reklam (display)" değişkeni temel alınarak, özellik seçiminde diğer değişkenlere ait özelliklerin ileri özellik seçimi ve geri özellik eleme yöntemleri ile doğru şekilde belirlenmesi algoritmaların deney veri setine uygulanmasında karmaşıklık durumlarını azaltmış ve daha az özellik kullanılarak yüksek doğruluk oranlarının elde edilmesini sağlamıştır. Uygulamalı analiz açısından yapılan çalışmada, değişken önem analizinden sonra özelliklerin sayısını çalışmada olduğu gibi azaltarak yüksek doğruluk değerlerinin elde edilebilmesi, en önemli ve doğru seçilmiş özelliklerden bir set kullanılarak oluşturulan bir modele dayalı analizlerin gerçekleştirilmiş olması anlamında değerlidir.

- ✓ Bu çalışma; gerçek dünya sisteminden elde edilen verilerin bütçeleme kararlarında satış tepkisinin modellemesi için özellik mühendisliği ve makine öğrenmesi algoritmalarının yararlılığını göstermektedir.

Çalışmada deney veri setinin analizi sonucu elde edilen bulgularda gradyan artırma makinelerinin ağaç temelli diğer algoritmalara kıyasla daha iyi performans

sergilemesi durumu literatürde son yıllarda elde edilen bulgular ile uyumludur (Molnar, 2018; Jang, 2019; Antipov ve Pokryshevskaya, 2020). Diğer taraftan, uygulamalı analiz açısından değişken önem analizinden sonra özelliklerin sayısının azaltılarak yüksek doğruluk değerlerinin elde edilmesi literatürde beklenen bir durumdur (Sun vd., 2008; Ali vd. 2009; Ferreira vd., 2015; Yang ve Zhang, 2018).

Tez konusu ile bağlantılı olarak ve çalışmanın sürdürülebilirliği açısından gelecekte yapılması önerilen çalışmalar aşağıda sunulmuştur:

- ✓ Deney veri setindeki veri sayısının artırılarak veya yeni veriler eklenerek sınıflandırma algoritmalarına uygulanabilir. Daha fazla sayıda veri ile bir bütçe destek sistemi modeli oluşturulabilir.
- ✓ Deney veri setine farklı sınıflandırma algoritmaları uygulanabilir.
- ✓ Deney veri setinde ek özellikler seçilerek veya eleme sayısını artırarak model performansları ölçülebilir ve mevcut performans sonuçlarıyla karşılaştırılabilir.
- ✓ Deney veri setindeki seçilen ürünlerin kategori sayısının, en çok satış yapan marka sayısının ve bu markaların en iyi ürün sayısının daha fazla sayıda tutularak sınıflandırma algoritmalarına uygulabilir.
- ✓ Deney veri setinde kullanılan algoritmaların işletmeler tarafından verilen bütçe kararları ile uyum gösterip göstermediği araştırılabilir.
- ✓ Bu çalışmanın altyapısı kullanılarak yeni algoritmalar ile zenginleştirilmiş bir bütçe karar mekanizmasının işletme gerçekleriyle uyumlu şekilde geliştirilmesi sağlanabilir.
- ✓ Reklama ilişkin gerçek giderlerin de veri setinde yer almasıyla pazarlamada bütçe verimliliğin kantitatif olarak birim cinsinden gözlenebilmesi sağlanabilir.

Satış tahmininin temel sorunlarından biri, verilerin gözlemsel niteliği nedeniyle talebin her zaman karşılandığının ve bu nedenle satışların gerçek talebi yansıttığının varsayılması gerektiğidir. Çalışma kapsamında ele alınan uzun raf ömürlü ürünler için yapılan varsayım gerçekçi olsa da, talep edilen miktar için

özellikle bozulabilen ürünlerin satılan birim sayısından daha fazla olabileceği durumu göz önünde bulundurulmalıdır (Ozhegov ve Teterina, 2018). Bu nedenle, perakendecilerin yaptıkları satışlarda mümkün olduğunca sansürlenmiş olan bu talepleri yansıtıp yansıtmadığını (en azından, her hafta bir süre için SKU'nun stokta olup olmadığını ve ne kadar süreyle stokta olmadığını takip ederek) hesaba katmaları gerekmektedir. Çalışma kapsamındaki bir diğer sınırlılık ise, seçilen ürünlerin kategori sayısının (dört), en çok satış yapan marka sayısının (üç) ve bu markaların en iyi ürün sayısının (beş) satış ve promosyon bilgilerinin az sayıda tutulması olmuştur.

Büyük verili satış ve promosyon bilgilerinin işlenerek pazarlamanın etkin olabileceği müşteri kitlesinin belirlenmesine yönelik yapılan tez çalışmasında uygulanan adımlar bütçe planlaması ve verimliliği için örnek bir çalışmadır. Makine öğrenmesi algoritmaları için model performansı ölçümünden elde edilen değerlere ilişkin doğruluk, duyarlılık, kesinlik, seçicilik ve F-Ölçütü oranlarına bakılarak değerlendirmeler yapılmıştır.

Sonuç olarak, gerçek dünya verilerinden elde edilen ve yapılan satışlar üzerinden birkaç özellik grubunun etkisi kullanılarak pazarlama için ayrılacak bütçenin verimliliğinin arttırılmasına yönelik tez çalışmasında en iyi tahminin yapıldığı sınıflandırma algoritmasının belirlenerek veri bilimine katkı sağlanması ve rehberlik etmesi amaçlanmıştır. Yapılan çalışmanın altyapısının daha da geliştirildiği bir modelin işletmeler tarafından kullanılarak iş dünyasına katkı sağlanması en büyük dileğimizdir.

KAYNAKÇA

Ailawadi, K. L., Harlam, B. A., César, J. ve Trounce, D. (2007). Practice Prize Report: Quantifying and Improving Promotion Effectiveness at CVS. Marketing Science, 26, 566-575.

Ali, Ö. G., Sayın, S., Van Woensel, T. ve Fransoo, J. (2009). SKU Demand Forecasting in the Presence of Promotions. Expert Systems with Applications, 36, 12340-12348.

Amit, Y. ve Geman, D. (1997). Shape quantization and recognition with randomized trees. Neural Computation. 9(7), 1545-1588.

Andrews, R. L., Currim, I. S., Leeflang, P. ve Lim, J. (2008). Estimating the SCAN*PRO Model of Store Sales: HB, FM or just OLS? International Journal of Research in Marketing, 25, 22-33.

Antipov, E. A. ve Pokryshevskaya, E. B. (2020). Interpretable machine learning for demand modeling with high-dimensional data using Gradient Boosting Machines and Shapley values. Journal of Revenue and Pricing Management, 1-10.

Bajari, P., Nekipelov, D., Ryan, S. P. ve Yang. M. (2015). Machine Learning Methods for Demand Estimation. The American Economic Review, 105, 481-485.

Ben-Gal, I., Dana, A., Shkolnik, N. ve Singer, G. (2014). Efficient Construction of Decision Trees by the Dual Information Distance Method. Quality Technology & Quantitative Management. 11(1), 133-147.

Bengio, Y., Courville, A. ve Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.

Berthold, M. R., Cebron, N., Dill, F., Di Fatta, G., Gabriel, T. R., Georg, F., Meinl, T., Ohl, P., Sieb, C. ve Wiswedel, B. (2006). KNIME: the Konstanz Information Miner. In: Workshop on Multi-Agent Systems and Simulation (MAS&S), 4th Annual Industrial Simulation Conference (ISC), 05-07 June 2006, Palermo, İtalya, 58-61.

Bradlow, E. T., Gangwar, M., Kopalle, P. ve Voleti, S. (2017). The Role of Big Data and Predictive Analytics in Retailing. Journal of Retailing, 93, 79-95.

Bramer, M. (2007). Principles of Data Mining. Springer, London, UK.

Breiman, L. (1997). Arcing The Edge. Technical Report 486. Statistics Department, University of California, Berkeley.

Breiman, L. (2001). Random Forests. Machine Learning. 45(1), 5-32.

Brotherton, T., Jahns, G., Jacobs, J. ve Wroblewski, Dariusz. (2000). Prognosis of Faults in Gas Turbine Engines. In Aerospace Conference Proceedings, 6, 163-171.

Brownlee, J. (2016). Crash Course On Multi-Layer Perceptron Neural Networks. Machine Learning Mastery, https://machinelearningmastery.com/neural-networks-crash-course/, Erişim Tarihi: 19/09/2020.

Camargo, L. S. ve Yoneyama, T. (2001). Specification of Training Sets and the Number of Hidden Neurons for Multilayer Perceptrons. Neural Computation, 13, 2673-2680.

Castellano, G., Fanelli, A., ve Pelillo, M. (1997). An iterative pruning algorithm for feedforward neural networks. IEEE Transactions on Neural Networks, 8, 519-531.

Chawla, N. V., Japkowicz, N. ve Kotcz, A. (2004). Special Issue on Learning from Imbalanced Data Sets. Sigkdd Explorations, 6(1), 1-6.

Chen, K. Y., Chen, L. S., Chen M. C. ve Lee, C. L. (2011). Using SVM Based Method for Equipment Fault Detection in A Thermal Power Plant. Computers in Industry, 62(1), 42-50.

Cui, G., Wong, M. L. ve Lui, H.-K. (2006). Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. Management Science, 52(4), 597-612.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function Mathematics of Control, Signals, and Systems, 2(4), 303-314.

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

Dunnhumby (2019). A Time Series Analysis: Breakfast at the Frat. https://www.dunnhumby.com/source-files/, Erişim Tarihi: 01.08.2020.

Ferreira, K. J., Lee, B. H. A. ve Simchi-Levi, D. (2015). Analytics for An Online Retailer: Demand Forecasting and Price Optimization. Manufacturing & Service Operations Management, 18, 69-88.

Friedman, J. H. (1999). Greedy Function Approximation: A Gradient Boosting Machine. Technical Report, Deparment of Statistics, Stanford University.

Gadaleta, F. (2019). Entropy in Machine Learning. Amethix. https://amethix.com/entropy-in-machine-learning/, Erişim Tarihi: 06/09/2020.

Gareth, J., Witten, Hastie, D.T. ve Tibshirani, R. (2015). An Introduction to Statistical Learning. Springer, New York, 315.

Gedenk, K. (2018). Retailer promotions. In Handbook of Research on Retailing, ed. K. Gedenk. Cheltenham: Edward Elgar Publishing.

Gray, M. L. ve Suri, S. (2019). Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. Houghton Mifflin Harcourt, 7.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. ve Tatham., R. L. (2006). Multivariate Data Analysis. Pearson Prentice Hall Upper Saddle River, Volume 6.

Han, J., Kamber, M. ve Pei, J. (2012). Data Mining Concepts and Techniques (3rd Ed.), Morgan Kaufmann Publishers, San Francisco, CA, USA.

Hastie, T., Tibshirani, R. ve Friedman, J. (2001). The Elements of Statistical Learning. Springer, 269-272.

Hastie, T., Tibshirani, R. ve Friedman, J. (2008). The Elements of Statistical Learning (2nd ed.). Springer.

Hastie, T., Tibshirani, R. ve Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.

Ho, T.K. (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 278-282.

Ho, T.K. (1998). The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 20(8), 832-844.

Ho, T.K. (2002). A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors. Pattern Analysis & Applications 5, 102-112.

Hu, X., Bhanu, N., Flores, E. L., Prateek, S. ve Macarena, R. L. (2019). Steering AI and advanced ICTs for knowledge societies: a Rights, Openness, Access, and Multi-stakeholder Perspective. UNESCO Publishing, 64.

Hyafil, L. ve Rivest, R. L. (1976). Constructing Optimal Binary Decision Trees is NP-complete. Information Processing Letters. 5 (1), 15-17.

Jahnke, P. (2015). Machine Learning Approaches for Failure Type Detection and Predictive Maintenance. Master Thesis. Department of Computer Science, Technical University of Darmstadt, 83.

Jang, H. (2019). A Decision Support Framework for Robust R&D Budget Allocation Using Machine Learning and Optimization. Decision Support Systems. 121, 1-12.

Kleinberg, E. (1990). Stochastic Discrimination. Annals of Mathematics and Artificial Intelligence. 1(1–4), 207-239.

Klimek, L. (2020). Simple and Foolproof ways to Shrink, Compress, and Accelerate your Deep Learning, Neural Network, etc. Artificial Intelligence Models. PiPrograming, https://piprogramming.org/articles/Simple-and-Foolproof-ways-to-Shrink-Compress-and-Accelerate-your-Deep-Learning-Neural-Network-etc-Artificial-Intelligence-Models-0000000015.html, Erişim Tarihi: 16/09/2020.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica, 31, 249-268.

Leif, J. (2013). What is the difference between labeled and unlabeled data?, Stack Overflow, https://stackoverflow.com/questions/19170603/what-is-the-difference-between-labeled-and-unlabeled-data/19172720#19172720, Erişim Tarihi: 07/09/2020.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. ve Liu, H. (2017). Feature Selection: A Data Perspective. ACM Computing Surveys (CSUR), 50(6), 94:1-45.

Li, Z. (2018). Bayesian Methods for Graphical Models with Limited Data (Doctoral dissertation).

Liu Q. ve Wu Y. (2012). Supervised Learning. In: Seel N.M. (eds) Encyclopedia of the Sciences of Learning. Springer, Boston, MA.

Liu, X. Y., Wu, J. ve Zhou, Z.H., 2009, Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, 39(2), 539-550.

Ma, S. ve Fildes, R. (2017). A Retail Store SKU Promotions Optimization Model for Category Multi-Period Profit Maximization. European Journal of Operational Research, 260, 680-692.

Ma, S., Fildes, R. ve Huang, T. (2016). Demand Forecasting with High Dimensional Data: The Case of SKU Retail Sales Forecasting with Intra-and Inter-Category Promotional Information. European Journal of Operational Research, 249, 245-257.

Mason, L., Baxter, J., Bartlett, P. L. ve Frean, M. (1999a). Boosting Algorithms as Gradient Descent. In S.A. Solla and T.K. Leen and K. Müller (ed.). Advances in Neural Information Processing Systems 12. MIT Press, 512-518.

Mason, L., Baxter, J., Bartlett, P. L. ve Frean, M. (1999b). Boosting Algorithms as Gradient Descent in Function Space. Technical report, RSISE, Australian National University, 1999.

Mehtaa, D. ve Raghavan, V. (2002). Decision tree approximations of Boolean functions. Theoretical Computer Science. 270(1–2): 609-623.

Molnar, C. (2018). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub.

Murphy, K. (2012). Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA.

Neocleous, C. ve Schizas, C. (2002). Artificial Neural Network Learning: A Comparative Review, Springer-Verlag Berlin Heidelberg, 300-313.

Ozhegov, E. ve Teterina D. (2018). The ensemble method for censored demand prediction. High. Sch. Econ. Res. Pap. No. WP BRP 200.

Parekh, R., Yang, J. ve Honavar, V. (2000). Constructive Neural Network Learning Algorithms for Pattern Classification. IEEE Transactions on Neural Networks, 11(2), 436-451.

Peng, Y., Dong, M. ve Zuo, M. J. (2010). Current Status of Machine Prognostics in Condition-Based Maintenance: A Review. The International Journal of Advanced Manufacturing Technology, 50(1-4), 297-313.

Peres, D. J., Iuppa, C., Cavallaro, L., Cancelliere, A. ve Foti, E. (2015). Significant wave height record extension by neural networks and reanalysis wind data. Ocean Modelling. 94: 128-140.

Piryonesi, S. M. ve El-Diraby T. E. (2020a). Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index. Journal of Infrastructure Systems. 26(1), 04019036.

Piryonesi, S. M. ve El-Diraby T. E. (2020b). Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. Journal of Transportation Engineering, Part B: Pavements. 146(2), 04020022.

Powers, D. M. W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001. School of Informatics and Engineering, Flinders University, Adelaide, Australia.

Preiswerk, F. (2018). Shannon Entropy in the Context of Machine Learning and AI. https://medium.com/swlh/shannon-entropy-in-the-context-of-machine-learning-and-ai-24aee2709e32, Erişim Tarihi: 06/09/2020.

Provost, F. 1964-(2013). Data science for business : [what you need to know about data mining and data-analytic thinking]. Fawcett, Tom. (1st ed.). Sebastopol, Calif.: O'Reilly.

Rokach, L. ve Maimon, O. (2005). Top-down Induction of Decision Trees Classifiers-A Survey. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews. 35 (4), 476-487.

Rosenblatt, F. (1961). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC.

Rouse, M. (2019). Data Labeling. TechTarget, https://whatis.techtarget.com/definition/data-labeling, Erişim Tarihi: 07/09/2020.

Rumelhart, D.E., Geoffrey, E.H. ve Williams, R. J. (1986). Learning Internal Representations by Error Propagation. David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the Microstructure of Cognition, Volume 1: Foundation. MIT Press.

Sezer, Ü. (2008). Karar Ağaçlarının Birliktelik Kuralları ile İyileştirilmesi, Yüksek Lisans Tezi, Kocaeli Üniversitesi Bilgisayar Mühendisliği, Kocaeli, 29-30.

Shannon, C. E. (2001). A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 5(1), 3-55.

Shmueli, G. (2010). To explain or to predict?. Statistical Science, 25(3), 289-310.

Siddique, M. N. H. ve Tokhi, M. O. (2001). Training Neural Networks: Backpropagation vs. Genetic Algorithms. IEEE International Joint Conference on Neural Networks, 4, 2673-2678.

Sun, Z.-L., Choi, T.-M., Au, K.-F. ve Yu., Y. (2008). Sales Forecasting Using Extreme Learning Machine with Applications in Fashion Retailing. Decision Support Systems, 46, 411-419.

Tareq, S. U., Noor, M. H. ve Bepery, C. (2020). Framework of dynamic recommendation system for e-shopping. International Journal of Information Technology, 12(1), 135-140.

Tu, P. Y. L., Yam, R., Tse, P. W. ve Sun, A. O. W. (2001). An Integrated Maintenance Management System for An Advanced Manufacturing Company. The International Journal of Advanced Manufacturing Technology, 17(9), 692-703.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspective, 28, 3-27.

Weigend, A. S., Rumelhart, D. E. ve Huberman, B. A. (1991). Generalization by weight-elimination with application to forecasting. In: R. P. Lippmann, J. Moody, & D. S. Touretzky (eds.), Advances in Neural Information Processing Systems 3, San Mateo, CA: Morgan Kaufmann.

Witten, I. H. ve Frank, E. (2016). Data Mining: Practical machine learning tools and techniques (2nd Ed.), Morgan Kaufmann Publishers, San Francisco, CA, USA.

Wu, X., Kumar, V., Quinlan, R.J.; Ghosh, J. Yang, Q. Motoda, H., McLachlan, G.J.; Ng, A., Liu, B., Yu, P. S.; Zhou, Z.H. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems. 14(1), 1-37.

Yam, J. ve Chow, W. (2001). Feedforward Networks Training Speed Enhancement by Optimal Initialization of the Synaptic Coefficients. IEEE Transactions on Neural Networks, 12, 430-434.

Yang, D. ve Zhang, A. N. (2018). Forecast UPC-Level FMCG Demand, Part IV: Statistical Ensemble. In: 2018 IEEE International Conference on Big Data (Big Data), 3180-3185.

Yen, G. G. ve Lu, H. (2000). Hierarchical genetic algorithm based neural network design. In: IEEE Symposium on Combinations of Evolutionary Computation and Neural Networks, 168-175.

Yu, F.W., Ho, W.T., Chan, K.T. ve Sit, R.K.Y. (2018). Logistic Regression-Based Optimal Control for Aircooled Chiller, International Journal of Refrigeration, 85, 200-212.

Yu, F.W., Ho, W.T., Ho, Chan, K.T., Sit ve Chan, R.K.Y. (2018) Logistic regression-based optimal control for aircooled Chiller, International journal of refrigeration, 85, 200-212.

Zabokrtsky, Z. (2015). Feature Engineering in Machine Learning.

Zadeh, L. A. (1965). Fuzzy Sets. Information and Control, 8(3), 338-353.

Zheng, A. ve Casari, A. (2018). Feature Engineering for Machine Learning. O'Reilly Media, Inc., Sebastopol, CA, USA.

Zhou, Z. J., Hu, C. H., Zhang, B. C., Xu, D. L. ve Chen, Y. W. (2013). Hidden Behavior Prediction of Complex Systems Based on Hybrid Information. IEEE Transactions on Cybernetics, 43(2), 402-411.

Zhu, R., Zeng, D. ve Kosorok, M.R. (2015). Reinforcement Learning Trees. Journal of the American Statistical Association. 110(512), 1770-1784.