ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

AR-GE VE TASARIM MERKEZLERİNİN İSTATİSTİKSEL OLARAK DEĞERLENDİRİLMESİ: VERİ MADENCİLİĞİ YÖNTEMLERİ İLE HİBRİT KARAR VERME

Gözde ULU METİN

İSTATİSTİK ANABİLİM DALI

Ankara
2024

ÖZET

Doktora Tezi

AR-GE VE TASARIM MERKEZLERİNİN İSTATİSTİKSEL OLARAK DEĞERLENDİRİLMESİ: VERİ MADENCİLİĞİ YÖNTEMLERİ İLE HİBRİT KARAR VERME

Gözde ULU METİN

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Prof. Dr. Özlem TÜRKŞEN

Veri-bilgi keşfi sürecinde işlenmiş veriden anlamlı bilgilerin elde edilmesi amacıyla veri madenciliği yöntemleri uygulanır. Elde edilen bilgilere dayalı kararların alınması stratejik karar verme süreçlerinde önemli bir role sahiptir. Bu tez çalışmasında, karar verme süreçlerine istatistiksel bakış açısı kazandırmak amacıyla veri madenciliği sınıflandırma yöntemleri ile hibrit karar verme yaklaşımı önerilmiştir. Çalışmada yapılan tüm analizler için Python 3.11.3 programı ve kütüphaneleri kullanılmıştır. İşlenmiş veri setindeki değişkenlerin farklı önem derecelerine sahip olduğu göz önünde bulundurularak değişkenler Analitik Hiyerarşi Süreci (AHP) yöntemiyle subjektif olarak, ENTROPY ve CRITIC yöntemleriyle objektif olarak ağırlıklandırılmıştır. Ağırlıklandırılmış veri setine veri madenciliği sınıflandırma yöntemlerinden Lojistik Regresyon (Logistic Regression-LR) analizi, *k*-En Yakın Komşu (*k*-Nearest Neighbour-*k*NN) algoritması, Destek Vektör Makineleri (Support Vector Machines-SVM) ve Rastgele Orman (Random Forest-RF) algoritması uygulanmıştır. Sınıflandırma yöntemlerinin performansı, 5-kat çapraz doğrulama ile elde edilen doğruluk, kesinlik, duyarlılık, F_1-Skor ve AUC performans ölçütleri kullanılarak hesaplanmıştır. Elde edilen değerlere göre sınıflandırma yöntemlerinin tercih sıralaması, çok ölçütlü karar verme yöntemleri (TOPSIS, COPRAS, CODAS, EDAS, MABAC ve GRA) ile belirlenmiştir. Çalışmada önerilen hibrit karar verme yaklaşımının uygulanabilirliğinin gösterilmesi amacıyla Ar-Ge ve Tasarım merkezlerine ait gerçek veri seti kullanılmıştır. Üç farklı yöntem ile ağırlıklandırılmış gerçek veri seti için RF algoritmasının öncelikli olarak tercih edilebilir olduğu sonucuna ulaşılmıştır. Sonuçların genelleştirilebilmesi amacıyla çeşitli senaryolar ile simülasyon çalışması yapılmıştır. Simülasyon ile üretilen veri setlerinde gerçek veri setine benzer olarak değişkenler, subjektif olarak ağırlıklandırılıp uygulamalar yapılmıştır. Önerilen hibrit karar verme yaklaşımına göre, hem gerçek veri seti hem de simülasyon uygulamaları için RF algoritmasının diğer veri madenciliği yöntemlerinden daha iyi sınıflandırma performansına sahip olduğu görülmüştür.

Haziran 2024, 115 sayfa

Anahtar Kelimeler: Veri madenciliği sınıflandırma yöntemleri, Performans ölçütleri, Çok ölçütlü karar verme, Hibrit karar verme

ABSTRACT

PhD Thesis

STATISTICAL ASSESSMENT OF R&D AND DESIGN CENTERS: HYBRID DECISION MAKING WITH DATA MINING METHODS

Gözde ULU METİN

Ankara University
Graduate School of Natural and Applied Sciences
Department of Statistics

Supervisor: Prof. Dr. Özlem TÜRKŞEN

In the data-information discovery process, data mining methods are applied to obtain meaningful information from the processed data. Making decisions based on the obtained information has an important role in strategic decision-making processes. In this thesis, a hybrid decision-making approach is proposed with data mining classification methods in order to bring a statistical perspective to decision-making processes. Python 3.11.3 program and libraries were used for all analyses in the study. Given that the variables in the processed data set exhibit varying degrees of importance, the variables were subjectively weighted using the Analytical Hierarchy Process (AHP) method and objectively weighted using the ENTROPY and CRITIC methods. The weighted data set was analyzed using the Logistic Regression (LR), *k*-Nearest Neighbour (*k*NN), Support Vector Machines (SVM) and Random Forest (RF) algorithms. The performance of the classification methods was calculated using the accuracy, precision, sensitivity, F_1-Score and AUC performance measures obtained by 5-fold cross-validation. The preference ranking of the classification methods was determined by multi-criteria decision-making methods (TOPSIS, COPRAS, CODAS, EDAS, MABAC and GRA). In order to demonstrate the applicability of the hybrid decision-making approach proposed in the study, a real data set of R&D and Design centres was used. It was concluded that the RF algorithm is preferable for the real data set, weighted with three different methods.. In order to generalise the results, simulation study were carried out with various scenarios. In the data sets generated by the simulation, similar to the real data set, the variables were subjectively weighted and applications were made. According to the proposed hybrid decision approach, it was observed that RF algorithm has better classification performance than other data mining methods for both real data set and simulation applications.

June 2024, 115 pages

Key Words: Data mining classification methods, Performance metrics, Multi-criteria decision making, Hybrid classification

TEŞEKKÜR

Doktora eğitimim boyunca bana yol gösteren, bilgi ve tecrübesiyle her zaman ilham kaynağım olan, değerli öğütleri ve yapıcı eleştirileri ile gelişimime katkıda bulunan, hem akademik hem de mesleki alanda başarıya ulaşmam için desteğini hiçbir zaman esirgemeyen, sadece bir danışman olarak değil aynı zamanda iyi bir yol arkadaşı olarak yanımda olan, övgüye değer emekleri ile büyük katkı sağlayan tez danışmanım Sayın Prof. Dr. Özlem TÜRKŞEN'e (Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü) içten teşekkürlerimi sunarım.

Tez izleme sürecimde deneyimleri ve değerli bilgileriyle bana rehberlik eden Sayın Prof. Dr. Halil AYDOĞDU'ya (Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü), bu sürecin daha verimli ve öğretici geçmesi için bana değerli zamanını ayırarak uygulama alanındaki destekleriyle büyük katkı sağlayan Sayın Prof. Dr. Hülya OLMUŞ'a (Gazi Üniversitesi Fen Fakültesi İstatistik Bölümü), kıymetli görüşleri ve önerileriyle teze katkıda bulunan Sayın Prof. Dr. Cemal ATAKAN'a (Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü) ve Sayın Prof. Dr. Nimet YAPICI PEHLİVAN'a (Selçuk Üniversitesi Fen Fakültesi İstatistik Bölümü) ilgi ve yardımları için çok teşekkür ederim.

Tez çalışmasının uygulama kısmında kullanılan veri setinin temin edilmesindeki katkılarından ötürü Sanayi ve Teknoloji Bakanlığı Ar-Ge Teşvikleri Genel Müdürlüğü'ndeki değerli yöneticilerime, çalışmam boyunca her zaman yanımda olan Ar-Ge ve Tasarım Merkezleri Dairesi Başkanlığı'ndaki çalışma arkadaşlarıma destekleri konusunda teşekkürü bir borç bilirim.

Hayatım boyunca beni destekleyen, yorucu çalışmalarım boyunca motivasyonumu hep yükselten, doktora tez sürecinde gösterdikleri ilgi ve hoşgörülerinden dolayı anneme, babama ve kardeşlerime sonsuz teşekkür ederim.

Eğitim ve öğrenme konusundaki tüm çabalarımı koşulsuz destekleyen, beni her daim motive eden, bu uzun ve zorlu yolculuk boyunca sabır ve anlayış gösteren sevgili eşim Umutcan METİN'e tüm kalbimle teşekkür ederim.

Gözde ULU METİN
Ankara, Haziran 2024

İÇİNDEKİLER

KISALTMALAR DİZİNİ

AHP	Analitik Hiyerarşi Prosesi
Ar-Ge	Araştırma Geliştirme
AUC	Area Under the Curve
CV	Cross Validation (Çapraz Doğrulama)
CODAS	Combinative Distance based Assesment
COPRAS	Complex Proportional Assessment
CRITIC	The Criteria Importance Through Intercriteria Correlation
EDAS	Evaluation based on Distance from Average Solution
GRA	Grey Relationship Analysis
KMO	Kaiser Meyer Olkin
*k*NN	*k*- Nearest Neighbour (*k*-En Yakın Komşu)
LR	Lojistik Regresyon
MABAC	Multi Attributive Border Approximation Area Comparison
MCDM	Multi Criteria Decision Making (Çok Ölçütlü Karar Verme)
ML	Machine Learning (Makine Öğrenimi)
RF	Random Forest (Rastgele Orman)
ROC	Receiver Operator Characteristic Curve
SVM	Support Vector Machines (Destek Vektör Makineleri)
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ VE ÖNCEKİ ÇALIŞMALAR

Son yıllarda bilgi teknolojisinin gelişimi ile verinin depolanma kapasitesi artarak büyük hacimli veri yığınları oluşmuştur. Teknolojinin avantajları sayesinde elde edilen veriler kolaylıkla depolanmaktadır. Depolanan verilerin geleneksel yöntemlerle analiz edilmesi oldukça zordur. Ham halde depolanan verilerin işlenerek analize hazır hale getirilmesi gerekir.

Veri-bilgi keşfi sürecinde veri ön işleme olarak tanımlanan ham verinin işlenmesi aşaması veri analizi öncesi oldukça kritik öneme sahiptir. Veri ön işlemede, betimsel istatistikler ve veri görselleştirme kullanılarak veri seti hakkında genel bir bilgi elde edilir. Farklı ölçme düzeylerine sahip değişkenler (öznitelik, özellik) içeren veri setinde, veri temizleme, veri bütünleştirme, veri seçme ve veri dönüşümü işlemleri yapılarak temiz, eksiksiz veri seti oluşturulur.

Veri setindeki çok sayıdaki değişkenin anlamlı biçimde daha az sayıda değişkene indirgenerek, değişkenler arasında örtülü biçimde bulunan ilişki yapısının belirlenmesinde çok değişkenli istatistiksel yöntemlerden yararlanılır. Temel Bileşenler Analizi, Faktör Analizi, Diskriminat Analizi ve Kümeleme Analizi yaygın biçimde kullanılan çok değişkenli istatistiksel yöntemlerdir. Temel Bileşenler Analizi, veri setindeki değişkenleri en iyi biçimde temsil eden daha az sayıda yeni temel bileşenler oluşturarak veri boyutunu azaltan bir yöntemdir. Faktör Analizi, veri setindeki değişkenler arasındaki ilişkiyi açıklamayı ve ilişkili değişkenleri faktörlerde toplayarak boyut indirgemeyi amaçlayan çok değişkenli bir istatistiksel yöntemdir. Diskriminant Analizi, bir gözlemin sahip olduğu özelliklere göre doğru sınıfa atanmasında ayırma fonksiyonu kullanan çok değişkenli istatistiksel bir yöntemdir. Kümeleme Analizi ise, uzaklık ölçüsü yardımıyla veri setindeki gözlemleri birbirine olan benzerliklerine göre kümelere ayırarak her bir kümenin içindeki gözlemlerin benzerlik düzeyini artırmayı ve kümeler arasındaki farklılıkları belirginleştirmeyi amaçlayan bir yapı oluşturur.

Veri setindeki değişkenlerin veri analizi sonuçlarına etkisi farklılık gösterebilir. Her bir değişken, veri seti içerisinde farklı bir önem derecesine ve ağırlığa sahip olabilir.

Değişkenlerin gerçek önem değerlerini dikkate alarak yapılan veri analizi sonucu daha anlamlı bilgiler elde edilir. Veri setindeki değişkenlerin önem ağırlıklarının belirlenmesinde Çok Ölçütlü Karar Verme (Multi Criteria Decision Making-MCDM) yöntemlerinden yararlanılır. Bu yöntemler, ağırlıklandırma ve alternatiflerin sıralanması başlıkları altında değerlendirilebilir. Literatürde ağırlıklandırma sürecinde kullanılan çeşitli MCDM yöntemleri mevcuttur. Örneğin, AHP, ENTROPY, CRITIC, SWARA vb. Ağırlıklandırmada kullanılan bu yöntemler subjektif ya da objektif olabilir. Her iki yaklaşım kullanılarak değişkenler ağırlıklandırıldığında, veriden daha anlamlı bilgi elde edilmesi sağlanır. Veri ön işlemede ham veriye uygulanan keşfedici veri analizleri sonucu işlenmiş veri elde edilir. Bu aşamadan sonra veri madenciliği yöntemleri ile veri-bilgi keşfi sürecine devam edilir.

Veri madenciliği yöntemleri, veriden anlamlı bilginin elde edilmesine imkân sağlayan denetimli ve denetimsiz öğrenmeden oluşan istatistiksel, analitik ve algoritmik yöntemler içerir. Veri setinde bağımsız (açıklayıcı) değişkenlerle birlikte en az bir bağımlı (açıklanan) değişken varsa veri setinin eğitilmesi için denetimli öğrenme algoritmaları kullanılır. Denetimli öğrenme modelinde eğitim verisi kullanılarak yeni gelecek gözlemler için tahmin ya da sınıflandırma yapılır. Denetimli öğrenme, regresyon ve sınıflandırma yöntemlerini kapsar. Denetimsiz öğrenme de sadece bağımsız değişkenlerin olduğu veri setlerine uygulanan bir öğrenme türüdür. Denetimsiz öğrenme modelinde, gözlemler arasındaki yapı ve örüntü keşfedilir. Aynı gözlem değerine sahip birimler kullanılarak benzer gruplar bir araya getirilir. Kümeleme ve birliktelik kuralları, denetimsiz öğrenme yöntemleridir.

Çeşitli sınıflandırma yöntemlerini kapsayan denetimli öğrenme, gerçek dünyada karşılaşılan karmaşık sınıflandırma problemlerinin çözümünde veri madenciliği sınıflandırma yöntemlerinden Lojistik Regresyon (Logistic Regression-LR) analizi, *k*-En Yakın Komşu (*k*-Nearest Neighbour-*k*NN) algoritması, Karar Ağaçları (Decision Tree-DT), Destek Vektör Makineleri (Support Vector Machines-SVM), Rastgele Orman (Random Forest-RF) algoritması, Yapay Sinir Ağları (Artificial Neural Network-ANN), Regresyon Ağaçları (Regression Tree), Torbalama (Bagging) ve XG Boost (XGB)

kullanılır. Uygulanan yöntemler sonucu oluşturulan modeller yorumlanarak bilgi keşfi süreci tamamlanır.

Veri madenciliği yöntemlerinin sınıflandırma başarısının değerlendirilmesinde sınıflandırma performans ölçütleri dikkate alınır. Performans ölçütlerinin hesaplanmasında öncelikle veri seti; eğitim, doğrulama ve test seti olmak üzere parçalanır. Çapraz doğrulama (Cross Validation-CV), model performansının değerlendirilmesinde ve modelin genellenebilirliğinin ölçülmesinde uygulanılan bir yöntemdir. CV'de eğitim seti, gruplardaki gözlem sayısı eşit olacak biçimde k sayıda alt gruba ayrılır. $k-1$ tane grup modelin eğitilmesi, kalan grup ise modelin doğrulanması için kullanılır. k kez yapılan bu işlem, k-kat CV olarak adlandırılır. Böylece her alt grup en az bir kez doğrulama verisi olarak kullanılır. k-kat CV ile tek bir veri setine bağlı kalmadan modelin geliştirilmesi ve daha iyi performans ölçütlerinin elde edilmesi sağlanır. Model performansının artırılması amacıyla yardımcı ayarlanabilir parametreler kullanılır. Her bir sınıflandırma yöntemi kendine özgü ayarlanabilir parametreler içerir. Bu ayarlanabilir parametrelerin belirlenmesi verinin doğru sınıflandırılabilmesinde büyük önem taşır. Ayarlanabilir parametrelerle modelin, eğitim veri setine daha iyi uyarlanması sağlanır. Ayarlanabilir parametreler araştırmacı tarafından belirlenir. Belirlenen ayarlanabilir parametreler, ızgara arama ile optimize edilir. Izgara arama ile ayarlanabilir parametreler kümesinde tüm olası parametreler kombinasyonları denenerek en iyi kombinasyonun elde edilmesi amaçlanır. Performansın doğru bir biçimde değerlendirilebilmesi için eğitim setinden bağımsız test veri seti kullanılarak performans ölçütleri hesaplanır. Doğruluk, kesinlik, duyarlılık, F_1-Skor ve AUC sınıflandırma yöntemlerinin başarısını ölçmek için en çok uygulanan performans ölçütleridir.

Hesaplanan performans değerleri göz önünde bulundurularak en uygun sınıflandırma yöntemine karar verilmesi gerekir. Çok sayıda ölçüt ve yöntemleri değerlendirmek, sıralamak ya da seçmek amacıyla MCDM kullanılır. Sınıflandırma performans sonuçları bir karar matrisi olarak ele alınarak MCDM yöntemleri ile karar verme süreci daha sistematik bir biçime dönüştürülür. Bu aşamada uygulanacak MCDM yöntemleri problemin yapısı ve araştırmacının tercihine göre seçilir. Farklı tercihlere göre seçilen MCDM yöntemleri karar matrisinin normalleştirilmesine (ölçeklendirilmesine) göre üç

yaklaşımla değerlendirilebilir: (*i*) Toplama dayalı; (*ii*) Orana dayalı ve (*iii*) Min-max değerine dayalı. Toplama dayalı, TOPSIS (Technique for Order Preference by Similarity), COPRAS (Complex Proportional Assessment), ARAS (Additive Ratio Assessment) ve MOORA (Multi-Objective Optimization on the basis of Ratio Analysis); Orana dayalı (EDAS (Evaluation based on Distance from Average Solution), CODAS (Combinative Distance-based Assesment)) ve WASPAS (Weighted Aggregated Sum Product Assessment); min-max değerine dayalı MABAC (Multi-Attributive Border Approximation Area Comparison), GRA (Grey Relationship Analysis)) ve MAUT (Multi Attribute Utility Theory) normalleştirme yaklaşımı kullanılan MCDM yöntemlerindendir.

Veri madenciliği yöntemleri ile ilgili teorik ve farklı alanlardaki uygulamaları içeren önceki çalışmalar çizelge 1.1'de özetlenmiştir.

Çizelge 1.1 Veri madenciliği yöntemleri ile ilgili teorik ve farklı alanlardaki uygulamalar

Yazar ve Yıl	***Yöntem***	***Amaç***
Frawley vd. (1992)	Veri-bilgi keşfi süreci	Veri-bilgi keşfi kavramını ilk defa ortaya koymuşlardır. Ayrıca, veri tabanı kavramı, keşif yöntemleri, keşfedilen bilginin biçimi ve kullanımı ile ilgili bir çerçeve sunmuşlardır.
Fayyad vd. (1996)	Veri-bilgi keşfi süreci	Veri-bilgi keşfi sürecini geliştirmişlerdir. Bu süreci, düşük kalitedeki verilerden üst düzey bilginin çıkarım süreci olarak tanımlamışlardır.
Li ve Ruan (2007)	Veri-bilgi keşfi süreci	Veri-bilgi keşfi sürecini, verinin depolanması, erişilmesi, ölçeklendirilmesi, görselleştirilmesi ve insan-makine etkileşimi ile modellenmesini kapsayan büyük hacimli veriden bilgi elde edilmesine kadar olan süreç olarak tanımlamışlardır.
Gao vd. (2016)	Veri-bilgi keşfi süreci	Veri-bilgi keşfi sürecini, veriden yeni, potansiyel olarak yararlı ve anlaşılabilir bilgi elde edilmesi olarak ifade etmişlerdir. Bilgi keşfini, veri madenciliği tekniklerinin kullanımına ve uzman rehberliğine ihtiyaç duyan çok yönlü bir süreç olarak ifade etmişlerdir.
Romero vd. (2013)	Veri ön işleme	Veri ön işlemenin tüm veri madenciliği yöntemleri için gerekli olduğu fakat veri ön işleme adımlarının bilinen uygulanma sırasının araştırmacının seçimi ve problemin amacına göre değiştirilebileceğini önermişlerdir.

Çizelge 1.1 Veri madenciliği yöntemleri ile ilgili teorik ve farklı alanlardaki uygulamalar (devam)

Çınar ve Silahtaroğlu (2012)	Veri ön işleme	Çalışmalarında bir anket veri seti üzerinde veri madenciliği yöntemlerini uygulayarak gizli kalmış örüntüleri ve nedenlerini keşfetmişlerdir.
Bilalli vd. (2019)	Veri ön işleme	Veri ön işlemenin, yapılacak analizin kalitesini ve etkinliğini, önemli ölçüde etkilediğini sınıflandırma yöntemlerini kullanarak göstermişlerdir.
Çetin ve Yıldız (2022)	Veri ön işleme	Veri ön işlemenin veri analizi süreçlerindeki kritik rolünü vurgulayarak uygun yöntem seçiminin analiz sonuçlarının doğruluğu ve güvenilirliği üzerinde önemli bir etkiye sahip olduğunu ortaya koymuşlardır. Veri ön işlemenin, karar verme süreçlerinin iyileştirilmesine katkıda bulunarak model performansını ve karar destek sistemlerinin etkinliğini artırma potansiyeline sahip olduğunu belirtmişlerdir.
Sıkı ve Acartürk (2019)	Betimsel istatistikler ile veri ön işleme	Türkiye'deki ilaç sektöründe faaliyet gösteren firmaların Sanayi ve Teknoloji Bakanlığı tarafından Ar-Ge merkezi belgesi verildikten sonraki patent ve yenilik faaliyetlerini değerlendirmişlerdir. Çalışmalarında, Ar-Ge merkezlerine verilen desteklerin, yenilik faaliyetlerini, proje sayılarını ve patent başvurularını olumlu yönde etkilediği sonucuna varılmıştır.
Gülyaz ve Ertürk (2020)	Betimsel istatistikler ile veri ön işleme	Doğu Marmara bölgesinde faaliyet gösteren Ar-Ge merkezlerinin bilgi yönetimi uygulamaları ile inovasyon yeteneği arasındaki ilişkiyi incelemişlerdir. Bilgi yönetiminin Ar-Ge merkezlerinin inovasyon yeteneği ile olumlu ve anlamlı bir ilişkisinin bulunduğu, özellikle bilginin elde edilmesi ve paylaşılmasının, inovasyon yeteneğinin, öğrenme, üretim, pazarlama ve stratejik planlama kriterleri bakımından önemli olduğu sonucuna varılmıştır.
Ekinci (2021)	Betimsel istatistikler ile veri ön işleme	Gaziantep ilinde bulunan dokuz (9) Ar-Ge merkezinin performanslarının ölçülebilmesi amacıyla anket çalışması yapmıştır. Otuz üç (33) göstergeden oluşan Ar-Ge merkezi performans endeksleri kullanılarak betimsel istatistiklere yer verilip sonuçlar yorumlanmıştır.

Çizelge 1.1 Veri madenciliği yöntemleri ile ilgili teorik ve farklı alanlardaki uygulamalar (devam)

Mete ve Dağdeviren (2017)	Faktör analizi ve regresyon analizi	Otomotiv yan sanayi sektöründe bulunan Ar-Ge merkezlerinin bilgi yönetimi etkinliğini ortaya koyacak bir model oluşturmaya çalışmışlardır. Faktör analizi sonucu bilgi yönetimi ile Ar-Ge performansı arasında pozitif yönlü bir korelasyon bulunmuştur. Korelasyon analizi sonucunda, bilgi yönetimi uygulamalarını kapsayan faktörlerin bazıları ile Ar-Ge performansı arasında ilişki olduğu ancak regresyon analizinde bu faktörlerin Ar-Ge performansını etkilemediği belirtilmiştir. Yapılan istatistiksel test sonucuna göre firmalar arasında ölçek ve faaliyet süreleri bakımından fark olmadığı sonucuna ulaşılmıştır.
Cebeci vd. (2021)	Faktör analizi ve regresyon analizi	Türkiye'de Ar-Ge merkezi belgesine sahip farklı sektörlerde faaliyet gösteren işletmelerde takım algısı ve öğrenme yöneliminin Ar-Ge faaliyetlerine etkisini, faktör analizi ve regresyon analizi ile incelemişlerdir. Takım algısı ve öğrenme yönelimi ile Ar-Ge proje başarısı arasında pozitif bir ilişki olduğu sonucuna ulaşılmıştır.
Yeşildal ve Kamasak (2021)	Faktör analizi ve regresyon analizi	Yüz kırk beş (145) Ar-Ge merkezinin bilgi işleme yeteneklerinin inovasyon üzerindeki etkisini araştırmak üzere anket çalışması yapmışlardır. Bu amaçla elde edilen değişkenler, faktör analizi yapılarak gruplandırılmış regresyon analizi sonucu merkezlerin bilgi işleme yetenekleri ile inovasyon performansı arasında pozitif bir ilişki olduğu tespit edilmiştir.
Soyal vd. (2022)	Korelasyon ve regresyon analizi	Türk savunma ve havacılık sanayisinde algılanan kritik başarı faktörlerinin Ar-Ge projelerinin başarısına etkisi araştırmışlardır. Yapılan korelasyon ve regresyon analizi sonucu, projelerin performansını en çok etkileyen faktörler sırasıyla sorun giderme ve proje misyonu olarak belirlenmiş olup algılanan kritik başarı faktörleri ile proje performansı arasında anlamlı bir ilişki olduğu gösterilmiştir.
Aghdaie ve Alimardani (2015)	AHP ve TOPSIS	Çalışmalarındaki hedef pazar seçim probleminde alternatifleri değerlendirerek AHP ve TOPSIS içeren yenilikçi bir hibrit MCDM yöntemi önermişlerdir. AHP ile her değişken ağırlığı hesaplanarak TOPSIS ile hedef pazar alternatifleri sıralanmıştır.

Çizelge 1.1 Veri madenciliği yöntemleri ile ilgili teorik ve farklı alanlardaki uygulamalar (devam)

Keleş ve Tunca (2015)	AHP	Bir Ar-Ge firmasının kuruluş aşamasında, işletmelerin görüşüne göre önemli değişkenlerin AHP yöntemi ile ağırlıklarını belirlemişlerdir. Mevcut programlar değerlendirilerek sonuçlar ile girişimciler için bir ölçek oluşturulmuştur.
Arslan ve Belgin (2020)	AHP ve GRA	İmalat sanayisindeki yüksek ve orta-yüksek teknoloji alanlarını önceliklendirmek için değişkenlerin ağırlıklandırılmasında AHP uygulamışlardır. Çalışma sonucunda MCDM yöntemlerinden GRA ile elde edilen sıralamaya göre ilgili sektörlere sağlanan Ar-Ge, yenilik ve girişimcilik desteklerinde öncelik tanınması önerilmiştir.
Salman ve Peker (2021)	ENTROPY ve ARAS	Savunma sanayinde faaliyet gösteren Ar-Ge merkezlerinin performansını etkileyen değişken ağırlıklarını ENTROPY ile belirleyerek ARAS yöntemiyle Ar-Ge merkezlerinin sıralamasını yapmışlardır. Sıralama sonucuna göre en yüksek değişken ağırlığına Ar-Ge gelirlerinin, en düşük değişken ağırlığına ise Ar-Ge harcamalarının sahip olduğu gözlenmiştir.
Zafar vd. (2021)	ENTROPY, CRITIC, WSM, TOPSIS ve VIKOR	ENTROPY ve CRITIC'in bir kombinasyonu olan yeni bir ağırlıklandırma yaklaşımı önermişlerdir. On altı (16) blockchain benimseme kriterlerini içeren veri seti, MCDM yöntemlerinden WSM, TOPSIS ve VIKOR ile değerlendirilmiştir. WSM, TOPSIS ve VIKOR yöntemlerinin değerlendirilmesinde Spearman'ın sıra korelasyon katsayısı kullanılmıştır. Bu yöntemlerle çeşitli blockchain platformları değerlendirilerek sıralanmıştır.
Çetin ve Kuvat (2022)	ENTROPY, CRITIC ve COPRAS	Türkiye'de 2017-2019 yılları arasındaki bölgesel gelişmişlik düzeylerini etkileyen sekiz (8) ekonomik göstergenin sıralanmasını amaçlamışlardır. Bunun için, ENTROPY ve CRITIC ile ağırlıklandırılan göstergeler COPRAS ile sıralanmıştır. Yapılan Spearman korelasyon analizi sonuçlarına göre iki yöntemle elde edilen sıralamalar arasında yüksek ilişkinin varlığı gösterilerek sonuçlar karşılaştırılmıştır.

Çizelge 1.1 Veri madenciliği yöntemleri ile ilgili teorik ve farklı alanlardaki uygulamalar (devam)

Eşiyok vd. (2023)	ENTROPY, CRITIC ve EDAS	2010-2020 yılları arasında G7 ülkeleri ve Türkiye için yıllık ülke performans sıralaması yapmışlardır. ENTROPY ve CRITIC ile ağırlıklandırılan değişkenler EDAS ile sıralanmıştır. Yapılan duyarlılık analizine göre, ülkelerin sıralanmasında ENTROPY ve CRITIC ağırlık katsayıları bakımından anlamlı bir farklılık bulunmamaktadır.
Yu vd. (2014)	Veri madenciliği yöntemleri	Veri madenciliği yöntemlerinin, büyük ölçekli verilerden anlamlı öngörüler sunarak kurumlarda bilinçli kararlar alınmasını ve etkili stratejiler sağladığını belirtmişlerdir.
Patel vd. (2015)	RF, SVM, Naive Bayes ve ANN	Hindistan borsalarındaki hisse senetleri ve hisse senedi fiyat endekslerinin hareket yönünü tahmin etmek amacıyla sınıflandırma yöntemlerini kullanmışlardır. 2003-2012 yıllarını kapsayan on (10) sürekli değişkenin bulunduğu veri seti için RF, SVM, Naive Bayes ve ANN uygulanmıştır. Doğruluk ve F1-Skor performans ölçütlerine göre RF daha iyi performans göstermiştir.
Dirican (2019)	SVM, RF ve ANN	İki sınıflı bağımlı değişkeni bulunan bir gerçek veri setinde SVM, RF ve ANN yöntemlerinin sınıflandırma performanslarını karşılaştırmıştır. Aynı sınıflandırma yöntemleri, yapay veri seti üzerinde de gerçekleştirilmiş, her iki yaklaşımda da SVM'nin RF ve ANN'ye göre doğruluk performans ölçütü bakımından daha yüksek değere sahip olduğu görülmüştür.
Sıtkı (2020)	RF, Çok katmanlı algılayıcılar, Naive Bayes, *k*NN, LR, Karar Ağaçları ve ZeroR algoritması	On sekiz (18) farklı hastalığın teşhisine yönelik sınıflandırma çalışması amacıyla RF, Çok katmanlı algılayıcılar, Naive Bayes, *k*NN, LR, Karar Ağaçları ve ZeroR algoritması kullanmıştır. Doğruluk ölçütüne göre, RF'nin daha iyi performansa sahip olduğu görülmüştür.
Zan (2021)	*k*NN, SVM, Regresyon Ağaçları, Torbalama, RF ve XGB	Meteorolojik faktörlerin radon gazı üzerindeki etkilerinin değerlendirilmesi için denetimli öğrenme algoritmalarından çoklu doğrusal regresyon, *k*NN, SVM, Regresyon Ağaçları, Torbalama, RF ve XGB yöntemlerini uygulamıştır. 5-kat CV sonucu elde edilen belirtme katsayısı, hata kareler ortalaması, hata kareler ortalamasının karekökü ve ortalama mutlak hata değerlerine göre RF en iyi performansı gösterdiğini belirtmiştir.

Çizelge 1.1 Veri madenciliği yöntemleri ile ilgili teorik ve farklı alanlardaki uygulamalar (devam)

Rafiei-Sardooi vd. (2021)	SVM, RF ve Desteklenmiş Regresyon Ağaçları	Kentsel sel riski değerlendirmesinde MCDM ve Makine Öğrenimi (Machine Learning-ML) yöntemlerini birleştirerek hibrit bir yaklaşım sunmuşlardır. Çalışmada sel tehlikesini modellemek için SVM, RF ve Desteklenmiş Regresyon Ağaçları kullanılmıştır. Model performansları AUC değerleri ile ölçülmüş olup RF diğer yöntemlerden daha iyi performans göstermiştir. Önerilen hibrit yaklaşım ile kentsel sel riski olan bölgelerin haritalanması yapılmıştır.
Osmanoğlu (2021)	SVM, Naive Bayes, kNN, LR, Karar Ağaçları ve RF	Sağlık alanında kalp yetersizliği mortalitesinin tahminlenmesinde SVM, Naive Bayes, kNN, LR, Karar Ağaçları ve RF algoritmalarını uygulamıştır. Uygulama sonuçlarına göre doğruluk, duyarlılık, özgüllük ve AUC performans ölçütleri bakımından RF'nin daha iyi performans gösterdiğini ifade etmişlerdir.
Başer vd. (2021)	LR, kNN, Karar Ağaçları, Naive Bayes ve RF	Diyabet hastalığının sınıflandırılmasında LR, kNN, Karar Ağaçları, Naive Bayes ve RF algoritmalarının performanslarını karşılaştırmışlardır. Analiz sonucunda, duyarlılık, özgüllük, kesinlik ve AUC performans ölçütleri bakımından RF'nin en iyi sınıflandırma performansını sağladığını göstermişlerdir.
Zengin (2022)	XGB, HGAM, CatBoost, GAM, AdaBoost, kNN, Karar Ağaçları, RF, LR ve SVM	Bireylerin dijital bankacılık uygulamalarına karşın tutumları incelemek istemiştir. XGB, HGAM, CatBoost, GAM, AdaBoost, kNN, Karar Ağaçları, RF, LR ve SVM algoritmaları uygulanmıştır. Doğruluk performans ölçütüne göre RF algoritmasının en iyi performansı gösterdiği görülmüştür.
Ghosh vd. (2022)	RF, SVM ve XGB	Hindistan'da bulunan bir nehir havzasındaki taşkınlıkların modellenmesinde RF, SVM ve XGB kullanmışlardır. Hesaplanan performans ölçütleri Friedman ve Wilcoxon işaret testi ile test edilmiştir. Friedman testine göre XGB en iyi performans gösteren model iken Wilcoxon işaret testine göre ise XGB ile RF arasında performansları bakımından istatistiksel olarak anlamlı bir fark bulunmamıştır.

Çizelge 1.1 Veri madenciliği yöntemleri ile ilgili teorik ve farklı alanlardaki uygulamalar (devam)

Xie vd. (2023)	TOPSIS, Karar Ağaçları, RF ve XGB	Birçok kategorik değişkenli veri setlerinin modellenmesinde karşılaşılan zorluklara bir yaklaşım önermişlerdir. Önerilen yaklaşımda, gerçek bir veri setindeki değişkenlerin seçimi TOPSIS ile belirlenmiştir. Modelin sonucunu etkileyen en önemli değişkenlerin belirlenerek değişken seçimi sonrası model tahmin doğruluğunun korunması amaçlanmıştır. Ayrıca önerilen yaklaşım, üç simülasyon modeli ile değerlendirilmiştir. Karar Ağaçları, RF ve XGB uygulanarak performansları hesaplanmıştır. Önerilen yaklaşımın XGB yönteminden daha iyi performansla RF algoritmasına benzer bir performans gösterdiği sonucuna ulaşılmıştır.

Veri ön işleme, bilgi keşfi sürecinin en kritik aşamasıdır. Veri ön işlemede, verinin temizlenmesi, bütünleştirilmesi, dönüştürülmesi ve standartlaştırılması işlemleri yapılır. Veri ön işleme aşamasında, veri setinin yapısını tanımlamak, boyut azaltmak ya da kümelemek amacıyla keşfedici veri analizi yöntemleri kullanılır. Veri ön işlemede, veri setindeki değişken ağırlıklarının belirlenmesi ve subjektif değerlendirmelerin bilgiye dahil edilmesi amacıyla MCDM yöntemleri kullanılır. Veri madenciliği yöntemleri, karar destek sistemlerinde kritik bir role sahiptir. Veri madenciliğinde temel olarak istatistiksel modeller, matematiksel algoritmalar ve makine öğrenimi teknikleri kullanılmaktadır. Veri madenciliği sınıflandırma yöntemlerinin performanslarının değerlendirilmesinde performans ölçütleri kullanılır. Uygun sınıflandırma yönteminin seçilmesinde MCDM yöntemleri uygulanır. Veri madenciliği yöntemleri ile MCDM yöntemlerinin birlikte kullanıldığı çalışmalara literatürde rastlanılmamıştır. Bu tez çalışmasının özgünlüğü, veri madenciliği ve MCDM yöntemleri kullanılarak hibrit bir sınıflandırma yaklaşımının önerilmesidir.

Veri madenciliği sınıflandırma yöntemleri, kurumlarda yaygın bir biçimde karşılaşılan sınıflandırma problemlerinin çözümünde oldukça etkilidir. Özellikle bir ülkenin gelişimine katkıda bulunacak çalışmalarda, kurumlarda depolanmış verilerin işlenerek bilgi keşfedilmesi, gelecek çalışmalar hakkında öngörü oluşturulması ve karar verme sürecinin başarıya ulaşması bakımından gereklidir. Son yıllarda bilgi çağının

gereksinimlerinin karşılanması ve dünya çapında rekabet edilebilir bir seviyeye gelinebilmesinde yenilik çalışmalarının önemi oldukça büyüktür. Bu amaçla, birçok kurum ve kuruluş, gelişmesi öncelikli olarak görülen alanlarda yapılan yenilik faaliyetlerini, hibe, istisna ve muafiyet gibi teşvik mekanizmaları aracılığıyla desteklemektedir. Hedeflenen bu teşvik mekanizmasının başarıya ulaşması ve sürdürülebilir olması bakımından desteklenecek kişi ya da firmaların belirlenmesinde sınıflandırma çalışması oldukça önemlidir. Hızlı ve etkin karar almada iyileştirme sağlayacak istatistiksel bir yaklaşımın uygulanması hedeflenen başarının elde edilmesi bakımından değerlidir.

Çalışmada Sanayi ve Teknoloji Bakanlığı'ndan Ar-Ge ve Tasarım merkezlerine ilişkin ham veri seti temin edilerek veri temizliği yapılmıştır. Veri ön işleme aşamasında AHP ile subjektif, ENTROPY ve CRITIC ile objektif olarak belirlenen değişken ağırlıkları ile veri seti ağırlıklandırılmıştır. Ağırlıklandırılmış veri setine uygulanan LR, *k*NN, SVM ve RF veri madenciliği sınıflandırma yöntemlerinin performans ölçütleri hesaplanmıştır. Elde edilen performans ölçütleri bir karar matrisi biçiminde ele alınarak sınıflandırma yöntemlerine karar verilmesinde MCDM yöntemleri kullanılmıştır. Toplama dayalı TOPSIS ve COPRAS, orana dayalı EDAS ve CODAS ve min-max değerine dayalı MABAC ve GRA normalleştirme yaklaşımı kullanan MCDM yöntemleri uygulanmıştır. Elde edilen MCDM sıralama sonuçlarına göre sınıflandırma yöntemlerinden RF algoritmasının Ar-Ge ve Tasarım merkezlerinin sınıflandırılmasında öncelikli tercih edilebilir olduğu görülmüştür. Çalışmada çok sayıda yöntem etkileşimli olarak bir arada kullanıldığından Ar-Ge ve Tasarım Merkezlerinin faaliyetlerinin değerlendirilmesine ilişkin karar verme süreci hibrit bir yaklaşım olarak tanımlanmıştır.

Tez çalışmasının ikinci bölümünde, veri ön işlemede kullanılan kavramların tanımları ile keşfedici veri analiz yöntemlerine ilişkin teorik bilgiler sunulacaktır. Veri ön işleme aşamasında verideki önemine göre değişkenlerin ağırlıklarını belirlemek amacıyla uygulanan MCDM (AHP, ENTROPY ve CRITIC) yöntemleri hakkında detaylı açıklamalar verilecektir.

Bu çalışmanın özgün yanını oluşturan üçüncü bölümde, çalışmanın bir sınıflandırma problemi olması sebebiyle veri madenciliği sınıflandırma yöntemleri kullanılacaktır. Değişkenleri ağırlıklandırılmış veri setine uygun olan sınıflandırma yöntemlerinden LR, *k*NN, SVM ve RF algoritmalarının teorik açıklamaları, uygulama adımları ve ayarlanabilir parametreleri hakkında bilgi verilerek sınıflandırma yöntemlerine ilişkin detaylı anlatım yapılacaktır.

Dördüncü bölümde, sınıflandırma performans ölçütleri hakkında bilgi verilecektir. Karar verme süreci tanımlanarak, karar matrisine uygulanan normalleştirme yaklaşımlarına göre MCDM yöntemleri açıklanacaktır. Performans değerlendirmesi amacıyla kullanılan MCDM yöntemlerinden TOPSIS, COPRAS, CODAS, EDAS, MABAC ve GRA yöntemlerinin işleyişi hakkında açıklamalar yapılacaktır.

Beşinci bölümde, gerçek veri setinin keşfedici veri analizi yöntemleri ile veri ön işleme sürecinden geçirilmesi, değişken ağırlıklarının elde edilmesi, ağırlıklandırılmış gerçek veri setine LR, *k*NN, SVM ve RF sınıflandırma yöntemlerinin uygulanması, doğruluk, kesinlik, duyarlılık, F_1-Skor ve AUC performans ölçütlerinin hesaplanması ile performans ölçütlerine dayalı olarak TOPSIS, COPRAS, CODAS, EDAS, MABAC ve GRA kullanılarak sınıflandırma yöntemlerine karar verme süreçlerine ilişkin analiz sonuçları sunulacaktır. Gerçek veri setinden elde edilen sonuçların genelleştirilebilmesi amacıyla farklı senaryolar kurgulanarak simülasyon çalışması yapılmıştır. Simülasyon çalışmasında, gerçek veri setine yapılan benzer analizler uygulanarak sonuçları detaylı bir biçimde verilecektir.

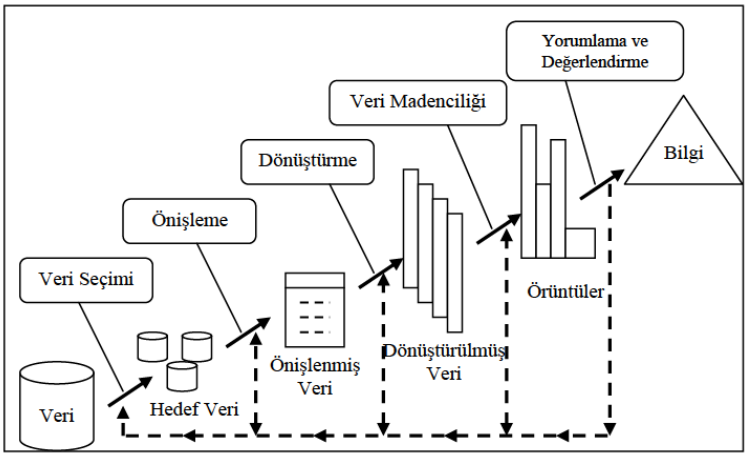
Altıncı bölümde, gerçek veri seti ve simülasyon çalışmasına ilişkin elde edilen sonuçlar hibrit karar verme yaklaşımı ile detaylıca yorumlanacaktır.

2. VERİ ÖN İŞLEME

Teknolojik gelişmelere bağlı olarak elektronik ortamda uygun biçimde depolanan ham veriden, anlamlı ve faydalı bilgiler elde edilmesi için ham verinin ön işlemeden geçirilmesi gerekir. Ham veri, araştırma, ölçüm veya gözlem yoluyla elde edilen işlenmemiş değerlerden oluşur. Büyük veri ise geleneksel istatistik yöntemlerle işlenmesi zor olan büyük hacimli ve karmaşık veri kümelerini ifade eder. Büyük veri, hacim (Volume), hız (Velocity), çeşitlilik (Variety), değişkenlik (Veracity), geçerlilik (Validity) ve değer (Value) olmak üzere 6V olarak bilinen temel özelliklerle tanımlanır (Sorhun 2021). Bu özellikler, verinin çok büyük miktarda ve kapasitede bulunması, hızlı bir biçimde üretilip toplanması, farklı format ve yapıdaki değerlerden oluşması, sürekli güncellenmesi, yüksek hızda değişimi nedeniyle geçerliliğinin belirsizleşmesi ve veriden elde edilen bilginin değerini ifade eder. Büyük veri yapılandırılmış, kısmen yapılandırılmış ve yapılandırılmamış olmak üzere üçe ayrılır. Yapılandırılmamış veri genellikle metin, görüntü, ses, video ve e-posta gibi biçimlerde, kısmen yapılandırılmış veri, belirli bir yapıya sahip olsa da karmaşık ve çeşitli formatlarda bulunurken, yapılandırılmış veri tablolar halinde bulunur.

Veri ambarı, çeşitli kaynaklardan gelen büyük miktardaki veriyi depolamak, yönetmek ve analiz etmek amacıyla tasarlanmış bir sistemdir (Yıldız 2014). Bu sistem içerisinde, meta veri verilerin tanımlandığı alanı, veri sözlüğü ise değişken tanımlarının yer aldığı bölümü oluşturur. Her bir veri elemanı bir gözlem olarak ifade edilir ve bu gözlemlere ait özellikler değişken olarak adlandırılır. Değişkenler sebep-sonuç ilişkisine, yapılarına ve aldıkları değerlere göre farklı biçimlerde tanımlanır (Erbaş 2016). Bağımlı değişken, bir çalışmanın amacını belirleyen yanıt veya hedef değişken olarak da bilinen çıktı değişkendir. Bağımsız değişken ise bağımlı değişkeni etkileyen veya açıklayan girdi değişkendir. Nicel değişken sayısal değerler, nitel değişken ise kategorik değerler ile ifade edilir. Nicel değişken, aldıkları değerlere göre sürekli ve kesikli olarak ikiye ayrılır. Sürekli değişken belirli bir aralıkta sonsuz değer alabilirken, kesikli değişken sayılarak elde edilebilen değerler ile tanımlanır.

En bilinen haliyle veri bilgi keşfi süreci şekil 2.1'de görülmektedir.



Şekil 2.1 Veriden bilgi keşfi süreci (Fayyad vd. 1996)

Şekil 2.1'de görülen veri bilgi keşfi sürecinde öncelikle veri ambarından bilgi elde edilmesi istenilen araştırma konusuna göre veri seçimi yapılır. Veri ön işleme aşamasında veri temizleme, veri bütünleştirme ve boyut azaltma çalışmaları gerçekleştirilir. Kayıp gözlem varsa tamamlanır, hatalı gözlem varsa çıkarılır. Eğer, veride dönüşüm gerekiyorsa, veri setindeki değişkenler üzerinde düzenlemeler yapılarak temiz veri olarak da adlandırılan dönüştürülmüş verilere ulaşılır.

Bilgi keşfi sürecinde, çeşitli veri yapıları ve farklı amaçların olması nedeniyle birçok disiplindeki yöntemlere ihtiyaç duyulmaktadır. Bu süreçte kullanılan bilgi keşfi yöntemleri şekil 2.2'de verilmiştir.

Şekil 2.2 Bilgi keşfi yöntemleri (Altunkaynak 2022)

Bilgi keşfi sürecinde şekil 2.2'de belirtilen yöntemlere ilişkin özet bilgiler aşağıda verilmiştir.

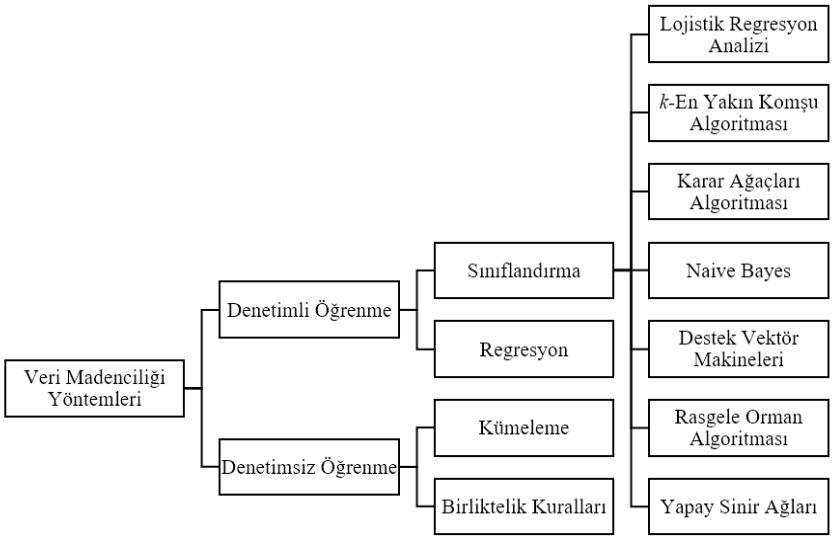
Yapay Zekâ, bilgisayar teknolojilerine, insana özgü zekâ ve öğrenme yeteneklerini kazandırmayı amaçlayan bir bilim dalıdır. Yapay zekâ sistemleri, veri analizi, örüntü tanıma, doğal dil işleme gibi alanlarda uygulanıp kendi deneyimlerinden öğrenerek geliştirilebilirler.

Makine Öğrenimi, bilgisayar sistemlerinin veri analizi yaparak örüntüleri öğrenmesine ve deneyimlerinden bilgi çıkarmasına imkân tanıyan bir yapay zekâ alt dalıdır. Denetimli ve denetimsiz öğrenme olmak üzere ikiye ayrılır.

Derin Öğrenme, yapay sinir ağları gibi karmaşık ve çok katmanlı yapıları kullanarak büyük veriden öğrenme yeteneği olan makine öğreniminin bir alt dalıdır. Bu özelliği sayesinde görüntü tanıma, doğal dil işleme ve diğer karmaşık problemlerde oldukça başarılıdır.

Veri bilimi, istatistik, matematik ve makine öğrenimi gibi disiplinleri kullanarak özellikle yapılandırılmamış verilerden oluşan büyük veri setlerini analiz eden, anlamlı bilgiler çıkaran ve bu bilgileri kullanarak karar vermeyi sağlayan bir bilimdir.

Veri madenciliği, makine öğrenimi ve istatistiksel yöntemleri birlikte kullanarak veriden yararlı bilgilerin elde edilmesini amaçlayan veri biliminin bir alt dalıdır. Veri madenciliği yöntemleri şekil 2.3'te verilmiştir.



Şekil 2.3 Veri madenciliği yöntemleri

Veri madenciliği yöntemleri, denetimli ve denetimsiz öğrenme olmak üzere şekil 2.3'te gösterildiği gibi iki başlıkta incelenir. Denetimli öğrenme regresyon ve sınıflandırma yöntemlerini içerir. Denetimli öğrenme ile verilerin tahmin ya da sınıflandırılması yapılarak öğrenme gerçekleşir. Denetimsiz öğrenme ise kümeleme ve birliktelik kurallarını içermektedir. Herhangi bir küme ya da kategoriye ait olduğu bilinmeyen veriler için denetimsiz öğrenme uygulanır.

İstatistik, araştırmacı tarafından deney veya gözlem yoluyla elde edilmiş veriyi, analiz etme, çözümleme ve yorumlama süreçlerini kullanarak sayısal bilgiler elde etmeyi amaçlayan bir bilim dalıdır. Tüm bilgi keşfi yöntemlerinin temeli istatistik bilimine dayanır.

Veri bilgi keşfinin en önemli aşaması veri ön işleme aşamasıdır (Oğuzlar 2003). Bu aşama veri temizliği, betimsel istatistikler ve veri görselleştirme, keşfedici veri analizi ve değişken ağırlıklarını belirleme başlıklarını içerir.

2.1 Veri Temizliği

Veri temizliğinin ilk aşamasında öncelikle ham veri seti değerlendirilir. Bu aşama, veri setinin genel durumunu anlamak ve temizleme sürecinde hangi adımların atılması gerektiğini belirlemek için kritik öneme sahiptir. Veri setinin yapısını anlamak için veri setindeki gözlem ve değişken sayısı belirlenir. Veri setindeki değişken türleri (nitel, nicel, kategorik vb.) incelenir. Benzer bilgi taşıyan değişkenler bütünleştirilir. Birçok veri kaynağının birleştirilmesinden kaynaklı oluşan verilerdeki format farklılıkları belirlenir. Veri setindeki tutarsız ve mantıksız değerler tespit edilir. Tekrar eden gözlemler belirlenerek bozuk veya hatalı girişler kontrol edilir. Bu aşamanın sonunda veri setinin yapısı ve içerdiği problem saptanarak izlenmesi gereken adımlar ve yöntemler belirlenir. Betimsel istatistikler ve veri görselleştirme kullanılarak keşfedici veri analizi yapılır.

2.2 Betimsel İstatistikler ve Veri Görselleştirme

Betimsel istatistikler, bir veri setinin genel özelliklerini anlamak, özetlemek, yorumlamak ve görsel olarak ifade etmek amacıyla elde edilen temel bilgilerdir. Betimsel istatistikler hesaplanmadan önce veri setindeki değişken türleri tespit edilir. Çizelge 2.1'de çok değişkenli bir veri seti verilmiştir (Türkşen 2023).

Çizelge 2.1 Çok değişkenli veri seti

	Bağımsız değişkenler				***Bağımlı değişken***
No.	X_1	X_2	…	X_m	Y
1	x_{11}	x_{12}	…	x_{1m}	y_1
2	x_{21}	x_{22}	…	x_{2m}	y_2
⋮	⋮	⋮	⋱	⋮	⋮
n	x_{n1}	x_{n2}	…	x_{nm}	y_n

m sayıda bağımsız değişken, bir bağımlı değişkenden oluşan n gözlemli bir veri seti çizelge 2.1'deki gibi gösterilir. Çizelge 2.1'de bağımsız değişkenler $X = \left[x_{ij} \right]_{n \times m}$ ve bağımlı değişken $Y = \left[y_i \right]_{n \times 1}$ biçimindedir, $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m$. Bağımlı ve

bağımsız değişkenler, nicel ya da nitel değerler alabilir. Nitel değişkenler kodlanarak sayısal biçimde ifade edilebilir.

Betimsel istatistiklerde, merkezi eğilim ölçüleri ve yayılım ölçüleri kullanılır. Merkezi eğilim ölçüleri, veri setindeki gözlemlerin toplandığı merkez bir noktanın belirlenmesidir. Bu merkez nokta, ortalama, ortanca (medyan) ya da mod değerleridir. Nitel değişkenler için mod, nicel değişkenler için ise ortalama, ortanca ve mod, merkezi eğilim ölçüleri olarak kullanılır. Yayılım ölçüleri ise veri setindeki değerlerin ne kadar yayıldığını ve değiştiğini belirlemek için kullanılır. Bu ölçüler; varyans, standart sapma, değişim katsayısı, basıklık katsayısı, çarpıklık katsayısı, açıklık ve çeyrekliklerdir.

Veri setine ait merkezi eğilim ölçüleri ve yayılım ölçüleri hesaplanarak grafikler ile görselleştirilir. Betimsel istatistiklerde en çok kullanılan grafik türleri, histogram, çubuk grafiği, kutu grafiği, Q-Q grafiği, saçılım grafiği ve pasta grafiğidir. Nicel verilerde, histogram, kutu grafiği, Q-Q grafiği, nitel verilerde ise saçılım grafiği ve pasta grafiği görselleştirme amacıyla kullanılır.

Veri setlerinde, kayıp gözlem, aykırı gözlem veya farklı türde değişkenlerin bulunması gibi sorunlarla karşılaşılabilir. Veri setlerinde ölçüm hatası, bilgi eksikliği, sistemsel kaynaklı sorunlar gibi nedenlerden dolayı eksik/kayıp gözlem oluşur. Bu problemin çözümünde, kayıp gözlemin bulunduğu satırları silme işlemi yapılabilir. Fakat bu durum, kayıp gözlem sayısının fazlalığı ya da veri setindeki gözlem sayısının az olduğu durumlarda bilgi kaybına neden olur. Kullanılacak diğer bir yöntem, kayıp gözlemin doldurulma işlemidir. Betimsel istatistikler kullanılarak kayıp gözlem doldurulur. Betimsel istatistiklerden mod ya da medyan kayıp gözlemin doldurulmasında çoğunlukla tercih edilir. Ölçüm ya da örnekleme hatası nedeniyle veri setinde bazen aykırı gözlem tespit edilir. Aykırı gözlem, ortalama etrafında dağılan verilerden aşırı büyük ya da küçük değerlerdir. Aykırı gözlem, yanlı tahmin, yüksek varyans ya da düşük performansa yol açabileceğinden tespit edilmesi gerekir. Aykırı gözlem tespitinde kutu grafiği kullanılır. Bu grafik ile verilerin normal dağılmamasına neden olan aykırı gözlemler görsel olarak tespit edilir.

Veri setinde birbirinden farklı yapıda bulunan değişkenlere, seçilen analiz yöntemine uygun bir biçimde veri dönüştürme işlemi yapılır. Verilerin daha kolay yorumlanabilmesi için logaritmik ya da karekök dönüşümü uygulanabilir. Veri setindeki metinsel yapıdaki değişkenlerin analize dahil edilebilmesi için kategorik duruma getirilir. Farklı ölçüm birimine sahip değişkenlerin, belirli bir aralığa indirgenmesi için normalleştirme ya da standartlaştırma işlemi gerekir.

Normalleştirme: Veri setindeki değişkenlerin birim farklılıklarının giderilmesi amacıyla belirli bir aralık içerinde kalacak biçimde dönüşüm yapılır. Verinin ortalaması (X_{ort}) minimum değeri $min(X)$ ve maksimum değeri $maks(X)$ kullanılarak

$$X^* = \frac{X - min(X)}{maks(X) - min(X)} \tag{2.1}$$

normalleştirilmiş X^* değerine dönüştürülür. Bu ölçeklendirme normalizasyonu ile veri setindeki değişkenler $[0, 1]$ veya genel olarak her değişkenin minimum değeri a ve maksimum değeri b olacak biçimde $[a, b]$ aralığında ölçeklendirilir. Normalleştirme ile değişkenler aynı ölçüm birimine getirilerek değişkenler arasında tutarlılık sağlanır. Farklı formüller ile elde edilen birçok normalleştirme yöntemi mevcuttur.

Standartlaştırma: Verinin ortalaması (X_{ort}) ve standart sapması (σ_X) kullanılarak

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{N}(X_i - X_{ort})^2}{N}}$$
$$X^* = \frac{X - X_{ort}}{\sigma_X} \tag{2.2}$$

standartlaştırılmış X^* değerine dönüştürülür (Erbaş 2016). Standartlaştırma ile değişken değerleri Z-skorlarına dönüştürülerek her bir gözlemin ortalama değerden kaç standart sapma uzaklıkta olduğu hesaplanır. Standartlaştırma, veri setindeki aykırı gözlemlere karşı daha duyarlıdır.

Betimsel istatistikler ile veri hazırlama aşamasından sonra verilerin analizini kolaylaştıran çok değişkenli istatistiksel yöntemlerden Faktör Analizi, Kümeleme

Analizi, Temel Bileşenler Analizi ve Diskriminant Analizi yaygın kullanılan yöntemlerdir. Bu çalışmada faktör analizi ile çok sayıda değişkenin anlamlı bir şekilde gruplandırılıp değişken sayısının azaltılması planlanmıştır.

2.3 Çok Değişkenli İstatistiksel Yöntemler ile Keşfedici Veri Analizi

Büyük hacimli veri setlerinde, değişken sayının fazla ve değişkenlerin birbiriyle ilişkili olması nedeniyle uygulanan tek değişkenli analizler anlamlı sonuçlar vermez. Bu nedenle çok değişkenli istatistiksel yöntemler kullanılır. Çok değişkenli istatistiksel yöntemler, değişkenler arasındaki karmaşık ilişkileri inceleyerek çok değişkenli bir veri setini daha az sayıda değişkene indirgeyen yöntemler bütünüdür. Çok değişkenli istatistiksel yöntemler, sınıflandırma, anlamlı değişken elde etme, değişkenlerin kovaryans ve korelasyonlarından yararlanarak bağımlılık yapısını ortaya koyma ve hipotezleri test etme amacıyla kullanılır (Tatlıdil 2002).

Faktör analizi, sıkça kullanılan çok değişkenli istatistiksel yöntemlerden biridir. Spearman (1904) çalışmasında faktör analizi yöntemini geliştirmiştir. Faktör analizi, çok sayıda değişken arasındaki ilişkiyi ve karmaşık örüntüleri ortaya çıkararak birbirinden bağımsız, daha az sayıda ve daha anlamlı değişkenler bulunmasını sağlar. Önemsiz değişkenleri de ortaya koyarak değişkenleri daha iyi yorumlar. Birbirinden etkilenen değişkenlerin kümelenmesini hedefler.

Faktör analizi, değişkenler arasındaki korelasyonlar ile faktör yapısını inceler (Harman 1976). Faktör analizinde, ham veri matrisi kullanıldığında varyans-kovaryans matrisinden, standartlaştırılmış veri matrisi kullanıldığında ise korelasyon matrisinden yararlanılır (Atakan 2022).

Faktör analizi değişkenler arasındaki ilişkinin doğrusal olmasıyla regresyon analizine benzer. m değişkenli $\boldsymbol{X}' = \left(X_1, X_2, \cdots, X_m\right)$ gözlenen rastgele vektörü için $E(\boldsymbol{X}) = \boldsymbol{\mu}$ ve $Cov(\boldsymbol{X}) = \Sigma$ olmak üzere faktör analizi modeli

$$X_j - \mu_j = \lambda_{j1}F_1 + \lambda_{j2}F_2 + \ldots + \lambda_{jt}F_t + \varepsilon_j, \quad j = 1, 2, \ldots, m \qquad (2.3)$$

biçiminde yazılır. Burada,

X_j : Gözlenen değişken

λ_{jt} : *j.* değişkenin *t.* ortak faktörü üzerindeki etkisine ilişkin faktör yükü

ε_j : *j.* değişken için ortak faktörler ile açıklanamayan kısım

F_t : Oluşturulan ortak faktör ve *t* ortak faktör sayısı

olarak tanımlanır.

Faktör analizi uygulamasında dikkate alınan temel hususlar şöyledir:

(*i*) Örneklem büyüklüğü için Comrey (1973) çalışmasında örneklem büyüklüğü için 100 zayıf, 300 iyi, 500 çok iyi, 1000 ise mükemmel olarak belirtmiştir. Tabachnick vd. (2007) çalışmasında en az 300 örneklem büyüklüğünü önermektedir.

(*ii*) Bartlett Değişken Yeterliliği Testi, değişkenler arasındaki ilişkiyi test ederek değişkenlerin faktör analizine uygunluğunu sınamak için kullanılır. Eğer, değişkenler arasında ilişki yoksa faktör analizine devam edilmez.

(*iii*) Kaiser-Meyer-Olkin Örneklem Yeterliliği Testi, örneklem büyüklüğünü test eder. Değişkenlerin iç tutarlılığını ölçmek için kullanılan bir yöntem olup $0 < \text{KMO} < 1$ arasında bir değer alır. KMO kriterinin 0.70 ve üzerinde bir değer olması istenir. Bu kriterin 0.50'nin altında olması önerilmez (Karagöz 2019). KMO değeri 0.50'den küçük olan değişken çıkarılarak analiz tekrarlanır.

(*iv*) Hesaplanan korelasyonlar bir korelasyon matrisinde yer alır. Faktör analizinde değişkenler arasında yüksek düzeyde korelasyon beklenir. Değişkenler arasındaki korelasyon azaldıkça, faktör analizinin güvenilirliği azalır. Korelasyon matrisinin determinantının, 0.00001'den büyük olması, çoklu bağlantının olmadığını gösteren genel bir kuraldır. Çoklu bağlantının varlığında, bu probleme neden olan gözlenen değişkenler analizden çıkarılarak test tekrarlanabilir (Yong ve Pearce 2013).

Faktör sayısının belirlenmesi faktör analizinde önemli bir yer tutar. Yamaç eğrisi grafiğinde özdeğer, bir faktör tarafından açıklanan toplam varyansı ifade eder. Varyans-

kovaryans matrisinden elde edilen özdeğerlerin $\frac{\sum_{j=1}^{t}\lambda_j}{\sum_{j=1}^{m}\lambda_j} \geq \frac{2}{3}$ sonucunu sağlayan en küçük *t* değeri uygun faktör sayısıdır. Özdeğerler korelasyon matrisinden elde edildiyse, özdeğeri 1 ve 1'den büyük olan faktörler modele alınır. Yamaç eğrisi, özdeğer sayısına göre faktör sayısının grafiğini vermektedir. Grafikte eğrinin yatıklaşmaya başladığı nokta faktör sayısının belirlenmesine yardımcı olur. Ortak varyans tablosu, her bir değişkenin diğer değişkenler ile paylaştığı ortak varyansları gösterir. Örneklem sayısının fazla olduğu durumda, ortak varyansı 0.3'ün altında olan değişkenler analizden çıkarılır (Yong ve Pearce 2013). Faktörlerin daha iyi yorumlanabilmesi için rotasyon yöntemi uygulanabilir. Rotasyon sayesinde daha basit, az sayıda ve yüksek faktör ağırlığına sahip faktörlerin bulunduğu bir modelin elde edilmesi amaçlanır. Dik ve eğik olmak üzere iki tane yöntem bulunmaktadır. Dik rotasyon türünde, eksenler 90 derecelik bir açıyla döndürülürken eğik döndürmede iki farklı açı bulunmaktadır. En önemli farklılık ise dik döndürmede faktörler ilişkisiz iken eğik döndürmede bu şart aranmamaktadır (Tatlıdil 2002). Yorumlama kolaylığı bakımından dik döndürme yöntemi en çok kullanılan yöntemdir.

2.4 Çok Ölçütlü Karar Verme Yöntemleri ile Değişken Ağırlıklarını Belirleme

Veri ön işlemede, elde edilecek anlamlı bilgi amacıyla veri seti ağırlıklandırılır. Çok ölçütlü karar verme yöntemleri (MCDM), değişkenlerin önem ağırlıklarının belirlenmesinde kullanılır. İnsan bilincinde mevcut olan doğal düşünce sürecini, modelin bir parçası olarak gören Analitik Hiyerarşi Süreci (Analytic Hierarchy Process-AHP), literatürde sıkça kullanılan yöntemlerden biridir. AHP ile karmaşık değişken yapıları, uzman görüşü dikkate alınarak değerlendirilir. AHP yönteminde kullanılan hiyerarşik yapı sayesinde veri setindeki değişkenler için önem ağırlığı oluşturulur. Uzman görüşüne ulaşılamadığı durumlarda değişken ağırlıkları objektif olarak belirlenir. Objektif olarak ağırlık belirlemede ENTROPY ve The Criteria Importance Through Intercriteria Correlation (CRITIC) yaygın biçimde kullanılan MCDM yöntemleridir. ENTROPY yöntemi, değişken ağırlıklarının belirlenmesinde uzman görüşüne gerek duyulmadan

veriye dayalı bilgiyi kullanır. CRITIC, uzman görüşlerinden faydalanmadan değişkenlerin standart sapmalarını ve değişkenler arası korelasyonu kullanarak değişkenlerin önem ağırlıklarını belirler. Bu çalışmada hem subjektif hem objektif olarak değişken ağırlığı hesaplayan MCDM yöntemleri seçilmiştir.

2.4.1 Subjektif yaklaşım

Subjektif karar verme ile uzman bilgisinden yararlanılarak değişken ağırlıkları hesaplanır. Bu amaçla, değişken ağırlığı belirlemek için öznel bilginin dikkate alındığı AHP yöntemi kullanılır.

Analitik Hiyerarşi Süreci (Analytic Hierarchy Process-AHP), Saaty (1977) tarafından karar verme problemlerinin çözümü için geliştirilen matematiksel bir yöntemdir. Bu yöntem, özellikle karmaşık yapıdaki karar verme problemlerinde hiyerarşi ölçeğini kullanarak problemi sadeleştirir. AHP, sezgisel ve subjektif değerlendirmelerin, matematiksel modellerle ifade edildiği analitik bir yöntemdir. Değişkenlerin önem düzeyine göre değerlendirilmesinde uzman görüşüne ihtiyaç duyar. Farklı tecrübe ve eğitime sahip uzman kişilerin görüşlerini birleştirerek karar verilmesini sağlar. Değişkenlerin ikili karşılaştırılması sırasında subjektif değerlendirme yapılması, problemi yalnızca sayısal verilere dayanmaktan kurtarır. AHP, hem nicel hem nitel değişkenleri aynı anda değerlendirir.

Saaty (1977) çalışmasında önerilen karşılaştırma ölçeğiyle birbirlerine göre önem değerleri dikkate alınarak karşılaştırma matrisleri oluşturulur. Veri setindeki her bir değişken kriter (ölçüt) olarak alınır.

Çizelge 2.2 AHP yönteminde hiyerarşi ölçeği

Önem Değerleri	Değer Tanımları
1	Her iki değişkenin eşit öneme sahip olması durumu
3	i. değişkenin, j. değişkenden daha önemli olması durumu
5	i. değişkenin, j. değişkenden çok önemli olması durumu
7	i. değişkenin, j. değişkene göre çok güçlü bir öneme sahip olması durumu
9	i. değişkenin, j. değişkene göre mutlak üstün bir öneme sahip olması durumu
2, 4, 6, 8	Ara değerler

Karşılaştırma matrisinin i. köşegeni üzerindeki bileşenler 1 değerini alır ve bu değer ilgili değişkenin kendisiyle kıyaslanması sonucunda ortaya çıkar. Değişkenlerin karşılaştırılması, birbirlerine göre sahip oldukları önem değerleri dikkate alınarak birebir ve karşılıklı yapılır. AHP işleyiş adımları aşağıdaki gibi tanımlanır.

Adım1: Çizelge 2.1'de verilen m sayıda değişken için çizelge 2.2'de verilen hiyerarşi ölçeği kullanılarak A karşılaştırma matrisi

$$A = \left[a_{ij} \right]_{m \times m} = \begin{bmatrix} 1 & a_{12} & \cdots & a_{1m} \\ 1/a_{12} & 1 & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1/a_{1m} & 1/a_{2m} & \cdots & 1 \end{bmatrix} \tag{2.4}$$

oluşturulur. Burada, a_{ij} i. değişkenin j. değişkene göre önem değerini, $1/a_{ij}$ j. değişkenin i. değişkene göre önem değerini ifade eder.

Adım2: A karşılaştırma matrisi

$$a_{ij}^{*} = \frac{a_{ij}}{\sum_{j=1}^{m} a_{ij}}, \; i = 1, 2, \ldots, m \tag{2.5}$$

biçiminde normalleştirilir.

Adım3: Normalleştirilmiş karşılaştırma matrisinin satır ortalamaları

$$w_j = \frac{1}{m} \sum_{j=1}^{m} a_{ij}^{*}, \; i = 1, 2, \ldots, m \tag{2.6}$$

j. değişkenin önem ağırlığıdır.

2.4.2 Objektif yaklaşım

Objektif karar vermede, subjektif karar vermenin aksine uzman bilgisi dikkate alınmadan değişken ağırlıkları veriye dayalı olarak hesaplanır. Bu amaçla çalışmada ENTROPY ve CRITIC yöntemleri kullanılmıştır.

Çizelge 2.1'de verilen n gözlemli bir veri setinde m sayıda bağımsız değişkenden oluşan bir X veri matrisi $n \times m$ boyutlu olarak tanımlanır ve

$$X = \left[x_{ij} \right]_{nxm} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{2.7}$$

biçiminde yazılır. Objektif olarak değişken ağırlığı elde edilmesinde X veri matrisi, n sayıda alternatif ve m sayıda kriterden oluşan $n \times m$ boyutunda karar matrisi olarak kullanılır.

ENTROPY

ENTROPY yöntemi, Shannon (1948) tarafından geliştirilen MCDM yöntemlerinde değişken ağırlıklarının belirlenmesinde kullanılır. Bu yöntemde uzman görüşüne gerek duyulmadan objektif olarak değişken ağırlıkları belirlenir.

Adım1: X karar matrisi eşitlik (2.7)'deki gibi tanımlanır.

Adım2: Karar matrisi.

$$x_{ij}^{*} = \frac{x_{ij}}{\sum_{j=1}^{m} x_{ij}}, \; i = 1, 2, \ldots, n \tag{2.8}$$

biçiminde normalleştirilir. Normalleştirilmiş karar matrisi $N = \left[x_{ij}^{*} \right]_{nxm}$ oluşturulur.

Adım3: Her bir değişken için ENTROPY değeri

$$E_j = -\frac{1}{ln(n)} \sum_{i=1}^{n} x_{ij}^{*} \, ln(x_{ij}^{*}) \;\; j = 1, 2, \ldots, m \tag{2.9}$$

hesaplanır.

Adım4: Her bir değişken için belirsizlik değeri

$$d_j = 1 - E_j, \quad j = 1, 2, \ldots, m \tag{2.10}$$

elde edilir.

Adım5: Değişken ağırlıkları

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j}, \quad j = 1, 2, \ldots, m \tag{2.11}$$

biçimindedir.

CRITIC

CRITIC (The Criteria Importance Through Intercriteria Correlation) yöntemi, Diakoulaki vd. (1995) tarafından geliştirilen MCDM yöntemlerinden biridir. Değişken ağırlıklarının belirlenmesinde değişkenlerin standart sapmaları ve değişkenler arası korelasyon kullanılır. Bu yöntem ile uzman görüşlerinden faydalanmadan değişken ağırlıkları belirlenir.

Adım1: X karar matrisi eşitlik (2.7)'deki gibi tanımlanır.

Adım2: Karar matrisi

$$x_{ij}^{*} = \frac{x_{ij}}{maks\, x_{ij}}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m \tag{2.12}$$

biçiminde normalleştirilir. Normalleştirilmiş karar matrisi $N = \left[x_{ij}^{*} \right]_{nxm}$ oluşturulur.

Adım3: s_j, j. değişkenin standart sapması ve r_{jt}, j. ve t. değişkenler arasındaki korelasyon katsayısı

$$s_j = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij}^{*} - \bar{x}_j^{*})^2}{n}}$$
$$r_{jt} = \frac{\sum_{i=1}^{n}(x_{ij}^{*} - \bar{x}_j^{*})(x_{it}^{*} - \bar{x}_t^{*})}{\sqrt{\sum_{i=1}^{n}(x_{ij}^{*} - \bar{x}_j^{*})^2 \sum_{i=1}^{n}(x_{it}^{*} - \bar{x}_t^{*})^2}}, \quad j,t = 1,2,\ldots,m \tag{2.13}$$

ile hesaplanır.

Adım4: Her bir değişkene özgü bilgi miktarı

$$C_j = s_j \sum_{t=1}^{m}(1 - r_{jt}), \quad j = 1,2,\ldots,m \tag{2.14}$$

olarak hesaplanır.

Adım5: Değişken ağırlıkları

$$w_j = \frac{C_j}{\sum_{j=1}^{m} C_j}, \quad j = 1,2,\ldots,m \tag{2.15}$$

biçimindedir.

Hem ENTROPY hem de CRITIC yöntemleri ile elde edilen değişken ağırlıkları $\boldsymbol{w} = \begin{bmatrix} w_1 & w_2 & \cdots & w_m \end{bmatrix}$, $j = 1,2,\ldots,m$ ile X veri matrisi ağırlıklandırılır. Buna göre, çizelge 2.1'de görülen çok değişkenli veri seti çizelge 2.3'teki biçimde ağırlıklandırılmış veri seti olarak elde edilir.

Çizelge 2.3 Çok değişkenli bir veri setinin ağırlıklandırılması

	Bağımsız değişkenler				***Bağımlı değişken***
No.	w_1X_1	w_2X_2	…	w_mX_m	Y
1	w_1x_{11}	w_2x_{12}	…	w_mx_{1m}	y_1
2	w_1x_{21}	w_2x_{22}	…	w_mx_{2m}	y_2
⋮	⋮	⋮	⋱	⋮	⋮
n	w_1x_{n1}	w_2x_{n2}	…	w_mx_{nm}	y_n

Buna göre, $\zeta_i = w_i X_i$, $i = 1,2,\ldots,m$ olarak adlandırıldığında çizelge 2.4'teki ağırlıklandırılmış veri seti olur.

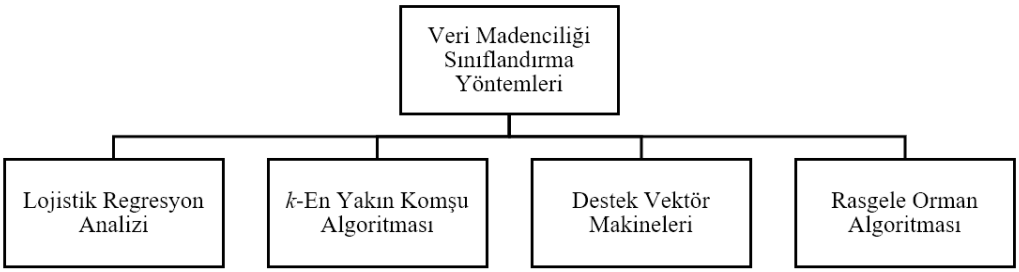
Çizelge 2.4 Ağırlıklandırılmış çok değişkenli bir veri seti

	Bağımsız değişkenler				***Bağımlı değişken***
No.	ζ_1	ζ_2	$\cdots$	ζ_m	Y
1	ζ_{11}	ζ_{12}	$\cdots$	ζ_{1m}	y_1
2	ζ_{21}	ζ_{22}	$\cdots$	ζ_{2m}	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	ζ_{n1}	ζ_{n2}	$\cdots$	ζ_{nm}	y_n

Böylece, çizelge 2.4'te verilen ağırlıklandırılmış veri setinin elde edilmesi ile veri ön işleme süreci tamamlanarak veri matrisi, sınıflandırma analizine hazır hale getirilmiş olur.

3. VERİ MADENCİLİĞİ YÖNTEMLERİ İLE SINIFLANDIRMA

Ön işleme süreci tamamlanan veri setine veri madenciliği sınıflandırma yöntemleri uygulanır. Veri madenciliği sınıflandırma yöntemleri, verideki gizli örüntülerin ortaya çıkarılarak gözlemlerin hangi sınıfa ait olduğunun tahmin edilmesini sağlayan denetimli öğrenme yöntemleridir. Denetimli öğrenmede, her bir girdiye karşılık bir çıktı değeri var olduğundan veri madenciliği sınıflandırma yöntemlerinde uygulanan algoritmaların öğrenme yoluyla verdiği cevaplar denetlenir. Uygulanan algoritmaların öğrenme modelinin geliştirilmesinde yardımcı ayarlanabilir parametreler kullanılır. Her bir sınıflandırma yöntemine ilişkin ayarlanabilir parametreler bulunur. Ayarlanabilir parametreler araştırmacı tarafından belirlenir (Geron 2017). Sınıflandırma yöntemlerinin model performansı optimize edilerek en iyi ayarlanabilir parametreler elde edilir.



Şekil 3.1 Veri madenciliği sınıflandırma yöntemleri

Bu çalışmada şekil 3.1'de verilen veri madenciliği sınıflandırma yöntemleri kullanılmıştır. Yöntemlere ilişkin açıklayıcı bilgiler sırasıyla aşağıda verilmiştir.

3.1 Lojistik Regresyon Analizi

Lojistik regresyon (Logistic Regression-LR), veri setindeki gözlemlerin bir sınıfa ait olup olmama olasılığını hesaplayarak yeni gelecek gözlemleri sınıflandıran bir veri madenciliği yöntemidir. LR analizinde, bağımlı değişkenin kategorik bir değişken olmasından dolayı en küçük kareler yöntemi ile tahmini yetersiz kalmaktadır. (Çokluk 2010). Model parametrelerinin tahmini, maksimum olabilirlik yöntemi uygulanarak hesaplanır.

$Y_i \in \{0,1\}$ iki sınıflı bir bağımlı değişkene sahip $\{(X_i, Y_i)\}_{i=1}^{n}$ biçimde gözlem çifti ile tanımlanmış bir veri seti olsun. $Y_i = 0$ ve $Y_i = 1$, $i = 1,2,\ldots,n$ olarak etiketlenmiş iki sınıf oluşturulacak doğrusal regresyon modeli

$$Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij} + \varepsilon_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{im} + \varepsilon_i, \; i = 1,2,\ldots,n \qquad (3.1)$$

biçiminde tanımlansın. Eşitlik (3.1) ile tanımlı model denklemi $(-\infty, +\infty)$ arasında değerler alacağından $0 < Y_i < 1$ değerini alacağı söylenemez. Bu nedenle, bağımlı değişkenin bir olasılık değerine dönüştürülmesi gerekir. Bu dönüşüm sırasında bağımlı değişken ile bağımsız değişken arasında doğrusal ilişkiyi sağlayacak link fonksiyonları kullanılır.

$lojit(p) = ln\dfrac{p}{(1-p)}$ dönüşümü yapılarak bağımlı değişkenin sınırları genişletilir. $\boldsymbol{X}_i = \boldsymbol{x}_i$ ve $t \in \mathbb{R}$ olmak üzere $t = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots + \beta_m x_{im}$, $p_i \in (0,1)$ için

$$\begin{aligned} y_i = 1 &\Rightarrow P(Y_i = 1) = P(Y_i = 1 \mid x_{i1}, x_{i2}, \ldots, x_{im}) = p_i \\ y_i = 0 &\Rightarrow P(Y_i = 0) = (1 - P(Y_i = 1 \mid x_{i1}, x_{i2}, \ldots, x_{im})) = p_i(1 - p_i) \end{aligned} \qquad (3.2)$$

olur. Bağımlı değişkenin olasılığı hesaplanarak lojit dönüşüm yapılırsa

$$\begin{aligned} lojit(p_i) &= ln\left(\frac{p_i}{1-p_i}\right) = t \\ e^{ln(\frac{p_i}{1-p_i})} &= e^t \\ \frac{p_i}{1-p_i} &= e^t \Rightarrow p_i = e^t(1-p_i) \\ p_i &= e^t - e^t p_i \Rightarrow p_i(1+e^t) = e^t \\ p_i &= \frac{e^t}{1+e^t}\frac{e^{-t}}{e^{-t}} \\ p_i &= \frac{1}{1+e^{-t}} \end{aligned} \qquad (3.3)$$

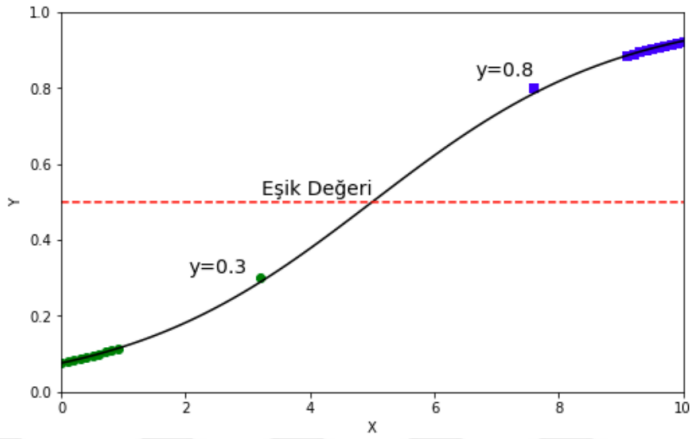
biçiminde lojit fonksiyonu elde edilir. $\zeta_i = w_i x_i$, $i = 1,2,\ldots,m$ biçiminde ağırlıklandırıldığında LR modeli

$$f_{LR}(\zeta_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1\zeta_{i1} + \beta_2\zeta_{i2} + \ldots + \beta_m\zeta_{im})}}, \; i = 1,2,\ldots,n \qquad (3.4)$$

olur. Burada, $\zeta_i = [\zeta_{i1} \quad \zeta_{i2} \quad \ldots \quad \zeta_{im}]$, $i = 1,2,\ldots,n$'dir. Eşitlik (3.4) bir olasılık değeri olup eşik değeri 0.5 alındığında sınıflandırma

$$f(\zeta_i) \geq 0.5, \ \Rightarrow \hat{Y}_i = 1$$
$$f(\zeta_i) < 0.5, \ \Rightarrow \hat{Y}_i = 0 \tag{3.5}$$

biçiminde yapılır. Şekil 3.2'de eşik değerine göre sınıflandırılma gösterilmiştir.



Şekil 3.2 LR ile iki değişkenli veri setinin sınıflandırılması

Şekil 3.2'de $f_{LR}(\zeta_i)$ fonksiyon değeri eşik değeri olan 0.5'ten küçük olanlar yeşil, büyük olanlar mavi olacak biçimde sınıflandırılmıştır. Model parametrelerinin tahmin edilmesinde maksimum olabilirlik yöntemi kullanılarak eşitlik (3.4)'te verilen model maksimize edilir. $f_{LR}(\zeta_i)$'in olabilirlik fonksiyonu

$$L(\beta_0, \beta_1, \ldots, \beta_m) = \prod_{i=1}^{n} f_{LR}(\zeta_i)^{y_i} (1 - f_{LR}(\zeta_i))^{(1-y_i)} \tag{3.6}$$

olmak üzere lnL fonksiyonu

$$lnL(\beta_0, \beta_1, \ldots, \beta_m) = \sum_{i=1}^{n} (y_i \ ln \ f_{LR}(\zeta_i) + (1 - y_i) \ ln(1 - f_{LR}(\zeta_i)) \tag{3.7}$$

yazılır. Model parametreleri

$$\underset{\beta_0, \beta_1, \ldots, \beta_m \in \mathbb{R}}{maks} \ lnL(\beta_0, \beta_1, \ldots, \beta_m)$$
$$g(\boldsymbol{\beta}) = c \sum_{j=1}^{m} \left| \beta_j \right|^t \geq 0, \quad t = 1, 2, \ c \geq 0 \tag{3.8}$$

biçiminde tanımlı eşitsizlik kısıtlı doğrusal olmayan optimizasyon probleminin çözümü Kuhn-Tucker koşullarından yararlanılarak tahmin edilir. Burada, $\boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \ldots \quad \beta_m]$ parametre vektörüdür.

LR Algoritması:

Adım1: Bağımlı değişken ve bağımsız değişkenler arasında lojit bir ilişki olduğu varsayılır.

Adım2: Bağımlı değişken, $lojit(p) = ln\frac{p}{(1-p)}$ dönüşümü kullanılarak bir olasılık değerine dönüştürülür. Dönüşüm sonrası, bağımsız değişken ve regresyon katsayılarına bağlı lojit fonksiyon elde edilir.

Adım3: Lojit fonksiyonda gerekli düzenlemeler yapılarak LR modeli elde edilir.

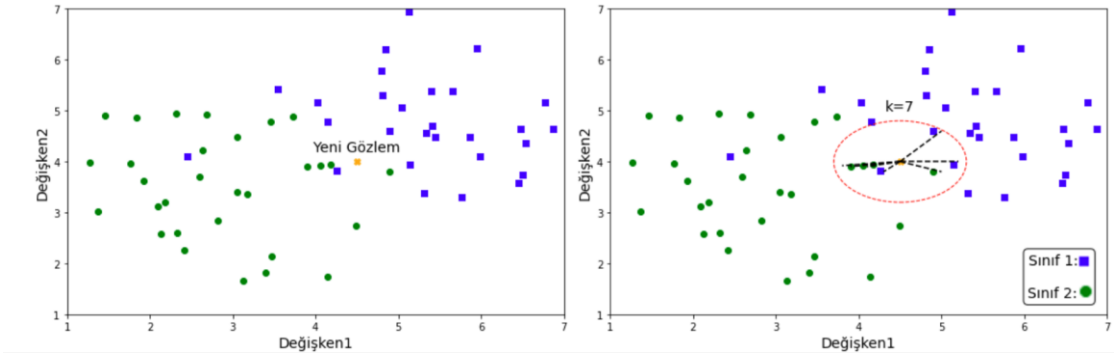
Adım4: LR ile bağımlı değişkenin alabileceği değerlerin gerçekleşme olasılığı hesaplanır. Eşik değeri belirlenerek gözlemler sınıflandırılır.

Adım5: Maksimum olabilirlik yöntemi kullanılarak LR katsayıları elde edilir.

LR'nin ayarlanabilir parametreleri, kısıtlama yapılacak yöntemi (Lasso $(t = 1)$, Ridge $(t = 2)$, None $(c = 0)$) belirleyen ceza değeri (penalty) ve kısıtlama oranı (c)'dır (Geron 2017).

3.2 *k*-En Yakın Komşu Algoritması

k-En Yakın Komşu Algoritması (*k*-Nearest Neigbors-*k*NN), veri madenciliği sınıflandırma yöntemlerinden biridir. *k*NN, tahmin modeli oluşturulmasına ihtiyaç duyulmaması nedeniyle genellikle tercih edilen bir sınıflandırma yöntemidir (Silahtaroğlu 2008, Özkan 2016). Bu yöntemde, sınıfı bilinmeyen yeni bir gözlem ile sınıfı bilinen gözlemlerin her biri arasındaki uzaklıklar hesaplanır. Yeni gözleme en yakın uzaklığa sahip olan *k* sayıda gözlemin seçimi yapılır. *k* sayıda gözleme ait sınıf arasından en çok tekrar eden sınıf, yeni gözlemin sınıfı olarak tahmin edilir. Şekil 3.3'te yeni gözlemin sınıflandırılması gösterilmiştir.



Şekil 3.3 *k*NN ile iki değişkenli veri setinin sınıflandırılması

Sınıfı bilinmeyen yeni bir gözlem sınıflandırılmak istendiğinde yeni gözlem ile diğer tüm gözlemler arasındaki uzaklıklar hesaplanır. Şekil 3.3'e göre optimal k komşu sayısı $k=7$ olarak belirlenmiştir. Yeni gözleme en yakın yedi (7) komşu gözleme ait sınıflar arasında en çok tekrar eden sınıf yeni gözlemin sınıfı olarak belirlenir.

Değişken sayısının fazla ya da gözlem sayısının büyük olduğu veri setlerinde, veri noktaları arasındaki uzaklıkların ölçülmesi oldukça zordur. Gözlemler arasındaki uzaklıkların belirlenmesinde farklı uzaklık ölçüm yöntemleri kullanılabilir.

Minkowski uzaklığı: m değişken sayılı veri setindeki $i.$ ve $j.$ gözlemler arasındaki uzaklıkların mutlak değerlerinin alınması ile

$$d_{ij}=\left(\sum_{t=1}^{m}\left|\boldsymbol{x}_{it}-\boldsymbol{x}_{jt}\right|^{p}\right)^{1/p},\ i,j=1,2,\ldots,n,\ i\neq j,\ p\geq 1 \qquad (3.9)$$

elde edilir. Bu uzaklık ölçütü genel bir uzaklık denklemi olup seçilen p değerine göre farklı uzaklık ölçütleri elde edilir. Eşitlik 3.9'da, $p=1$ için Manhattan uzaklığı, $p=2$ için Öklid uzaklık ölçütü hesaplanır. *k*NN algoritmasında, mevcut uzaklık ölçütleri arasında Öklid uzaklık formülü yaygın olarak kullanılır. $\zeta_i=w_i x_i,\ i=1,2,\ldots,m$ biçiminde ağırlıklandırıldığında *k*NN modeli

$$f_{kNN}(\zeta_i,\zeta_j)=\sqrt{\sum_{t=1}^{m}\left(\zeta_{it}-\zeta_{jt}\right)^2},\ i,j=1,2,\ldots,n,\ i\neq j \qquad (3.10)$$

biçiminde elde edilir. Burada, k ayarlanabilir parametrenin belirlenmesi önemlidir. En küçük uzaklığın elde edilmesi amaçlanır. Buna göre,

$$\min_{t\in\mathbb{R}} f_{kNN}(\zeta_i, \zeta_j) \tag{3.11}$$

biçiminde tanımlı kısıtsız doğrusal olmayan optimizasyon problemi çözülerek yeni gelecek gözleme en yakın uzaklıklar hesaplanır. Doğrusal olmayan optimizasyon problem çözümü için türeve dayalı optimizasyon yöntemleri kullanılır (Türkşen 2023).

kNN Algoritması:

***Adım1*:** Yeni gözlem ile diğer tüm gözlemler arasındaki uzaklıklar hesaplanır.

Adım2: Hesaplanan uzaklıklar, en küçükten en büyüğe doğru sıralanarak uzaklık matrisi oluşturulur.

Adım3: Optimal k parametresi belirlenir. Bu parametre yeni gözleme en yakın komşu sayısını ifade eder.

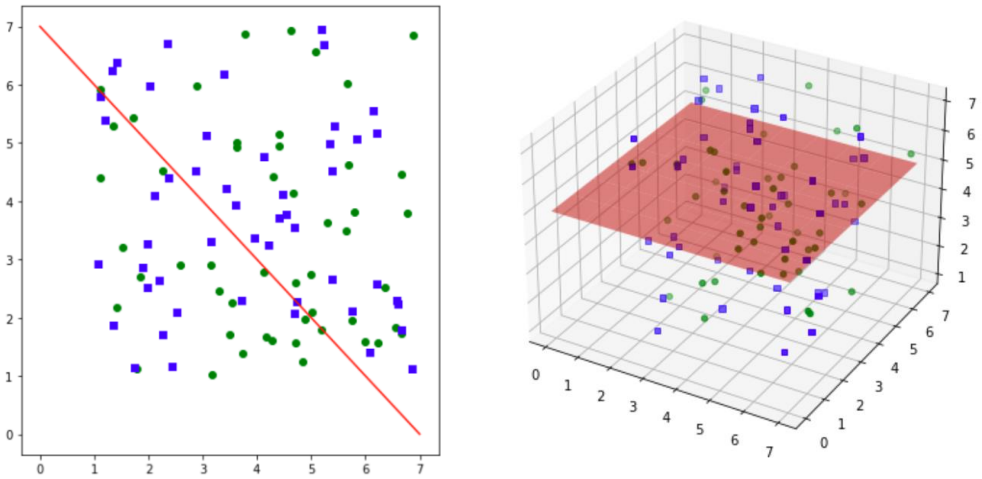
Adım4: Yeni gözleme en yakın uzaklıklara sahip k sayıda gözleme ait en çok tekrar eden sınıf seçilir.

Adım5: Seçilen sınıf, tahmin edilmek istenen yeni gözlem değerinin sınıfıdır.

3.3 Destek Vektör Makineleri

Destek Vektör Makineleri (Support Vector Machine-SVM), Cortes ve Vapnik (1995) tarafından geliştirilen veri madenciliği yöntemlerinden biridir. Bu yöntem, veriyi ayırmak için en uygun fonksiyonun tahmin edilmesi esasına dayanır. Hem doğrusal ve hem doğrusal olmayan sınıflandırmalar SVM ile yapılabilmektedir. SVM, verileri etkili bir biçimde sınıflandırmak için matematiksel analiz, doğrusal cebir ve optimizasyon prensiplerini kullanır. Bu nedenle daha karmaşık bir algoritma yapısına sahiptir. Aykırı değer tespitinde de kullanılır (Geron 2021).

(a) (b)



Şekil 3.4 SVM ile sınıflandırılma (a) İki değişkenli verinin bir doğru ile sınıflandırılması, (b) Üç değişkenli verinin bir düzlem ile sınıflandırılması

SVM ile veriler şekil 3.4 (a)'daki gibi iki değişkenli doğru, şekil 3.4 (b)'deki gibi üç değişkenli düzlem, daha yüksek değişken sayısında ise hiperdüzlem kullanılarak sınıflandırılır.

3.3.1 Destek vektör makineleri ile doğrusal sınıflandırma

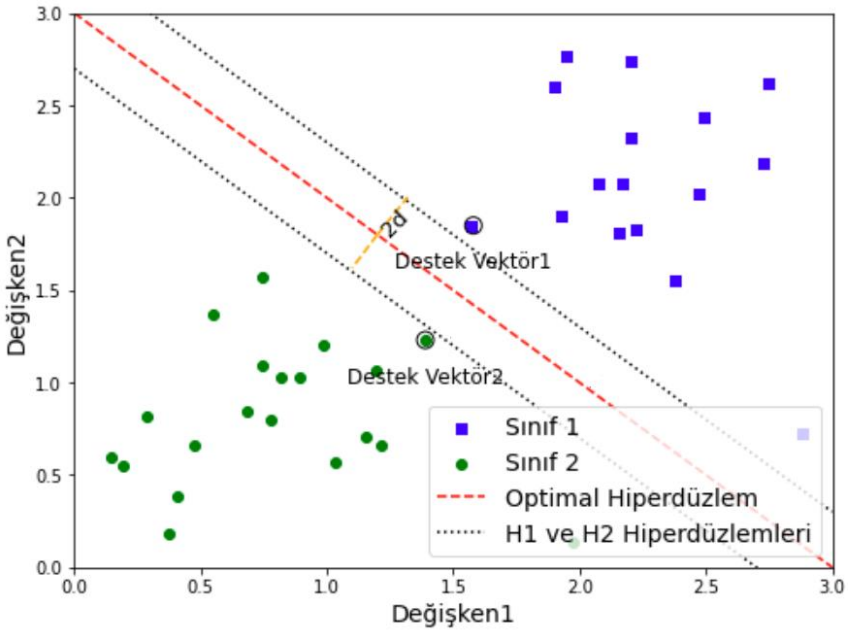
$Y_i \in \{-1,1\}$ iki sınıflı bağımlı değişken olmak üzere $\{(X_i, Y_i)\}_{i=1}^{n}$ biçimde gözlem çifti ile tanımlanmış bir veri seti olsun. $Y_i = -1$ ve $Y_i = 1$ olarak etiketlenmiş iki sınıf H hiperdüzlemi ile ayrılır. $\boldsymbol{X}_i = \boldsymbol{x}_i$ olmak üzere iki sınıf arasındaki karar sınırı olan H hiperdüzlemi

$$H\ düzlemi: \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + \beta_0 = 0 \tag{3.12}$$

elde edilir. Burada, nokta çarpımları

$$\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle = \boldsymbol{\beta}^T \boldsymbol{x} = \sum_{j=1}^{m} \beta_j x_j \tag{3.13}$$

biçiminde ifade edilir.



Şekil 3.5 SVM ile iki değişkenli veri setinin sınıflandırılması

İki sınıfı ayıran hiperdüzlem hattı şekil 3.5'te görülmektedir. H düzlemi burada optimal hiperdüzlem olup H_1 düzlemi ve H_2 düzlemi sırasıyla

$$\begin{aligned} H_1\,düzlemi &: \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + \beta_0 - 1 = 0 \\ H_2\,düzlemi &: \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + \beta_0 + 1 = 0 \end{aligned} \tag{3.14}$$

biçiminde ifade edilir. $\boldsymbol{\beta}$ ağırlık vektörü, hiperdüzleme dik yönde olup eğimi belirlemektedir. β_0 sabit terimdir. H_1 düzlemi ve H_2 düzlemi üzerindeki gözlemler, sınırları belirleyen destek vektörleri olarak tanımlanmaktadır. Destek vektörlerinin seçimi yapılırken H_1 düzlemi ve H_2 düzlemleri ile H düzlemi arasındaki uzaklık olan kenar payı (d)

$$d = \frac{\left| \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + \beta_0 \right|}{\|\boldsymbol{\beta}\|} = \frac{1}{\|\boldsymbol{\beta}\|} \tag{3.15}$$

biçiminde elde edilir. H düzlemine kenar payı uzaklıkları d olan H_1 ve H_2 düzlemlerinin denklemleri eşitlik (3.14) kullanılarak

$$\begin{aligned} H_1\,düzlemi &: \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + \beta_0 - d = 0 \\ H_2\,düzlemi &: \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + \beta_0 + d = 0 \end{aligned} \tag{3.16}$$

biçiminde yazılır. Kenar payı uzunluğunun en büyük olacak biçimde, gözlemlerin sınıflandırılması amaçlanır. Sınıflandırma işlemi yapılırken karşı sınıfta en az sayıda

farklı sınıftan gözlemin olması ve H_1 düzlemi ile H_2 düzlemi üzerinde en az birer destek vektörünün bulunması istenir (Metlek ve Kayaalp 2020). Kenar payı $2d = \frac{2}{\|\boldsymbol{\beta}\|}$ değerinin en büyük yapılması amaçlanır. Buna göre

$$\min_{\beta \in \mathbb{R}} \frac{1}{2}\|\boldsymbol{\beta}\|^2 \tag{3.17}$$

biçiminde yazılır (Fletcher 2009; Pehlivanoğlu 2017; Arlı vd. 2022).

Sabit kenar payı ile SVM sınıflandırması

Sabit kenar payı ile sınıflandırmada, her sınıfa ait gözlem diğer sınıfta olmayacak biçimde gözlemler sınıflandırılır. Veriyi doğrusal bir biçimde ayıran H hiperdüzlemi

$$\left.\begin{array}{l} \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \beta_0 \geq 1, \; y_i = 1 \\ \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \beta_0 \leq -1, \; y_i = -1 \end{array}\right\} \; y_i\left(\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \beta_0\right) \geq 1, \; i = 1, 2, \ldots, n \tag{3.18}$$

biçimindedir. Sabit kenar payı ile SVM sınıflandırması

$$\begin{aligned} &\min_{\beta \in \mathbb{R}} \; f(\boldsymbol{\beta}) = \frac{1}{2}\langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle \\ &g(\boldsymbol{\beta}, \beta_0) = y_i\left(\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \beta_0\right) - 1 \geq 0 \,, \; i = 1, 2, .., n \end{aligned} \tag{3.19}$$

uygun bir biçimde birleştirilip eşitsizlik kısıtlı doğrusal olmayan optimizasyon problemi gibi yazılır. Optimal hiperdüzlem parametrelerinin tahmin edilmesi için Lagrange çarpanlarından yararlanılır. Lagrange çarpanı $\lambda_i \geq 0$ olmak üzere Lagrange fonksiyonu

$$L(\boldsymbol{\beta}, \beta_0) = \frac{1}{2}\langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle - \sum_{i=1}^{n} \lambda_i y_i \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \beta_0 + \sum_{i=1}^{n} \lambda_i \tag{3.20}$$

elde edilir. Denklem sistemi çözülerek

$$\begin{aligned} &\frac{\partial L(\boldsymbol{\beta}, \beta_0)}{\partial \boldsymbol{\beta}} = 0 \;\; \Rightarrow \boldsymbol{\beta} = \sum_{i=1}^{n} \lambda_i y_i \boldsymbol{x_i} \\ &\frac{\partial L(\boldsymbol{\beta}, \beta_0)}{\partial \beta_0} = 0 \;\; \Rightarrow \sum_{i=1}^{n} \lambda_i y_i = 0 \end{aligned} \tag{3.21}$$

$\boldsymbol{\beta}$ ve β_0 elde edilir. Elde edilenler eşitlik (3.20)'de yerine konulup Kuhn-Tucker koşullarından yararlanılarak problem çözülür. $\zeta_i = w_i x_i, \; i = 1, 2, \ldots, m$ biçiminde ağırlıklandırıldığında SVM ile

$$f_{SVM}(\boldsymbol{\zeta}) = sign(\boldsymbol{\beta}^T \boldsymbol{\zeta} + \beta_0) \tag{3.22}$$

biçiminde sınıflandırılır.

Esnek kenar payı ile SVM sınıflandırması

Bazı durumlarda gözlemleri keskin bir biçimde sınıflandırmak mümkün değildir. Cortes ve Vapnik (1995) esnek kenar payı yaklaşımını geliştirmişlerdir. Esnek kenar payında, bazı gözlemlerin karşı sınıfa geçmesine izin verilir. Böylece daha esnek bir sınıflandırma yapılır. Esnek kenar payı kullanılarak

$$y_i\left(\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \beta_0\right) - 1 + \xi_i \geq 0, \quad \xi_i \geq 0, \ i = 1, 2, \ldots, n \tag{3.23}$$

hiperdüzlemin pozisyonu değiştirilir. $\xi_i \geq 0, \ i = 1, 2, \ldots, n$ olmak üzere ξ_i esnek değişkeni gözlemlerin hatalı sınıflandırılmasına izin vererek destek vektörlerin konumlarının değiştirilmesini ve daha esnek bir sınıflandırma yapılmasını sağlar. Esnetmenin kabul edilebilir bir büyüklükte olması önemlidir. Esnek kenar payı ile toplam hatalı sınıflandırılan gözlem sayısının $(\sum_{i=1}^{n} \xi_i)$ en az olması istenir. $c \in \mathbb{R}^+$ olmak üzere c parametresi ile $\sum_{i=1}^{n} \xi_i$ sınırlandırılır. c parametresinin çok küçük seçilmesi, geniş bir kenar payı nedeniyle sınır ihlalini artırır. c parametresinin çok büyük seçilmesi ise daha dar kenar payına neden olup sınır ihlal toleransını oldukça azaltır. Esnek kenar payı ile SVM sınıflandırması

$$\begin{aligned} &\min_{\beta \in \mathbb{R}} \ f(\boldsymbol{\beta}) = \frac{1}{2}\langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle + c\sum_{i=1}^{n} \xi_i \\ &g_1(\boldsymbol{\beta}, \beta_0) = y_i\left(\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \beta_0\right) - 1 + \xi_i \geq 0 \\ &g_2(\xi) = \xi_i \geq 0, \ i = 1, 2, .., n \end{aligned} \tag{3.24}$$

biçiminde tanımlı eşitsizlik kısıtlı doğrusal olmayan optimizasyon probleminin çözümü ile elde edilir. Lagrange fonksiyonu

$$L(\boldsymbol{\beta}, \beta_0, \xi) = \frac{1}{2}\langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle + c\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \lambda_i y_i\left(\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \beta_0 - 1 + \xi_i\right) - \sum_{i=1}^{n} \alpha_i \xi_i \tag{3.25}$$

olur. Lagrange çarpanı $0 \leq \lambda_i \leq c$ olmak üzere denklem sistemi çözülerek Lagrange fonksiyonu λ parametresine bağlı

$$L(\lambda)=\sum_{i=1}^{n}\lambda_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_i y_j\lambda_i\lambda_j\left\langle \boldsymbol{x}_i,\boldsymbol{x}_j\right\rangle,\ i,j=1,2,\ldots,n,\ i\neq j \tag{3.26}$$

biçimde yazılır. Kuhn-Tucker koşullarından yararlanılarak model parametreleri tahmin edilir.

3.3.2 Destek vektör makineleri ile doğrusal olmayan sınıflandırma

Veri seti bazen doğrusal olarak sınıflandırılamayabilir ya da sınıflandırılması anlamlı olmayabilir. Doğrusal sınıflandırma mümkün olmadığında değişken eklenerek veri seti daha fazla değişkene sahip bir yapıya dönüştürülür. Bu dönüştürme ile verilerin doğrusal olarak ayrılabilir hale gelmesi amaçlanır. Fakat bu durum, matematiksel olarak hesaplama zorluğu getirmektedir. Bu durumun çözümünde çekirdek (kernel) fonksiyonu kullanılır.

Çekirdek fonksiyonu hesaplama zorunluluğu olmadan iç çarpım sonuçlarını dönüştürerek doğrusal olmayan sınıflandırma problemlerini çözmek için kullanılan matematiksel bir fonksiyondur. Çekirdek fonksiyonu, doğrudan hesaplama gerektirmeden fazla değişkenli verinin iç çarpım sonuçlarını dönüştürerek fonksiyonda kolaylık sağlar (Ayhan ve Erdoğmuş 2014). Çekirdek fonksiyonu ile doğrusal olarak sınıflandırılamayan verinin boyutu artırılarak bir düzlem ile ayırmak mümkündür. Bu nedenle, doğrusal olmayan sınıflandırma yönteminde çekirdek fonksiyonuna ihtiyaç duyulmaktadır.

ϕ dönüşüm fonksiyonu, az sayıda değişkenden daha fazla sayıda değişkene geçiş işlemi olarak tanımlanır.

$\boldsymbol{x}_i,\boldsymbol{x}_j\in\mathbb{R}^2$ olmak üzere $\boldsymbol{x}_i=(x_{i1},x_{i2}),\ \boldsymbol{x}_j=(x_{j1},x_{j2}),\ i,j=1,2,\ldots,n$ olsun.

$\phi:\mathbb{R}^2\rightarrow\mathbb{R}^3$ dönüşüm fonksiyonu kullanılarak

$$\begin{aligned}(x_{i1},x_{i2})&\xrightarrow{\phi}(x_{i1},x_{i2},x_{i3})=x_{i1}{}^2+\sqrt{2}x_{i1}x_{i2}+x_{i2}{}^2\\(x_{j1},x_{j2})&\xrightarrow{\phi}(x_{j1},x_{j2},x_{j3})=x_{j1}{}^2+\sqrt{2}x_{j1}x_{j2}+x_{j2}{}^2\end{aligned} \tag{3.27}$$

daha fazla değişken ile temsil edilir. Dönüşüm sonrası değişken vektörlerinin iç çarpımları

$$\left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle = \left\langle \phi_{x_i}, \phi_{x_j} \right\rangle$$
$$\left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle = (x_{i1}{}^2 + \sqrt{2}x_{i1}x_{i2} + x_{i2}{}^2)(x_{j1}{}^2 + \sqrt{2}x_{j1}x_{j2} + x_{j2}{}^2) = (x_{i1} + x_{j2})^2 \quad (3.28)$$
$$\left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle = \left(\left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle\right)^2$$

biçiminde yazılır. Eşitlik (3.28)'de, $\boldsymbol{x}_i$ ve $\boldsymbol{x}_j$ vektörlerinin çarpımı, dönüşüm öncesi iç çarpımlarına eşittir. Böylece dönüşüme gerek duyulmadan sadece girdi değişkenlerinin iç çarpımı kullanılarak işlem yapılır (Uğuz 2021). Çekirdek fonksiyonu

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left\langle \phi_{x_i}, \phi_{x_j} \right\rangle = \left(\left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle\right)^2 \quad (3.29)$$

biçiminde tanımlanır. Literatürde sıkça kullanılan çekirdek fonksiyonları çizelge 3.1'de verilmiştir.

Çizelge 3.1 Yaygın kullanılan çekirdek fonksiyonları

***Doğrusal*:** $K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle\right)$
***Polinom*:** $K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle + 1\right)^m, \ m > 1, \ m \in Z^+$
***Dairesel (Radyal) Tabanlı*:** $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|^2), \ \gamma = \frac{1}{2}\sigma^{-2}$
***Sigmoid*:** $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\gamma \boldsymbol{x}_i^T \boldsymbol{x}_j + r), \ \gamma, r \in \mathbb{R}$

Bir SVM doğrusal sınıflandırma probleminin amaç fonksiyonunda $\left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle$ iç çarpımı yer alıyorsa, yerine uygun bir $K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ çekirdek fonksiyonu yazılır.Eşitlik (3.26)'da gerekli düzenleme yapıldığında

$$L(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \lambda_i \lambda_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right), \ i, j = 1, 2, \ldots, n, \ i \neq j \quad (3.30)$$

olur.

SVM Algoritması:

Adım1: Optimal hiperdüzlem denklemi oluşturulur.

Adım2: Her bir gözlem H hiperdüzlem denkleminde yerine konularak H hiperdüzlemin altında ya da üstünde kalacak şekilde ayrılır.

Adım3: Sınıflandırma işlemi yapılırken karşı sınıfta en az sayıda farklı sınıftan gözlemin olması, H_1 düzlemi ile H_2 düzlemi üzerinde en az birer destek vektörünün bulunması istenir.

Adım4: Kenar payının en büyük olması amaçlanır. Eğer, sabit kenar payı ile sınıflandırma yapılamıyorsa hatalı değişkenlere izin veren esnek değişken $\xi_i,\ i=1,2,\ldots,n$ ile onu sınırlandıran c parametresi belirlenerek esnek sınıflandırma yapılır.

Adım5: Doğrusal olarak sınıflandırma yapılamıyorsa uygun çekirdek fonksiyon yardımıyla doğrusal olmayan sınıflandırma yapılır.

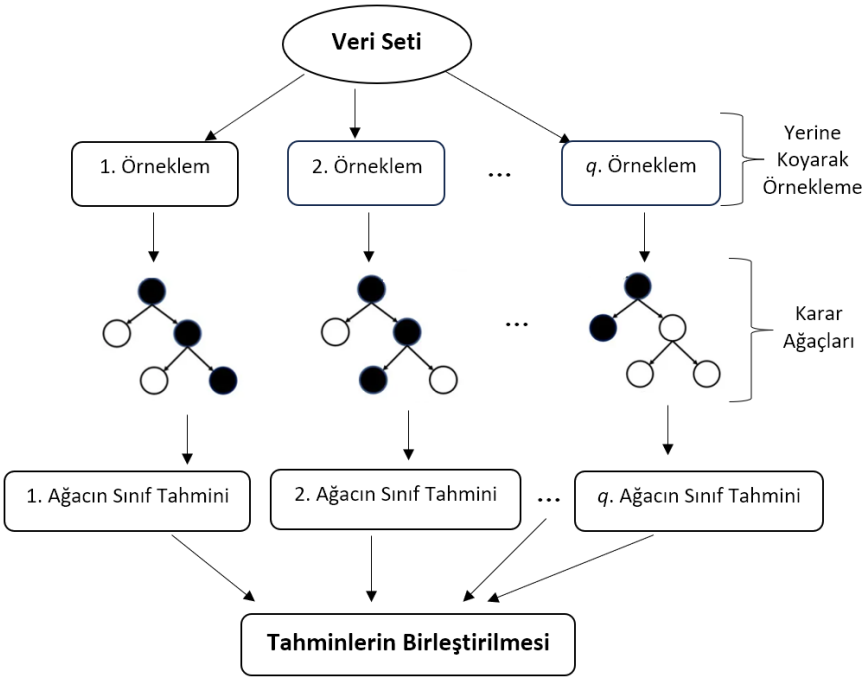
SVM'de kullanılan ayarlanabilir parametreler çizelge 3.2'de verilmiştir.

Çizelge 3.2 SVM'de kullanılan ayarlanabilir parametreler

c **:** Kısıtlama oranı
Kernel**:** Çekirdek fonksiyon (Polynomial, Rbf, Sigmoid, Linear)
Gamma: c oranını düzenleyici değer

3.4 Rastgele Orman Algoritması

Rastgele Orman (Random Forest-RF) Algoritması, Breiman (2001) tarafından geliştirilen bir sınıflandırma yöntemidir. RF, bir veri setinin rastgele seçilmiş alt kümelerinde eğitilen birden fazla karar ağacını içerir. RF, tek bir karar ağacının sağladığı tahminlere dayanmak yerine ormandaki her bir ağacın tahminlerini birleştirir. Şekil 3.6'da RF ile sınıflandırılma gösterilmiştir.



Şekil 3.6 RF ile sınıflandırma

Şekil 3.6'da veri setinden yerine koyarak örnekleme yöntemi (Bootstrap) ile seçilen her örneklemden çeşitli karar ağaçları oluşturulur. Her bir karar ağacı kendi sınıf tahminini yaparak nihai sonuç bu tahminlerin çoğunluk oyu ile belirlenir. Tek bir karar ağacıyla kıyaslandığında ormandaki ağaç sayısının artması modelin performansını artırır.

RF yönteminin temel avantajlarından biri, çeşitli değişken türlerine (kategorik ve sürekli) sahip farklı büyüklükte örneklem çapı olan veri setlerinde etkili ve esnek bir biçimde uygulanabilmesidir (Korkmaz vd. 2018). RF, aşırı öğrenme riskini azaltmaktadır. RF, en güçlü makine öğrenmesi modellerinden biri olarak görülmektedir (Geron 2021).

RF, veri setindeki m sayıdaki tüm değişkenleri kullanmak yerine her bir düğümde $s = \sqrt{m}$ sayıda değişkenleri rastgele seçer. Rastgele seçilen değişkenler arasından en iyisi seçilerek her bir düğüm dallara ayrılır. Her bir düğümde Gini ya da Entropi indeksi kullanılarak değişken seçimi yapılır. Her örneklem için ikili bölünme yapısı ile ağaç oluşturulur.

p_l , T. düğümdeki l. sınıfa ilişkin ilgilenilen durumun olasılığı olsun. Entropi indeksi

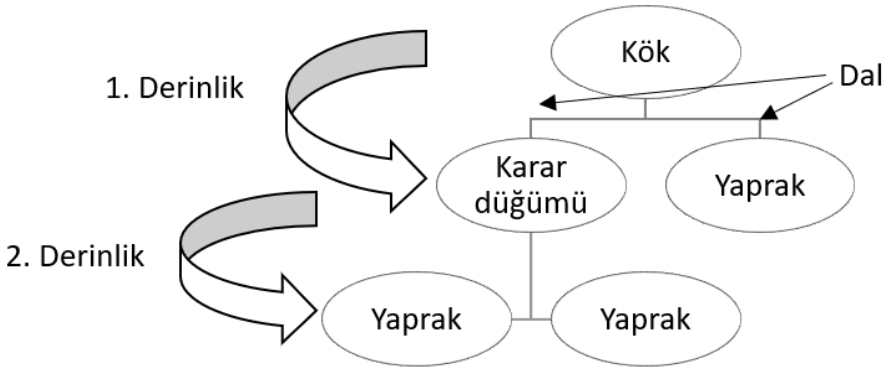
$$Entropi_T = -\sum_{l=1}^{2} p_l \log_2^{p_l} \tag{3.31}$$

biçiminde hesaplanır. Entropi, safsızlık ölçütü olarak bilinir. Veri kümesinde ne kadar düzensizlik var ise entropi o kadar yüksek olur. (Erden 2021). Veri kümesindeki tüm gözlemlerin sadece bir sınıfa atanabildiği durumda entropi değeri sıfır (tam saflık), sınıflara atanabilme olasılığı eşit olduğunda entropi değeri bire eşit olur. Gini indeksi ise, bir düğümdeki yanlış sınıflandırılma olasılığını ölçer. Gini indeksi

$$Gini_T = 1 - \sum_{l=1}^{2} p_l^2 \tag{3.32}$$

biçiminde hesaplanır. Düşük bir Gini indeks değeri, düğümün saflığını (bir sınıfın baskın olduğunu) ifade eder. Gini ya da Entropi indeks ile rastgele alt kümeler içerisinde rastgele değişken de aranarak modele fazladan rastgelelik getirilir. Bu adım ile daha iyi bir model elde edilmesi sağlanır.

Bir ağaçtaki her bir değişken bir düğüm ile temsil edilir. Bir ağacın yapısı göz önünde bulundurulduğunda şekil 3.7'deki gibi en üst yapıda kök, köklerin bölünmesiyle düğüm ve dallar, en sonunda yapraklar bulunur.



Şekil 3.7 RF'de bir ağacın yapısı

B oluşturulan ağaç sayısı olmak üzere $(b = 1, 2, \ldots, B)$, b ağacın sınıf tahmini $\hat{C}_b(\boldsymbol{x})$ 'dir. $\zeta_i = w_i x_i,\ i = 1, 2, \ldots, m$ biçiminde ağırlıklandırıldığında RF modeli

$$\hat{C}_{RF}^{B}(\boldsymbol{\zeta}) = \text{çoğunluk oylaması}\left\{\hat{C}_b(\boldsymbol{\zeta}_i)\right\}_1^B,\ i = 1, 2, \ldots, n \tag{3.33}$$

biçiminde yazılır. Ağaç performansının değerlendirilmesinde hata oranının minimum yapılması istenir (Arlı vd. 2022). Buna göre

$$\min_{\zeta \in \mathbb{R}} E_{OOB} = \frac{1}{n} \sum_{B=1}^{B} I(y_i - \hat{C}_{RF}^{B}(\zeta)) \tag{3.34}$$

biçiminde tanımlı kısıtsız doğrusal olmayan optimizasyon problemi çözülür.

RF Algoritması:

Adım1: Eğitim veri setinden yerine koyarak örnekleme yöntemi (Bootstrap) ile q sayıda örneklem seçilir.

Adım2**:** Her bir örneklemin 2/3'ü (In-bag (IB)) ile ağaç oluşturulur. Geriye kalan 1/3'ü (Out-of-bag (OOB)) ağacın performansını değerlendirilir.

Adım3**:** Oluşturulacak ağaç sayısı araştırmacı tarafından belirlenir. Her örneklem için ikili bölünme yapısı ile ağaç oluşturulur.

Adım4: Her düğüm noktasında, m değişken sayılı veri seti için $s < m$ olmak üzere $s = \sqrt{m}$ sayıda değişkenler rastgele seçilir. s sayıda değişkenler arasından Entropi ya da Gini indeksine dayalı bilgi kazancı en fazla olan değişken, dalları ayırıcı değişken olarak belirlenir. Her dal bir yaprakla sonuçlanana kadar işleme devam edilir.

Adım5: Her örneklemden bir sınıf tahmini yapılır. Tüm tahminlerin çoğunluk oylamasına göre sınıflandırma yapılır.

RF için ayarlanabilir parametreler Çizelge 3.3'te verilmiştir.

Çizelge 3.3 RF'de kullanılan ayarlanabilir parametreler

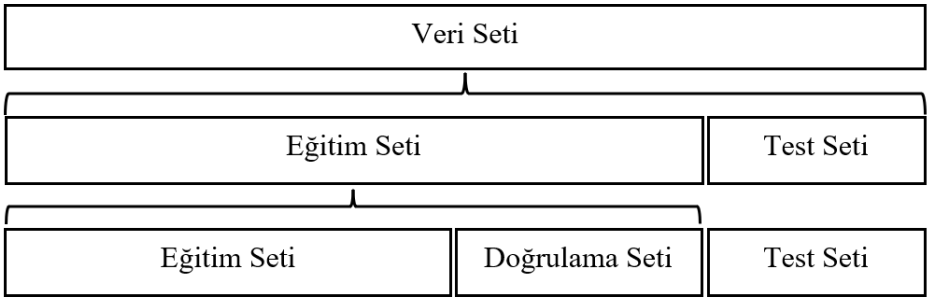
Maksimum derinlik (Max dept)**:** Ardışık soru sayısı
Maksimum değişken (Max features)**:** Her düğümde değerlendirilen değişken sayısı
Minumum yaprak örneği (Min samples leaf)**:** Bir yapraktaki minimum gözlem sayısı
Minimum örneklem (Min samples split): Bir düğüm bölünmeden önce gerekli gözlem sayısı
***n* tahmin edici (*n* estimators):** Ormanda oluşan ağaç sayısı (*B*)
Kriter (Criterion)**:** İndeks hesaplama yöntemi (Gini, Entropy)

4. SINIFLANDIRMA PERFORMANS ÖLÇÜTLERİ VE KARAR VERME

Veri madenciliği sınıflandırma yöntemlerinin performansları bakımından karşılaştırılabilmesi için ilgili ölçütlerin hesaplanması gerekir. Bunun için sınıflandırma performans ölçütlerine ilişkin hesaplamalar verilmiştir. Hesaplanan performans ölçütlerine göre uygun sınıflandırma yönteminin seçiminde MCDM yöntemleri kullanılmıştır.

4.1 Sınıflandırma Performans Ölçütleri

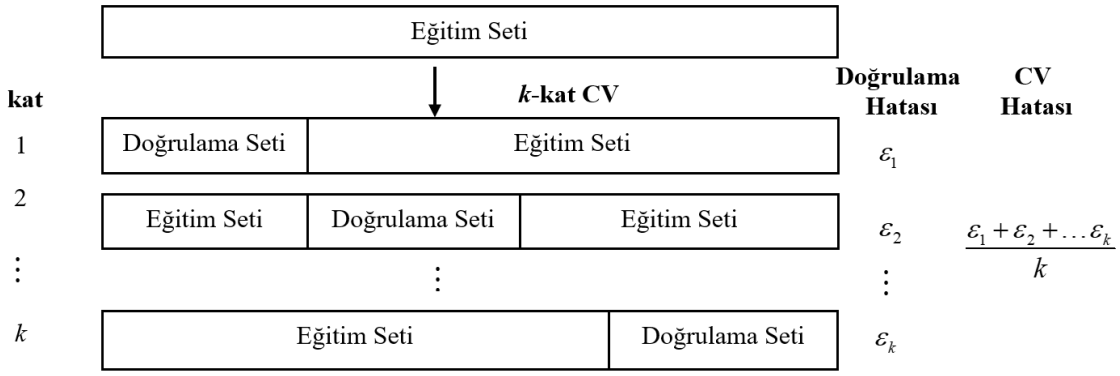
Sınıflandırma yöntemleri ile elde edilen modellerin tahmin performanslarının değerlendirilebilmesi için çeşitli performans ölçütleri bulunmaktadır. Sınıflandırma performans ölçütleri hesaplanmadan önce şekil 4.1'de verilen biçimde veri seti, eğitim, doğrulama ve test veri seti olmak üzere üç parçaya bölünür. Eğitim ve test veri setlerinin bölünme oranı genellikle %80 eğitim seti ile %20 test seti olmak üzere ayrılır. Eğitim veri setinin ise %20'si doğrulama seti olacak biçimde bölünür.



Şekil 4.1 Veri setinin eğitim, doğrulama ve test veri seti olarak bölünmesi

Eğitim veri seti ile sınıflandırma yöntemi eğitilerek model oluşturulur. Doğrulama veri seti, uygulanan yöntemin ayarlanabilir parametrelerini düzenleyerek en iyi modeli seçmek için kullanılır. Test veri seti ile öğretilen yöntemin performansı test edilir. Bazen, her ne kadar model eğitilse de verilen bilgi tam olarak öğrenilemez. Model eğitim sürecini ezberleyerek eğitim veri seti için doğru tahminde bulunsa da öğrenilen bilgiyi test verisi için genelleştiremez. Bu nedenle test performansını düşürür. Bu duruma aşırı öğrenme problemi denir. Modelin, eğitim veri setine duyarlılığı çapraz doğrulama ile sağlanır. Veri seti, her biri eşit olacak biçimde k sayıda alt gruba bölünür. Bu gruplar kat (fold) olarak

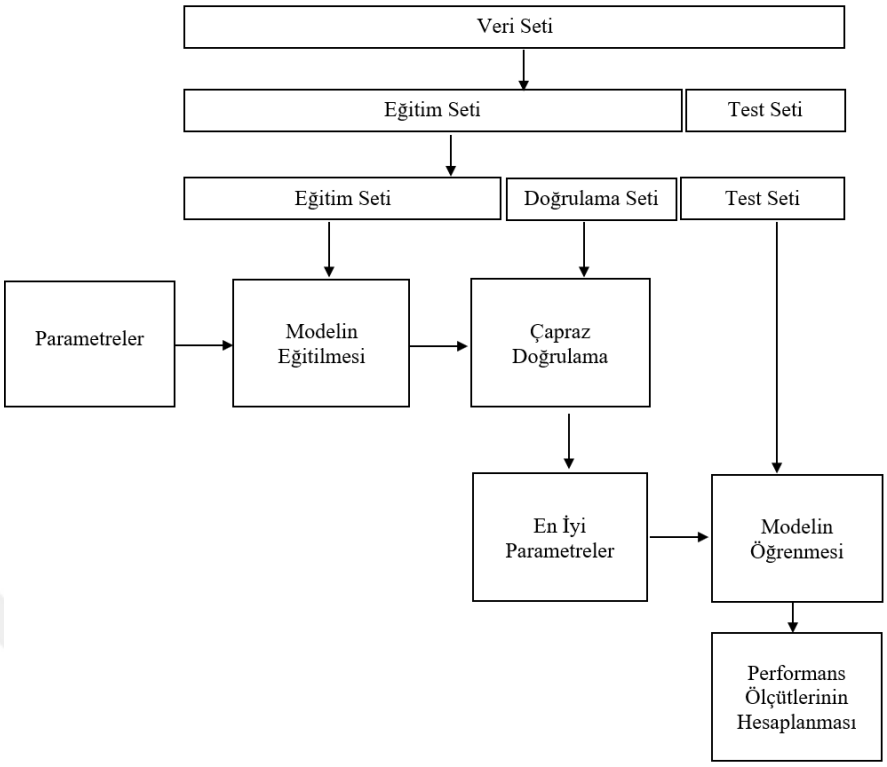
adlandırılır. Bölünen her bir gruptan birisi doğrulama diğer $k-1$ sayıda grup eğitim verisi olarak kullanılır. Her bir grup için doğrulama yapılarak doğrulama hatası hesaplanır. Şekil 4.2'de gösterilen biçimde hesaplanan k tane hatanın ortalamasıyla k-kat çapraz doğrulama (k-fold Cross Validation-CV) gerçekleştirilir.



Şekil 4.2 k-kat çapraz doğrulama örneği

Her bir kat için elde edilen doğrulama hatalarının benzer olması modelin eğitim veri setine duyarlı olmadığını gösterir. Model seçimi aşamasında en iyi parametrelerin belirlenebilmesi için ayarlanabilir parametreler kullanılır. Her bir sınıflandırma yöntemine özgü ayarlanabilir parametreler bulunmaktadır. Araştırmacı tarafından seçilen ayarlanabilir parametreler, en uygun değeri belirleyerek modelin sınıflandırma performansını etkiler. Her bir yöntem için belirlenen ayarlanabilir parametreler kullanılarak çeşitli olası modeller oluşturulur.

CV'nin amacı, mevcut verilerden öğrenilen modelin performansını tahmin ederek modelin genellenebilirliğini ölçmek ve farklı modeller arasından mevcut verilere uyan en iyi modeli bulmaktır (Refaeilzadeh vd. 2009). En iyi modelin elde edilmesinde ızgara arama kullanılır. Izgara arama ile CV sonuçlarına dayanarak ayarlanabilir parametreler ile oluşturulan tüm olası model kombinasyonları denenerek en iyi performans gösteren ayarlanabilir parametre kombinasyonu seçilir. Bu süreç şekil 4.3'te verilmiştir. Bu çalışmada, analizlerin hesaplamaları için Python 3.11.3 programı ve kütüphaneleri kullanılmıştır.



Şekil 4.3 Veri setinin öğrenme ve test süreci

Seçilen modelin başarısı daha önce hiç kullanılmamış olan test veri seti ile ölçülür. Uygulanan sınıflandırma yöntemlerinin değerlendirilebilmesi amacıyla performans ölçütleri çizelge 4.1'de verilen karışıklık matrisi kullanılarak hesaplanır.

Çizelge 4.1 Karışıklık matrisi

		Tahmin Sınıfı	
	Sınıf	***Pozitif***	***Negatif***
Gerçek Sınıf	***Pozitif***	Doğru Pozitif (*DP*)	Yanlış Negatif (*YN*)
	Negatif	Yanlış Pozitif (*YP*)	Doğru Negatif (*DN*)

Karışıklık matrisi, sınıflandırma yöntemlerinde model performansını değerlendirmek için kullanılan bir tablo yapısıdır. Bu matris, modelin tahmin ettiği sınıflar ile gerçek sınıflar arasındaki karşılaştırmayı Doğru Pozitif (DP), Yanlış Pozitif (YP), Doğru Negatif (DN) ve Yanlış Negatif (YN) olarak gruplandırılmış gözlem sayısını kullanarak ifade eder.

Karışıklık matrisinde her sınıfa ait gözlem sayılarının birbirine yakın olduğu durumda veri seti dengeli bir dağılıma sahiptir. Dengeli dağılan veri setinde her sınıfın temsil oranı diğer sınıflarla neredeyse eşittir. Dengeli dağılan bir veri seti ile oluşturulan modelde her sınıfa eşit derecede önem verilerek sınıflandırma algoritmalarının performansı artar, çoğunluk sınıfların baskın olması önlenir. Karışıklık matrisi kullanılarak hesaplanan performans ölçütleri aşağıda verilmiştir.

Doğruluk (Accuracy): Doğru olarak sınıflandırılmış gözlemlerin toplam gözlem sayısına oranı

$$Doğruluk = \frac{DP + DN}{DP + DN + YP + YN} \tag{4.1}$$

biçiminde hesaplanır. Doğruluk, yaygın olarak kullanılan performans ölçütüdür (Erden 2021). Dengeli dağılıma sahip veri setlerinde daha iyi sonuç vermektedir. Bu nedenle dengeli dağılıma sahip olmayan veri setlerinin sınıflandırılmasında yeterli bir ölçüt değildir (Hossin ve Sulaiman 2015).

Duyarlılık (Sensitivity-Recall): Doğru olarak sınıflandırılmış pozitif gözlem sayısının toplam pozitif gözlem sayısına oranı

$$Duyarlılık = \frac{DP}{DP + YN} \tag{4.2}$$

biçimindedir. Modelin doğruları tahmin etme konusundaki performansını gösterir. Doğru pozitif oranı olarak da tanımlanır. Özellikle doğru olması gerekirken yanlış olarak sınıflandırılan negatif grubun önemli olduğu (tıp alanında hastalık teşhisi gibi) çalışmalarda kullanılır.

Özgüllük (Specificity): Doğru negatif oranı olarak da tanımlanan bu ölçüt

$$Özgüllük = \frac{DN}{DN + YP} \tag{4.3}$$

biçiminde hesaplanır. 1-*Özgüllük*, yanlış pozitif oranını vermekte olup, *ROC* eğrisinin yatay eksenini oluşturmaktadır.

Kesinlik (Precision): Pozitif olarak sınıflandırılan gözlemlerin ne oranda doğru olduğu

$$Kesinlik = \frac{DP}{DP + YP} \tag{4.4}$$

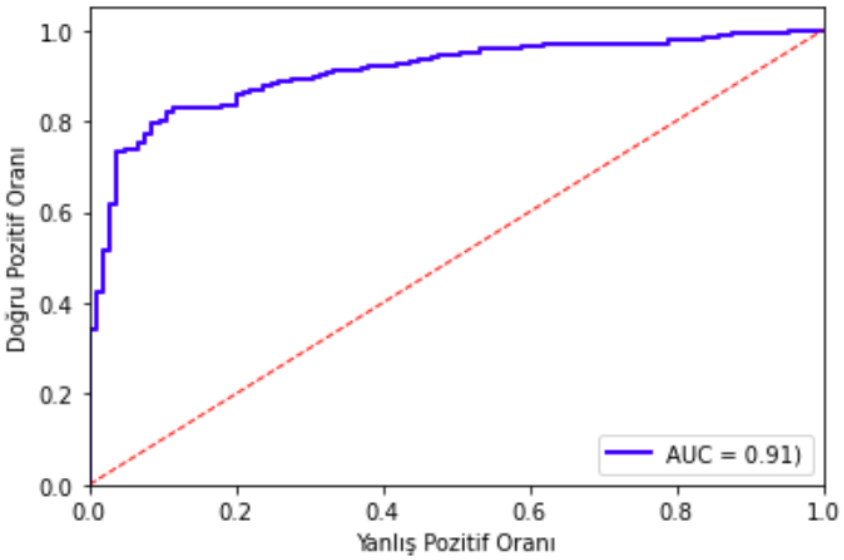
biçiminde hesaplanır. Doğru olarak belirlenen sınıfın performans ölçümünde daha iyi sonuç verir. Pozitif gözlemlerin doğru bir biçimde tanımlamanın önemli olduğu durumlarda kullanılır.

F_1-Skor (F_1-Score): Kesinlik ve duyarlılık ölçütünün harmonik ortalaması olan ölçüt

$$F_1 - Skor = 2\frac{(Duyarlılık)(Kesinlik)}{(Duyarlılık) + (Kesinlik)} \tag{4.5}$$

biçiminde tanımlanır. Hem kesinlik hem duyarlılığı benzer olan sınıfı ön plana çıkarır. Özellikle dengeli dağılıma sahip olmayan veri setlerinin değerlendirilmesinde yüksek performanslı bir değerlendirme ölçütüdür (Uğuz 2021). İkili sınıflandırma problemlerinde doğruluktan daha iyi bir performans göstermektedir (Joshi 2002).

ROC Eğrisi (Receiver operator characteristic curve): ROC eğrisinin yatay ekseni yanlış pozitif oranı (1-*Özgüllük*), dikey ekseni ise doğru pozitif oranı (*Duyarlılık*) ile oluşturulur. Duyarlılık, veri setinde tüm pozitif örneklerin performansını ve özgüllük ise, tüm negatif örneklerin performansını gösterir. Şekil 4.4'te verilen ROC eğrisinin yorumlanmasında eğrinin altında kalan alanı veren AUC ölçütü kullanılır.



Şekil 4.4 ROC eğrisi

Dengesiz dağılıma sahip veri setlerinde, yanlış pozitif oranı yeterli bir ölçüt olmadığından yerine kesinlik ölçütü kullanılarak Precision-Recall Curve (PCR) eğrisi oluşturulur (Fu vd.).

AUC (Area under the curve): AUC, ROC eğrisinin altında kalan tüm yamukların alanı toplanarak hesaplanır (Kumar vd. 2020). Dengesiz veri setleri için AUC ölçütü, PCR eğrisinin altında kalan alan toplamıdır. Bu alan ne kadar büyük ise sınıflandırma performansı o kadar yüksektir. AUC ölçütü, 0 ile 1 arasında değer alır.

4.2 Çok Ölçütlü Karar Verme Yöntemleri

Sınıflandırma performans değerlerinin hesaplanması aşamasından sonra uygun yöntemin belirlenmesi gerekir. Bu aşamada, çok ölçütlü karar verme (Multi Criteria Decision Making-MCDM) yöntemlerinden faydalanılacaktır. MCDM, birden çok kritere sahip farklı alternatifler arasından amaca en uygun olanın belirlenerek sıralanmasıdır. MCDM yöntemlerinin uygulama aşamasında öncelikle bir karar matrisi belirlenir. Bu karar matrisi, satırlarda alternatifler (sınıflandırma yöntemleri) ve sütunlarda kriterler (performans ölçütleri) olacak biçimde oluşturulur. n tane sınıflandırma yöntemi, m tane performans ölçütünden oluşan $n \times m$ boyutlu D karar matrisi $i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m$ olmak üzere

$$D = \left[d_{ij} \right]_{n \times m} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix} \tag{4.6}$$

biçimde elde edilir. Sınıflandırma yöntemlerinin karşılaştırılmasında farklı büyüklükte bulunan kriterlerin normalleştirilmesi gerekir. Çizelge 4.2'de verilen normalleştirme yaklaşımlarına uygun formüller kullanılarak V normalleştirilmiş karar matrisi

$$V = \left[v_{ij} \right]_{n \times m} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nm} \end{bmatrix} \tag{4.7}$$

elde edilir (Çelen 2014, Vafaei 2022). Literatürde farklı normalleştirme yaklaşımlarının uygulandığı çeşitli MCDM yöntemleri mevcuttur (Özçalıcı 2017; Kabak ve Çınar 2020; Demir 2021). Bu matris üzerinden seçilen MCDM yönteminin adımları uygulanarak belirlenen kriterler arasından amaca en uygun sınıflandırma yönteminin elde edilmesi amaçlanır.

MCDM yöntemlerinde kullanılan V normalleştirilmiş karar matrisinin elde edilmesinde uygulanan farklı normalleştirme yaklaşımlarına ait formüller Çizelge 4.2'de verilmiştir.

Çizelge 4.2 Normalleştirme yaklaşımları ve uygulama amaçlarına göre formüller

Yaklaşımlar	***Uygulama Amaçları***	***Formüller***
Toplama dayalı	*Fayda*	$v_{ij} = \frac{d_{ij}}{\sqrt{\sum_{k=1}^{n} d_{kj}}} \left(\frac{d_{ij}}{\sqrt{\sum_{k=1}^{n} d_{kj}^2}} \right)$
	Maliyet	$v_{ij} = \frac{1/d_{ij}}{\sqrt{\sum_{k=1}^{n} 1/d_{kj}}} \left(\frac{1/d_{ij}}{\sqrt{\sum_{k=1}^{n} 1/d_{kj}^2}} \right)$
Orana dayalı	*Fayda*	$v_{ij} = \frac{d_{ij}}{max d_{ij}}$
	Maliyet	$v_{ij} = 1 - \frac{d_{ij}}{max d_{ij}}$
Min-max değerine dayalı	*Fayda*	$v_{ij} = \frac{d_{ij} - min_j(d_{ij})}{max_j(d_{ij}) - min_j(d_{ij})}$
	Maliyet	$v_{ij} = \frac{max_j(d_{ij}) - d_{ij}}{max_j(d_{ij}) - min_j(d_{ij})}$

Çizelge 4.2'de verilen toplama dayalı, orana dayalı ve min-max değerine dayalı normalleştirme yaklaşımları kullanan MCDM yöntemleri aşağıda verilmiştir.

4.2.1 Toplama dayalı normalleştirme ile karar verme yöntemleri

Toplama dayalı normalleştirmede, karar matrisindeki her bir değer, ilgili sütunun toplam değerine bölünerek normalleştirilir. Normalleştirilmiş her bir değer, sütun içerisindeki

oransal payını gösterir. TOPSIS ve COPRAS, toplama dayalı normalleştirmenin uygulandığı MCDM yöntemleridir.

TOPSIS ve COPRAS yöntemlerine ilişkin uygulama adımları sırasıyla verilmiştir.

TOPSIS

TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution), Hwang ve Yoon (1981) tarafından geliştirilmiştir. Bu yöntemde, alternatiflerin (karar seçeneklerinin) değerlendirilmesi pozitif ideal çözüm ve negatif ideal çözüm olmak üzere iki temel noktaya dayanır. TOPSIS yöntemi, pozitif ve negatif ideal çözümlere olan uzaklıkları kullanarak amaca uygun ideal ve ideal olmayan çözüm kümeleri belirler.

Adım1: D karar matrisi eşitlik (4.6) ile belirlenir.

Adım2: V normalleştirilmiş karar matrisi eşitlik (4.7) ile elde edilir.

Adım3: Uygulama amacına uygun bir biçimde pozitif ideal ve negatif ideal çözümler elde edilir. Fayda amacı için, belirlenen kriterleri en büyük yapacak V matrisindeki sütunların en büyük değerleri, pozitif ideal çözüm $V^{+}=\left[v_j^{+}\right]_{1\times m}=\left[v_1^{+} \quad v_2^{+} \quad \cdots \quad v_m^{+}\right]$ ile V matrisindeki sütunların en küçük değerleri, negatif ideal çözüm $V^{-}=\left[v_j^{-}\right]_{1\times m}=\left[v_1^{-} \quad v_2^{-} \quad \cdots \quad v_m^{-}\right]$ elde edilir. Maliyet amacı için ise, V matrisindeki sütunların en küçük değerleri pozitif ideal çözüm, V matrisindeki sütunların en büyük değerleri negatif ideal çözüm değerlerinden oluşur.

Adım4: Öklid yaklaşımı kullanılarak her bir yönteme ilişkin pozitif ideal ve negatif ideal çözüm değerlerine olan uzaklıklar sırasıyla

$$\begin{aligned} S_i^{+} &= \sqrt{\sum_{j=1}^{m}(v_{ij}-v_j^{+})^2},\ i=1,2,\ldots,n \\ S_i^{-} &= \sqrt{\sum_{j=1}^{m}(v_{ij}-v_j^{-})^2},\ i=1,2,\ldots,n \end{aligned} \tag{4.8}$$

biçiminde hesaplanır.

Adım5: Her bir yöntem için ideal çözüme göreli yakınlık katsayısı olan TOPSIS katsayısı

$$C_i=\frac{S_i^{-}}{S_i^{+}+S_i^{-}},\ i=1,2,\ldots,n \tag{4.9}$$

olarak belirlenir. $0 \le C_i \le 1,\ i = 1,2,\ldots,n$ olmak üzere bire yakın TOPSIS katsayısı hesaplanan sınıflandırma yöntemi öncelikli olarak tercih edilir.

COPRAS

COPRAS (Complex Proportional Assesment) yöntemi, Zavadskas ve Kaklauskas (1996) tarafından geliştirilmiştir. Bu yöntemde alternatiflerin karşılaştırılmasında kriterlerin birbirine göre yüzdesel üstünlüğü kullanılır.

Adım1: D karar matrisi eşitlik (4.6) ile belirlenir.

Adım2: V normalleştirilmiş karar matrisi eşitlik (4.7) ile elde edilir.

Adım3: Sınıflandırma yöntemlerine en fazla fayda ve en az fayda sağlayacak ölçütlerin toplamı

$$\begin{aligned} S_i^+ &= \sum_{j=1}^{k} v_{ij},\ i = 1,2,\ldots,k \\ S_i^- &= \sum_{j=k+1}^{m} v_{ij},\ i = k+1,k+2,\ldots,n \end{aligned} \tag{4.10}$$

biçiminde hesaplanır.

Adım4: Sınıflandırma yöntemlerinin göreceli önem değerleri

$$Q_i = S_i^+ + \frac{\sum_{i=1}^{n} S_i^-}{S_i^- \times \sum_{i=1}^{n} \frac{1}{S_i^-}},\ i = 1,2,\ldots,n \tag{4.11}$$

hesaplanır.

Adım5: Sınıflandırma yöntemlerinin performans endeks değerleri

$$P_i = \frac{Q_i}{maks\{Q_i\}} \times 100,\ i = 1,2,\ldots,n \tag{4.12}$$

$0 < P_i \le 100,\ i = 1,2,\ldots,n$ olmak üzere performans endeksi en büyük olan sınıflandırma yöntemi öncelikli olarak tercih edilir.

4.2.2 Orana dayalı normalleştirme ile karar verme yöntemleri

Orana dayalı normalleştirmede, karar matrisindeki her bir değer, ilgili sütundaki en büyük değerine bölünerek normalleştirilir. Bu yöntem, değerlerin ideal bir noktaya göre oransal

benzerliğinin değerlendirilmesinde kullanılır. EDAS ve CODAS, orana dayalı normalleştirmenin uygulandığı MCDM yöntemleridir.

EDAS ve CODAS yöntemlerine ilişkin uygulama adımları sırasıyla verilmiştir.

EDAS

EDAS (Evaluation Based on Distance from Average Solution) yöntemi, Ghorabaee vd. (2015) tarafından geliştirilmiştir. Bu yöntemde, alternatiflerin ortalama çözüme olan pozitif ve negatif uzaklıkları kullanılarak en iyi alternatif belirlenir.

Adım1: D karar matrisi eşitlik (4.6) ile belirlenir.

Adım2: Her bir performans ölçütü için ortalama çözüm değeri

$$AV_j = \frac{1}{n}\sum_{i=1}^{n} d_{ij}, \quad j = 1,2,\ldots,m \tag{4.13}$$

hesaplanır.

Adım3: Ortalama çözümden pozitif ve negatif uzaklıklar fayda amacı için

$$\begin{aligned} PDA_{ij} &= \frac{maks\,(0,(d_{ij} - AV_j))}{AV_j} \\ NDA_{ij} &= \frac{maks\,(0,(AV_j - d_{ij}))}{AV_j} \end{aligned}, i = 1,2,\ldots,n, \quad j = 1,2,\ldots,m \tag{4.14}$$

ve maliyet amacı için

$$\begin{aligned} PDA_{ij} &= \frac{maks\,(0,(AV_j - d_{ij}))}{AV_j} \\ NDA_{ij} &= \frac{maks\,(0,(d_{ij} - AV_j))}{AV_j}, \end{aligned} \quad i = 1,2,\ldots,n, \quad j = 1,2,\ldots,m \tag{4.15}$$

biçiminde hesaplanır.

Adım4: Pozitif ve negatif uzaklıkların toplamı

$$\begin{aligned} SP_i &= \sum_{j=1}^{m} PDA_{ij} \\ SN_i &= \sum_{j=1}^{m} NDA_{ij} \end{aligned}, i = 1,2,\ldots,n, \quad j = 1,2,\ldots,m \tag{4.16}$$

biçimindedir.

Adım5: Pozitif ve negatif uzaklıkların toplamları

$$NSP_i = \frac{SP_i}{maks_i(SP_i)}$$
$$NSN_i = 1 - \frac{SN_i}{maks_i(SN_i)} \quad , i = 1, 2, \ldots, n \tag{4.17}$$

biçiminde normalleştirilir.

Adım6: Değerlendirme skoru

$$AS_i = \frac{NSP_i + NSN_i}{2}, \ i = 1, 2, \ldots, n \tag{4.18}$$

hesaplanır. $0 \le AS_i \le 1, \ i = 1, 2, \ldots, n$ olmak üzere değerlendirme skoru büyük olan sınıflandırma yöntemi öncelikli olarak tercih edilir.

CODAS

CODAS (Combinative Distance Based Assessment) yöntemi, Ghorabaee vd. (2016) tarafından geliştirilmiştir. Bu yöntemde alternatiflerin değerlendirilmesi negatif ideal çözüme olan uzaklıklara dayanır. Bu yöntemin diğer yöntemlerden farkı hem Öklid hem de Taxicab (Manhattan) uzaklıklarının kullanılmasıdır. CODAS yönteminde iki farklı uzaklık ölçüsü kullanıldığından sonuçlar daha duyarlıdır.

Adım1: D karar matrisi eşitlik (4.6) ile belirlenir.

Adım2: V normalleştirilmiş karar matrisi eşitlik (4.7) ile elde edilir.

Adım3: Her kriter için negatif ideal çözüm $V^- = \left[v_j^- \right]_{1 \times m} = \left[v_1^- \quad v_2^- \quad \cdots \quad v_m^- \right]$, V matrisinin sütunlarının en küçük değerleri ile elde edilir.

Adım4: Öklid uzaklığı kullanılarak her bir sınıflandırma yöntemi için

$$E_i = \sqrt{\sum_{j=1}^{m} (v_{ij} - v_j^-)^2}, \ i = 1, 2, \ldots, n \tag{4.19}$$

biçiminde negatif ideal çözüm değerlerine olan uzaklıklar hesaplanır.

Adım5: Taksicab uzaklığı kullanılarak her bir sınıflandırma yöntemi için

$$T_i = \sum_{j=1}^{m} \left| v_{ij} - v_j^- \right|, \ i = 1, 2, \ldots, n \tag{4.20}$$

biçiminde negatif ideal çözüm değerlerine olan uzaklıklar hesaplanır.

Adım6: $R = \left[h_{ik} \right]_{n \times n}$ göreli değerlendirme matrisi

$$h_{ik} = (E_i - E_k) + \psi(E_i - E_k) \times (T_i - T_k), \ i = 1, 2, \ldots, n, \ k = 1, 2, \ldots, n \tag{4.21}$$

elde edilir. $r=(E_i-E_k)$ olarak tanımlansın. ψ ile iki alternatif arasındaki karşılaştırmalı Öklid uzaklık değerleri için bir eşik fonksiyonu

$$\psi(r)=\begin{cases}0, & |r|<\tau\\ 1, & |r|\geq\tau\end{cases} \tag{4.22}$$

oluşturulur. τ araştırmacı tarafından belirlenen eşik değeridir. $0.01<\tau<0.05$ aralığında bir değer olup genellikle $\tau=0.02$ olarak alınır.

Adım7: Değerlendirme puanları

$$H_i=\sum_{k=1}^{n}h_{ik},\ i=1,2,\ldots,n \tag{4.23}$$

biçiminde hesaplanır. Düşük performans gösteren sınıflandırma yöntemi, negatif değerli değerlendirme puanına sahip olabilir (Ghorabaee vd. 2016). $H_i\in\mathbb{R},\ i=1,2,\ldots,n$ olmak üzere değerlendirme puanı büyük olan sınıflandırma yöntemi öncelikli olarak tercih edilir.

4.2.3 Min-max değerine dayalı normalleştirme ile karar verme yöntemleri

Min-max değerine dayalı normalleştirmede, öncelikle eşitlik (4.6)'da verilen D karar matrisindeki her bir değer ile ilgili sütunun minimum değerinin farkı hesaplanır. Bu fark, ilgili sütunun maksimum ve minimum değerleri arasındaki farka bölünerek normalleştirme yapılır. GRA ve MABAC, orana dayalı normalleştirmenin uygulandığı MCDM yöntemleridir.

GRA ve MABAC yöntemlerine ilişkin uygulama adımları sırasıyla verilmiştir.

Gri İlişkisel Analiz

Gri İlişkisel Analiz (Grey Relationship Analysis-GRA), Deng (1982) tarafından geliştirilmiştir. GRA, özellikle örneklem çapı küçük ve yetersiz bilgiye sahip veri seti olan karar verme problemlerinin çözümü için geliştirilmiştir.

Adım1: D karar matrisi eşitlik (4.6) ile belirlenir.

Adım2: Sınıflandırma yöntemlerinin karşılaştırılmasında bir referans değeri

$$d_0=\left(d_0(j)\right),\ j=1,2,\ldots,m \tag{4.24}$$

belirlenir. Bu değer, karar matrisindeki fayda amaçlı ölçüt için sütunda bulunan en büyük değer, maliyet amaçlı ölçüt için sütunda bulunan en küçük değer referans değeridir.

Adım3: Karar matrisi çizelge 4.2'ye göre uygulama amacına uygun durum kullanılarak eşitlik (4.7)'de verilen V normalleştirilmiş karşılaştırma matrisi elde edilir.

Adım4: Her bir sınıflandırma yöntemine ait değerler ile normalleştirilmiş referans değeri arasındaki mutlak fark

$$\Delta_{0i} = \left| v_{oj} - v_{ij} \right|, \ i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m \tag{4.25}$$

hesaplanır. Mutlak değer matrisi

$$\Delta = \left[\Delta_{ij} \right]_{n \times m} = \begin{bmatrix} \Delta_{11} & \Delta_{12} & \cdots & \Delta_{1m} \\ \Delta_{21} & \Delta_{22} & \cdots & \Delta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{n1} & \Delta_{n2} & \cdots & \Delta_{nm} \end{bmatrix} \tag{4.26}$$

biçiminde oluşturulur.

Adım5: GRA katsayı matrisi için

$$\begin{aligned} \gamma_{ij} &= \frac{\Delta_{min} + \eta \Delta_{maks}}{\Delta_{ij} + \eta \Delta_{maks}} \\ \Delta_{max} &= maks_i maks_j \Delta_{ij} \\ \Delta_{min} &= min_i min_j \Delta_{ij} \end{aligned} \tag{4.27}$$

hesaplanır. Genellikle literatürde $\eta = 0.5$ alınır (Özbek 2017; Yıldırım ve Önder 2018). GRA katsayı matrisi

$$\gamma = \left[\gamma_{ij} \right]_{n \times m} \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1m} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nm} \end{bmatrix} \tag{4.28}$$

oluşturulur.

Adım6: Her bir sınıflandırma yöntemi için GRA dereceleri

$$\Gamma_i = \frac{1}{m} \sum_{j=1}^{m} \gamma_{ij}, \ i = 1, 2, \ldots, n \tag{4.29}$$

hesaplanır. $0 \le \Gamma_i \le 1, \ i = 1, 2, \ldots, n$ olmak üzere GRA derecesi büyük olan sınıflandırma yöntemi öncelikli olarak tercih edilir.

MABAC

MABAC (Multi-Attributive Border Approximation Area Comparison), Pamučar ve Cirovic (2015) tarafından geliştirilmiştir. Bu yöntem ile her bir alternatif için sınır yakınlık alanına olan uzaklıklar belirlenerek alternatiflerin sıralaması yapılır.

Adım1: D karar matrisi eşitlik (4.6) ile belirlenir.

Adım2: Karar matrisi çizelge 4.2'ye göre uygulama amacına uygun durum kullanılarak normalleştirilmiş d_{ij}' değerleri elde edilir. d_{ij}' değerleri kullanılarak

$$v_{ij} = d_{ij}' + 1 \tag{4.30}$$

hesaplanır. Eşitlik (4.7)'de verilen V normalleştirilmiş karar matrisi elde edilir.

Adım3: Her bir performans ölçütü için sınır yakınlık alanları

$$g_j = (\prod_{i=1}^{n} v_{ij})^{(1/n)} \tag{4.31}$$

biçiminde hesaplanır. Performans ölçütlerinin sınır yakınlık alanlarının birleşimiyle

$$G = \left[g_1 \;\; g_2 \;\; \cdots \;\; g_m \right] \tag{4.32}$$

G vektörü oluşturulur.

Adım4: Sınıflandırma yöntemlerinin sınır yakınlık alanına olan uzaklıkları

$$Q = \left[q_{ij} \right]_{n \times m} = V - G = \begin{bmatrix} v_{11} - g_1 & v_{12} - g_2 & \cdots & v_{1m} - g_m \\ v_{21} - g_1 & v_{22} - g_2 & \cdots & v_{2m} - g_m \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} - g_1 & v_{n2} - g_2 & \cdots & v_{nm} - g_m \end{bmatrix} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ q_{21} & q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n1} & q_{n2} & \cdots & q_{nm} \end{bmatrix} \tag{4.33}$$

hesaplanır.

Adım5: Her bir sınıflandırma yöntemi için skor değerleri

$$S_i = \sum_{j=1}^{m} q_{ij}, \;\; i = 1, 2, \ldots, n, \;\; j = 1, 2, \ldots, m \tag{4.34}$$

elde edilir. Sınıflandırma yöntemleri performans ölçütleri bakımından daha düşük performans gösterdiğinde negatif değerler elde edilebilir. Pozitif değerler, üstün performansı, negatif değerler ise düşük performansı temsil eder (Torkayesh vd. 2023). $S_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$ olmak üzere skor değeri büyük olan sınıflandırma yöntemi öncelikli olarak tercih edilir.

5. UYGULAMA

Sanayi ve Teknoloji Bakanlığı'nın Ar-Ge Teşvikleri Genel Müdürlüğü bünyesindeki Ar-Ge ve Tasarım Merkezleri Dairesi Başkanlığı tarafından, Ar-Ge veya Tasarım merkezi kurarak yenilik ve tasarım faaliyetlerini yürütmek isteyen firmalara gerekli şartları sağlamaları halinde Ar-Ge veya Tasarım merkezi belgesi verilmektedir. Bu belgeyi almaya hak kazanan firmalar, çeşitli destek ve teşviklerden yararlanmaktadır. Bu kapsamda Türkiye'de bulunan çok sayıda Ar-Ge ve Tasarım merkezi teşviklerden yararlanarak faaliyetlerine devam etmektedir (Anonim 2016).

Ar-Ge Teşvikleri Genel Müdürlüğü tarafından Ar-Ge veya Tasarım merkezi belgesine sahip firmaların belirli dönemlerde faaliyet denetimi yapılarak destek ve teşviklerden yararlanmaya devam edip etmemesi yönünde karar verilmektedir. Gerçekleştirilen yenilik faaliyetlerinin başarıya ulaşması ve sürdürülebilir bir stratejik planlamanın yapılması bakımından faaliyetlerine devam edecek firmaların sınıflandırılması oldukça önemlidir. Bu nedenle, genel yöntemlerin yanı sıra bilimsel ölçütlere göre yapılan değerlendirme çalışmaları hem kamu kaynağının etkin kullanılması hem de hedeflenen başarının sağlanması bakımından gereklidir.

Çalışmanın bu bölümünde, Ar-Ge ve Tasarım merkezi belgesine sahip firmaların faaliyet çalışmalarına göre sınıflandırılması istenmiştir. Yapılan sınıflandırma çalışmasında gerçek veri seti ve simülasyon çalışmasıyla uygulama gerçekleştirilmiştir. Elde edilen sınıflandırma sonuçlarının performansları, MCDM yöntemleri ile değerlendirilmiştir. Çalışmada yapılan analizler için Python 3.11.3 programı ve kütüphaneleri kullanılmıştır.

5.1 Gerçek Veri Seti

Çalışmada, Ar-Ge ve Tasarım merkezi belgesi alan firmaların çalışmalarına göre faaliyetlerinin devam ettirilmesi ya da iptal edilmesi (belge iptali) durumlarına (faaliyet sonucuna) göre sınıflandırılması istenmiştir. Buna göre, yetkililerin bilgisi dahilinde Ar-Ge ve Tasarım Merkezleri Dairesi Başkanlığı'ndan veri seti talep edilmiştir. Veri seti,

2008-2021 yılları arasında faaliyeti devam eden 2334 adet Ar-Ge ve Tasarım merkezlerini kapsamakta olup uygun biçimde veri tabanından temin edilmiştir.

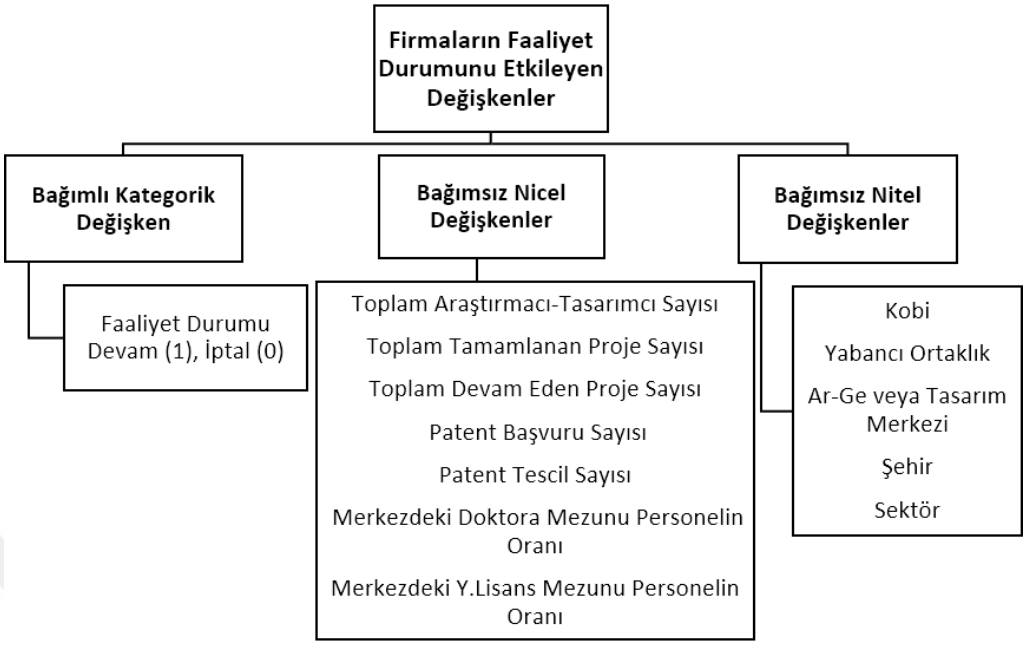
- ***Veri Temizliği***

Ar-Ge ve Tasarım Merkezleri Dairesi Başkanlığı'ndan elde edilen veri setinde her bir bilginin değerli olması ve bilgi kaybının istenmemesi nedeniyle örnekleme yapılmadan veri setinin tamamı değerlendirilmiştir. Ar-Ge ve Tasarım merkezlerine ait ham veri seti çizelge 5.1'de verilmiştir.

Çizelge 5.1 Ham veri seti

No.	*Faaliyet Durum*	*Toplam Y.Lisans Mezunu Personel Sayısı*	*Toplam Doktora Mezunu Personel Sayısı*	*Ar-Ge veya Tasarım Merkezi*	*Kobi*	*Araştırmacı -Tasarımcı Personel Sayısı*	…	*Toplam Tamamlanan Proje Sayısı*
1	Faaliyet Devam	62	1	Ar-Ge	Kobi Değil	272	…	62
2	İptal	7	0	Ar-Ge	Kobi	17	…	8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3	Faaliyet Devam	15	2	Tasarım	Kobi Değil	46	…	42

Çalışma kapsamında, ilgilenilen ham veri setine, öncelikli olarak veri ön işleme uygulanmıştır. Ham veri setinde sistemsel kaynaklı sorun nedeniyle bulunan tekrarlı ve hatalı gözlemler silinmiştir. Aynı bilgiyi içeren değişkenler birleştirilip bazı değişkenleri daha anlamlı hale getirmek amacıyla nicel değerli değişkenler oransal değişkenlere (Merkezdeki doktora mezunu personelin oranı ve Merkezdeki yüksek lisans mezunu personelin oranı) dönüştürülmüştür. Metinsel biçimde bulunan değişkenler (Faaliyet durumu, Kobi, Yabancı Ortaklık ve Ar-Ge veya Tasarım Merkezi) kategorik hale getirilerek düzenlenmiştir. Birbirinden farklı ölçüm birimine sahip değişkenler standartlaştırılmıştır. Firmaların faaliyet durumunu etkilediği düşünülen değişkenler şekil 5.1'de özet olarak sunulmuştur.



Şekil 5.1 Firmaların faaliyet durumunu etkileyen değişkenler

Şekil 5.1'e göre veri seti bir kategorik bağımlı değişken, yedi nicel ve beş nitel değişken olmak üzere on iki (12) bağımsız değişkenden oluşmaktadır. Bu değişkenler,

Y : Faaliyet Durumu

X_1 : Toplam Araştırmacı-Tasarımcı Sayısı

X_2 : Toplam Tamamlanan Proje Sayısı

X_3 : Toplam Devam Eden Proje Sayısı

X_4 : Patent Başvuru Sayısı

X_5 : Patent Tescil Sayısı

X_6 : Merkezdeki Doktora Mezunu Personelin Oranı

X_7 : Merkezdeki Y.Lisans Mezunu Personelin Oranı

X_8 : Kobi

X_9 : Yabancı Ortaklık

X_{10} : Ar-Ge veya Tasarım Merkezi

X_{11} : Şehir

X_{12} : Sektör'dür.

Veri setinde, veri temizliği yapıldıktan sonra 2334 merkez verisi kullanılmıştır. Çizelge 5.2'de, Ar-Ge ve Tasarım merkezlerine ait temizlenmiş veri seti verilmiştir.

Çizelge 5.2 Temizlenmiş veri seti

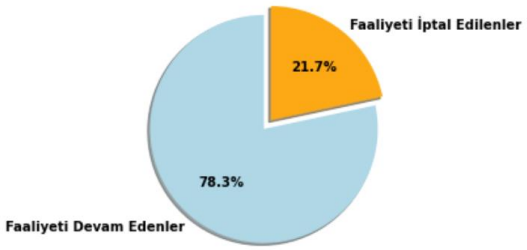
No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	Y
1	15	4	4	0	0	0.00	0.10	0	0	0	Ankara	Mimarlık	0
2	16	3	7	1	0	0.00	0.00	0	0	1	Kocaeli	İmalat	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2334	12	0	13	0	0	0.00	0.18	1	1	1	İstanbul	İlaç	0

- ***Betimsel istatistikler ve veri görselleştirme***

Betimsel istatistikler kullanılarak veri seti hakkında özet bilgiler şekil 5.2-5.4'te gösterilmiştir.

Bağımlı kategorik değişken

Y bağımlı değişkenine ait pasta grafiği şekil 5.2'de verilmiştir.



Şekil 5.2 Veri setinde kategorik bağımlı değişkene ait pasta grafiği

Şekil 5.2'ye göre, Ar-Ge ve Tasarım merkezlerinin yaklaşık %78'inin faaliyetinin devam etmekte olup yaklaşık %22'sinin ise faaliyeti iptal (belge iptali) edilmiştir. Buna göre, bağımlı değişkenin $Y=1$ olma olasılığı $p=0.78$ olarak da ifade edilir.

Bağımsız nicel değişkenler

$\{X_1, X_2, X_3, X_4, X_5, X_6, X_7\}$ nicel değişkenlerine ilişkin verilerin minimum, maksimum ve 1.çeyreklik (Q_1), 2.çeyreklik (Q_2) ve 3.çeyreklik değerleri (Q_3) çizelge 5.3'te verilmiştir.

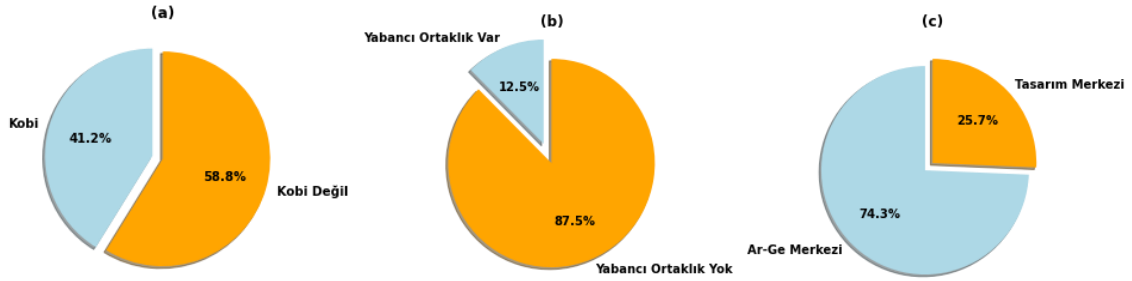
Çizelge 5.3 Veri setindeki nicel değişkenlere ilişkin betimsel istatistikler

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
min	1	0	0	0	0	0	0
Q_1	11	5	4	0	0	0	0.030
Q_2	15	12	6	0	0	0	0.080
Q_3	24	26	11	2	1	0	0.140
max	2710	1056	485	2127	693	0.43	0.710

Çizelge 5.3'ten, Ar-Ge veya Tasarım merkezi olan firmaların yarısında 15'ten fazla Araştırmacı-Tasarımcı personelin bulunduğu, en fazla Araştırmacı-Tasarımcı personeli bulunan firmada ise bu sayının 2710 olduğu, toplam tamamlanan ve toplam devam eden proje sayısı sıfır olan firmaların bulunduğu fakat bu firmaların faaliyette olmadığı, en fazla 2127 patent başvurusu yapıldığı, en fazla 693 patente tescil alındığı, en fazla doktora mezunu personelin oranı 0.43, en fazla y.lisans mezunu personelin oranı 0.71 olduğu görülmektedir.

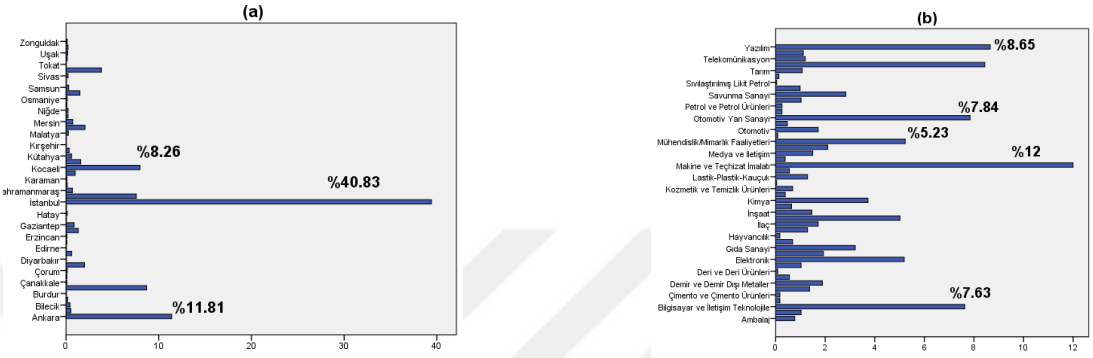
Bağımsız nitel değişkenler

$\{X_8, X_9, X_{10}, X_{11}, X_{12}\}$ nitel değişkenlerine ait grafikler şekil 5.3 ve şekil 5.4'te gösterilmiştir.



Şekil 5.3 (a) Kobi, (b) Yabancı Ortaklık ve (c) Ar-Ge veya Tasarım Merkezi bilgisine ait pasta grafiği

Şekil 5.3'e göre, Ar-Ge ve Tasarım merkezlerinin %58.8'i büyük ölçekli işletme sınıfında yer alırken %41.2'si Kobi'dir. Ar-Ge ve Tasarım merkezlerinin %87.5'inin yabancı ortaklığı bulunmazken %12.5'inin yabancı ortaklığı bulunmaktadır. Firmaların %73.3'ü Ar-Ge merkezi iken %25.7'si Tasarım merkezidir.



Şekil 5.4 Ar-Ge ve Tasarım merkezlerinin (a) Şehir ve (b) Sektör değişkenlerine göre yüzdelik değerleri

Şekil 5.4'e göre, Ar-Ge veya Tasarım merkezi kuran firmaların bulunduğu şehirlerin %40.83'ü İstanbul'da, %11.81'i Ankara'da, %8.26'sı Kocaeli'nde bulunmaktadır. Ar-Ge ve Tasarım merkezlerinin %12'si makine ve teçhizat imalatı sektöründe bulunurken %8.65'i yazılım, %7.84'ü ise otomotiv yan sanayi, %7.63'ü bilgisayar ve iletişim teknolojileri ve %5.23'ü mühendislik-mimarlık sektöründe faaliyet göstermektedir.

- ***Çok değişkenli istatistiksel yöntemler ile keşfedici veri analizi***

Çalışmada kullanılan veri setindeki değişkenlerin çok sayıda olması ve bu değişkenler arasında ilişki olduğunun düşünülmesi sebebiyle faktör analizi uygulanmıştır. Faktör analizi ile benzer özellikteki değişkenlerin gruplandırılması amaçlanmıştır. On iki (12) bağımsız değişkene sahip veri setinden niceliksel değer almayan şehir (X_{11}) ve sektör (X_{12}) değişkenleri hariç tutularak veri setine faktör analizi uygulanmıştır. Değişkenlerin ortak varyans tablosu çizelge 5.4'te verilmiştir.

Çizelge 5.4 On (10) değişkenin ortak varyans tablosu

Değişkenler	***Ortak varyans değerleri***
X_1	0.471
X_2	0.500
X_3	0.524
X_4	0.788
X_5	0.734
X_6	0.526
X_7	0.453
X_8	0.438
X_9	0.679
X_{10}	**0.263**

Çizelge 5.4'e göre Ar-Ge veya Tasarım Merkezi (X_{10}) değişkeni dışında tüm değişkenlerin ortak varyans değerleri 0.3'ten yüksektir. X_{10} değişkeni analizden çıkarılarak test tekrarlanır. Buna göre, X_{10} değişkeni analizden çıkarıldıktan sonra ortak varyans tablosu çizelge 5.5'te verilmiştir.

Çizelge 5.5 Dokuz (9) değişkenin ortak varyans tablosu

Değişkenler	***Ortak varyans değerleri***
X_1	0.472
X_2	0.514
X_3	0.553
X_4	0.804
X_5	0.757
X_6	0.541
X_7	0.445
X_8	0.460
X_9	0.705

Çizelge 5.5'e göre tüm değişkenlerin ortak varyans değerleri 0.3'ten yüksektir. Faktör analizinde değişkenler arasında yüksek korelasyon olması beklenir. Değişkenler

arasındaki korelasyon miktarı arttıkça faktör analizinin güvenilirliği artar. Verilere faktör analizi yapılmasının uygun olup olmadığını belirlemek için

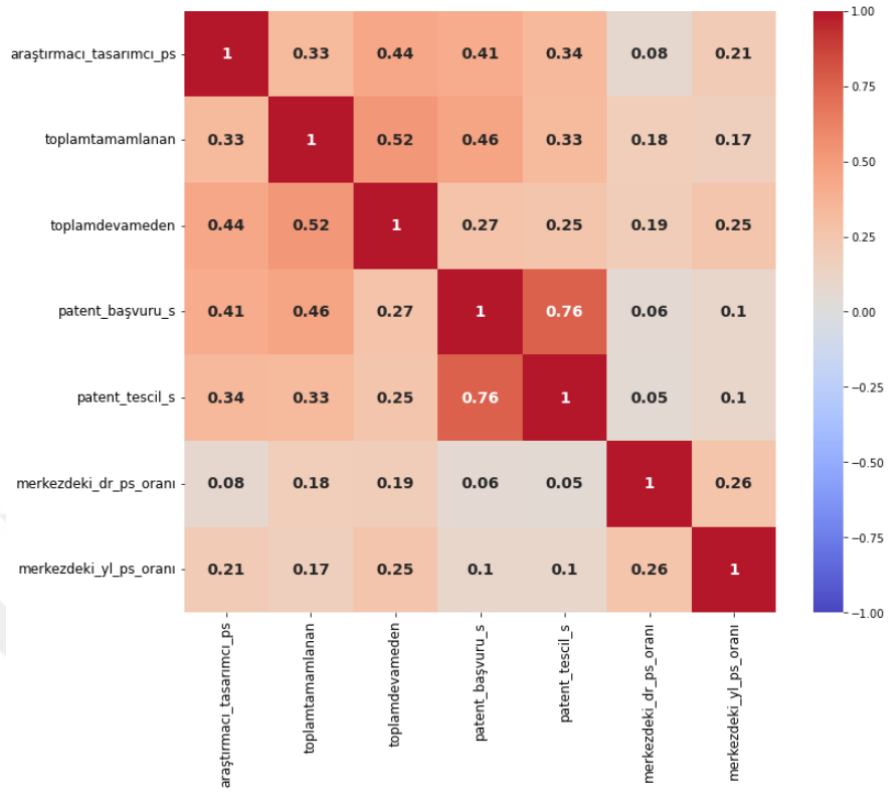
H_0: Değişkenler arasında ilişki yoktur (Korelasyon matrisi birim matristir).
H_S: Değişkenler arasında ilişki bulunmaktadır.

hipotezi test edilir. Çizelge 5.6'da verilen örneklem ve değişken yeterliliğine ait KMO ve Bartlett test sonuçlarına göre değişkenler arasında bir ilişkinin olduğu görülmektedir. Buna göre, örneklem ve değişkenler, faktör analizi için yeterlidir, sonucuna ulaşılır.

Çizelge 5.6 KMO örneklem ve Bartlett değişken yeterlilik testleri

Test İstatistiği	***Değeri***
KMO örneklem yeterliliği	0.698
Bartlett değişken yeterliliği testi	4639.268
sd	36
p-değeri	0.000

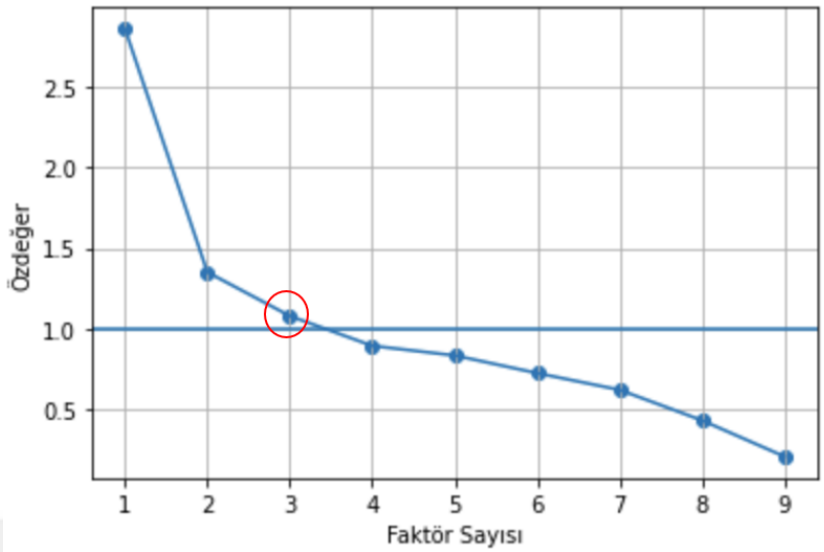
KMO değeri, faktör analizinde değişkenlerin iç tutarlılığını ölçmek için kullanılır. Çizelge 5.6'daki KMO değerinin 0.698 olması değişkenler arasında iyi düzeyde bir ilişki olduğunu göstermektedir. Şekil 5.5'te değişkenler arası korelasyonu gösteren ısı grafiği verilmiştir.



Şekil 5.5 Bağımsız değişkenlerin ısı grafiği

Şekil 5.5'e göre değişkenler arasındaki en yüksek ilişki (%76), patent başvuru değişkeni (X_4) ile patent tescil değişkeni (X_5) arasındadır. Başvurusu yapılan patentlerin daha sonra tescil edilmesi aralarındaki yüksek ilişkiyi desteklemektedir.

Çoklu bağlantının varlığının araştırılması için korelasyon matrisinin determinantı hesaplanır. Determinantın, 0.00001'den büyük olması, çoklu bağlantının olmadığını gösterir. *Determinant* = 0.143>0.00001 olduğundan çoklu bağlantı sorunu yoktur. Faktör analizi sonucu oluşan faktör sayısı şekil 5.6'da verilmiştir.



Şekil 5.6 Yamaç eğrisi

Şekil 5.6'daki yamaç eğrisinin yatay ekseni faktör sayısını ve düşey ekseni özdeğer sayısını gösterir. Buna göre, yamaç eğrisi, özdeğeri bir ve birden fazla olan üç (3) faktörü göstermektedir. Faktör analizi sonucunda dokuz (9) değişkenin üç faktör ile açıklandığı görülmektedir. Faktör analizi sonrası oluşan faktörler ve değişkenlere ait faktör yükleri çizelge 5.7'de verilmiştir.

Çizelge 5.7 Faktör analizi sonuçları

Değişkenler	F_1	F_2	F_3
Araştırmacı-Tasarımcı Personel Sayısı	0.605		
Toplam Tamamlanan Proje Sayısı	0.515	0.481	
Toplam Devam Eden Proje Sayısı	0.409	0.612	
Patent Başvuru Sayısı	**0.897**		
Patent Tescil Sayısı	**0.869**		
Merkezdeki Doktora Mezunu Personelin Oranı		**0.720**	
Merkezdeki Y. Lisans Mezunu Personelin Oranı		0.623	
Kobi			-0.632
Yabancı Ortaklık			**0.835**
Varyans Açıklama Oranı	0.263	0.184	0.135
Toplam Varyans Açıklama Oranı		0.583	

Faktör yükleri, faktörlerin değişkenlerin varyanslarına sağladığı katkıyı gösterir. Çizelge 5.7'e göre ilk faktör toplam varyansın, %26.3'nü, ikinci faktör %18.4'ünü, üçüncü faktör %13.5'ini açıklamaktadır. Tüm faktörler, değişkenlerdeki varyansın %58.3'ünü açıklamaktadır.

Faktör analizi sonucunda elde edilen bilgilerle Ar-Ge ve Tasarım merkezlerinin faaliyet durumunu etkileyen değişkenler gruplanarak belirlenmiştir. Elde edilen sonuçlar çizelge 5.8'de açıkça görülmektedir.

Çizelge 5.8 Faktör analizi sonucu değişkenler ve grupları

Gruplar	***Değişkenler***
Proje Süreç Bilgisi	Araştırmacı-Tasarımcı Personel Sayısı
	Toplam Tamamlanan Proje Sayısı
	Toplam Devam Eden Proje Sayısı
	Patent Başvuru Sayısı
	Patent Tescil Sayısı
Lisansüstü Eğitim Bilgisi	Merkezdeki Doktora Mezunu Personelin Oranı
	Merkezdeki Y. Lisans Mezunu Personelin Oranı
Merkez Bilgisi	Kobi
	Yabancı Ortaklık
	Ar-Ge veya Tasarım Merkezi
	Şehir
	Sektör

- ***Çok ölçütlü karar verme yöntemleri ile değişken ağırlıklarını belirleme***

Ar-Ge ve Tasarım Merkezleri Dairesi'nde çalışan uzmanların görüşleri dikkate alınarak hibrit bir sınıflandırma yaklaşımı uygulanması istenmiştir. Bu amaçla, nitel ve nicel değişkenleri değerlendirebilen hem objektif hem subjektif bilgileri karar sürecine dahil edebilen AHP yöntemi seçilmiştir. Ar-Ge ve Tasarım merkezlerinin faaliyetlerini etkileyen değişkenlerin önemlerine göre ağırlıklarının hesaplanması istenmiştir. AHP yönteminde her bir değişkenin birbirine göre doğrudan öneminin belirlenebilmesi amacıyla on beş (15) uzmanın görüşü alınmıştır. Uzman görüşleri 12*x*12 boyutlu

$U_1, U_2, \cdots, U_{15}$ matrisleri biçiminde EK 1'de verilmiştir. EK 1'de verilen matrislerin geometrik ortalaması alınarak eşitlik (2.4)'te verilen karşılaştırma matrisi ile

$$A = \begin{bmatrix} 1.00 & 2.00 & 1.00 & 2.00 & 1.00 & 0.50 & 1.00 & 9.00 & 3.00 & 4.00 & 5.00 & 4.00 \\ 0.50 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 6.00 & 3.00 & 2.00 & 4.00 & 3.00 \\ 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 5.00 & 3.00 & 2.00 & 4.00 & 2.00 \\ 0.50 & 1.00 & 1.00 & 1.00 & 0.50 & 0.33 & 0.33 & 6.00 & 3.00 & 2.00 & 4.00 & 2.00 \\ 1.00 & 1.00 & 1.00 & 2.00 & 1.00 & 1.00 & 1.00 & 7.00 & 3.00 & 3.00 & 6.00 & 3.00 \\ 2.00 & 1.00 & 1.00 & 3.00 & 1.00 & 1.00 & 5.00 & 9.00 & 5.00 & 4.00 & 6.00 & 5.00 \\ 1.00 & 1.00 & 1.00 & 3.00 & 1.00 & 0.20 & 1.00 & 8.00 & 4.00 & 3.00 & 5.00 & 3.00 \\ 0.11 & 0.16 & 0.20 & 0.16 & 0.14 & 0.11 & 0.12 & 1.00 & 0.33 & 0.50 & 1.00 & 1.00 \\ 0.33 & 0.33 & 0.33 & 0.33 & 0.33 & 0.20 & 0.50 & 3.00 & 1.00 & 1.00 & 2.00 & 1.00 \\ 0.25 & 0.50 & 0.50 & 0.50 & 0.33 & 0.25 & 0.66 & 2.00 & 1.00 & 1.00 & 1.00 & 1.00 \\ 0.20 & 0.25 & 0.25 & 0.25 & 0.16 & 0.16 & 0.20 & 1.00 & 0.50 & 1.00 & 1.00 & 1.00 \\ 0.25 & 0.33 & 0.50 & 0.50 & 0.33 & 0.40 & 0.33 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \end{bmatrix} \quad (5.1)$$

biçiminde oluşturulmuştur. Uzmanların belirttiği farklı görüşlerin birleştirilmesinde karşılaştırma matrisinin değerlerinin uç değerlerden etkilenmemesi için geometrik ortalama kullanılmıştır. Eşitlik (2.6) kullanılarak hesaplanan AHP değişken ağırlıkları çizelge 5.9'da verilmiştir.

Çizelge 5.9 AHP ile belirlenen değişken ağırlıkları

	Değişkenler	***w***
Nicel	Araştırmacı-Tasarımcı Sayısı (X_1)	0.1310
	Toplam Tamamlanan Proje Sayısı (X_2)	0.1020
	Toplam Devam eden Proje Sayısı (X_3)	0.1026
	Patent Başvuru Sayısı (X_4)	0.0800
	Patent Tescil Sayısı (X_5)	0.1218
	Merkezdeki Doktora Mezunu Personelin Oranı (X_6)	**0.1835**
	Merkezdeki Y. Lisans Mezunu Personelin Oranı (X_7)	0.1189
Nitel	Kobi (X_8)	0.0186
	Yabancı Ortaklık (X_9)	0.0390
	Ar-Ge veya Tasarım Merkezi (X_{10})	0.0404
	Şehir (X_{11})	0.0249
	Sektör (X_{12})	0.0373

Çizelge 5.9'a göre Ar-Ge ve Tasarım merkezlerinin faaliyet durumunu etkileyen en önemli değişkenlerin sırasıyla Merkezdeki Doktora Mezunu Personelin Oranı $(w_6 = 0.1835)$, Araştırmacı-Tasarımcı Sayısı $(w_1 = 0.1310)$ ve Patent Tescil Sayısı $(w_5 = 0.1218)$ olduğu uzmanlar tarafından düşünülmektedir. Ar-Ge veya Tasarım Merkezi olması $(w_{10} = 0.0404)$, Yabancı Ortaklık $(w_9 = 0.0390)$, Sektör $(w_{12} = 0.0373)$, Şehir $(w_{11} = 0.0249)$ ve Kobi $(w_8 = 0.0186)$ bilgilerinin Ar-Ge ve Tasarım merkezlerinin faaliyet durumunu etkileyen en az öneme sahip değişkenler olduğu görülmektedir. AHP sonucunda en az öneme sahip olduğu düşünülen merkezlere ait bilgileri içeren nitel değişkenlerin (Kobi, Yabancı Ortaklık Ar-Ge veya Tasarım Merkezi, Şehir, Sektör) toplam ağırlıkları 0.16'dır. Bu nedenle, bu nitel değişkenler veri madenciliği sınıflandırma yöntemlerinde dikkate alınmamıştır.

Veri ön işleme sonucunda veri seti, yedi (7) bağımsız nicel değişken ve bir (1) kategorik bağımlı değişken ile analize hazır hale getirilmiştir. Çizelge 5.10'da veri ön işleme sonucunda elde edilen veri seti verilmiştir.

Çizelge 5.10 İşlenmiş veri seti

No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y
1	15	4	4	0	0	0.00001	0.10	0
2	16	3	7	1	0	0.000001	0.00	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2334	12	0	13	0	0	0.00001	0.18	0

Eşitlik (2.2) ile standartlaştırılmış veri seti çizelge 5.11'de verilmiştir.

Çizelge 5.11 Standartlaştırılmış veri seti

No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y
1	-0.1616	-0.4583	-0.3806	-0.1242	-0.1399	-0.3562	0.0465	0
2	-0.1497	-0.4791	-0.1867	-0.1093	-0.1399	-0.3562	-1.1051	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2334	-0.1974	-0.5415	0.2010	-0.1242	-0.139	-0.3562	0.9678	0

Çizelge 5.9'da verilen AHP ile belirlenen değişken ağırlıkları kullanılarak veri setindeki değişkenler çizelge 2.3'teki biçimde ağırlıklandırılmıştır. AHP ile ağırlıklandırılmış veri seti çizelge 5.12'de verilmiştir.

Çizelge 5.12 AHP ile ağırlıklandırılmış veri seti

No.	ζ_1	ζ_2	ζ_3	ζ_4	ζ_5	ζ_6	ζ_7	Y
1	0.0211	-0.0467	-0.0390	-0.0099	-0.0170	-0.0653	0.0055	0
2	-0.0196	-0.0488	-0.0191	-0.0087	-0.0170	-0.0653	-0.1313	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2334	-0.0258	-0.0552	0.0206	-0.0099	-0.0170	-0.0653	0.1150	0

Çizelge 5.12'de veri seti, uzman görüşünün sınıflandırma yöntemlerine dahil edilmesi amacıyla çizelge 5.9 kullanılarak ağırlıklandırılmıştır.

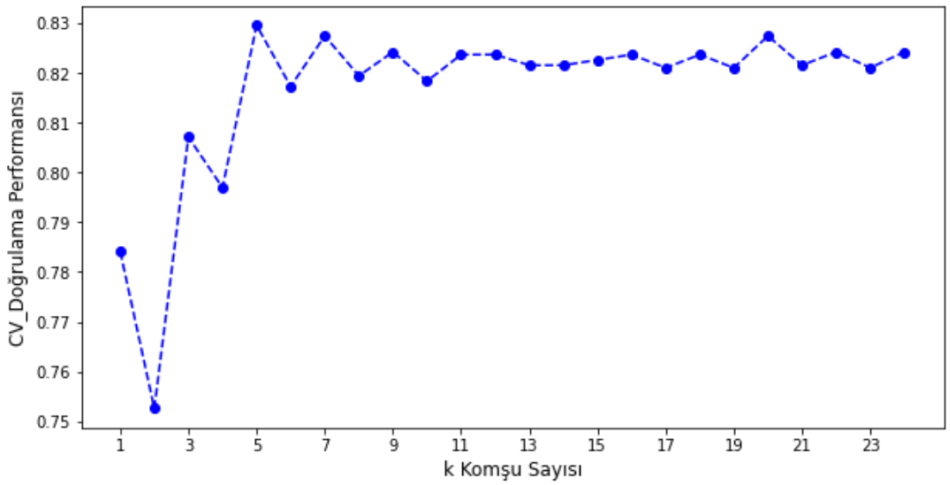
- ***Veri madenciliği yöntemleri ile sınıflandırma***

Çalışmanın bir sınıflandırma problemi olması sebebiyle veri madenciliği sınıflandırma yöntemlerinden LR, *k*NN, SVM ve RF kullanılmıştır. Sınıflandırma yöntemlerinin performansının ölçülebilmesi amacıyla veri seti, %80 eğitim seti ile %20 test seti olmak üzere ayrılmıştır. Python 3.11.3 sürümünde Scikit-learn kütüphanesinde sınıflandırma yöntemlerinin performansı 5-kat CV kullanılarak test edilmiş, ızgara arama yöntemi ile optimal parametreler elde edilmiştir. Araştırmacı tarafından belirlenen ayarlanabilir parametre değerleri çizelge 5.13'te verilmiştir.

Çizelge 5.13 Gerçek veri seti için sınıflandırma yöntemleri ayarlanabilir parametre değerleri

	Yöntemler	***Ayarlanabilir Parametreleri***
Araştırmacı ile belirlenen	**LR**	$c = \{1e-05,\ 3,\ 0.1\}$, $Penalty = \{l_1, l_2,\ None\}$
	***k*NN**	$Metric = \{Minkowski,\ Euclidean\}$, $Neighbors(k) = \{3,4,\ldots,25\}$
	SVM	$c = \{0,\ 0.1,\ 1,\ 10,\ 100\}$, $Gamma = \{0.001,\ 0.01,\ 1,\ 10\}$, $Kernel = \{Polynomial,\ Rbf,\ Sigmoid,\ Linear\}$
	RF	$Bootstrap = True$, $Criterion = \{Gini,\ Entropy\}$, $Max\_Depth = \{3,\ 5\}$, $Max\_Features = \{2,\ 3\}$, $Min\_Samples\_Leaf = \{5,\ 8,\ 10\}$, $Min\_Samples\_Split = \{3,\ 4,\ 5,\ 6\}$, $n\_Estimator = \{100,\ 200,\ 250,\ 500\}$

Sınıflandırma yöntemlerinden LR, *k*NN, SVM ve RF için çizelge 5.12'de verilen ayarlanabilir parametre değerleri araştırmacı tarafından belirlenir. Izgara arama yöntemi ile her bir sınıflandırma yöntemi için belirlenen ayarlanabilir parametre değerlerinin tüm olası kombinasyonları oluşturulur. Oluşturulan kombinasyonlar LR için eşitlik (3.4), *k*NN için eşitlik (3.10), SVM için eşitlik (3.22) ve RF için eşitlik (3.33)'da tanımlı optimizasyon modellerinde amaç fonksiyonunun optimal sonucunu veren ayarlanabilir parametre kombinasyonu Python 3.11.3 sürümünde seçilir.



Şekil 5.7 *k* sayısının belirlenmesi

*k*NN yönteminde k ayarlanabilir parametresinin belirlenmesinde CV'de en yüksek doğrulama performansına göre optimal k sayısı hesaplanır. Şekil 5.7'de optimal sayının $k=5$ olduğu görülmektedir.

Izgara arama ile seçilen sınıflandırma yöntemleri ayarlanabilir parametre değerleri çizelge 5.14'te verilmiştir.

Çizelge 5.14 Gerçek veri seti için ızgara uygulaması sonucu elde edilen ayarlanabilir parametre değerleri

	Yöntemler	***Ayarlanabilir Parametreleri***
Izgara arama ile seçilen	**LR**	$c=1e-05,\ Penalty=None$
	***k*NN**	$Metric=Euclidean,\ Neighbors(k)=5$
	SVM	$c=100,\ Gamma=10,\ Kernel=Rbf$
	RF	$Bootstrap=True,\ Criterion=Gini,\ Max\_Depth=3,$ $Max\_Features=3,\ Min\_Samples\_Leaf=5,$ $Min\_Samples\_Split=3,\ n\_Estimator=250$

Gerçek veri setinin %80'i eğitim seti (1868) %20 test seti (466) test seti olmak üzere ayrılır. Her bir sınıflandırma yöntemine göre test seti için öngörü değerleri çizelge 5.15'te verilmiştir.

Çizelge 5.15 Test seti için öngörü değerleri

No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y_{Test}	$\hat{Y}_{LR}$	$\hat{Y}_{kNN}$	$\hat{Y}_{SVM}$	$\hat{Y}_{RF}$
1	23	38	14	3	2	0.00	0.04	1	1	1	1	1
2	20	9	3	0	0	0.00	0.10	1	1	1	1	1
3	16	3	3	0	0	0.00	0.07	0	1	0	0	0
4	21	25	6	1	1	0.00	0.19	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
465	10	3	5	0	0	0.00	0.07	0	0	1	0	0
466	13	10	6	1	0	0.00	0.00	0	1	1	1	0

Çizelge 5.15'e göre Toplam Araştırmacı-Tasarımcı Sayısı (X_1) 13, Toplam Tamamlanan Proje Sayısı (X_2) 10, Toplam Devam Eden Proje Sayısı (X_3) 6, Patent Başvuru Sayısı (X_4) 1, Patent Tescil Sayısı (X_5) 0, Merkezdeki Doktora Mezunu Personelin Oranı (X_6) 0.00 ve Merkezdeki Y.Lisans Mezunu Personelin Oranı (X_7) 0.00 olan 466 numaralı merkezin faaliyeti iptal edilmiştir $(Y_{Test} = 0)$. Bu merkez, LR, *k*NN ve SVM ile sınıflandırıldığında sınıfı $(\hat{Y}_{LR} = \hat{Y}_{kNN} = \hat{Y}_{SVM} = 1)$ (faaliyeti devam ediyor) olarak öngörülmüştür. RF algoritması ile sınıflandırıldığında ise gerçek sınıfı ile benzer olarak faaliyet durumunun iptal $(\hat{Y}_{RF} = 0)$ olduğu öngörüsü elde edilmiştir.

- ***Sınıflandırma performans ölçütleri***

Sınıflandırma yöntemlerini değerlendirmek için performans ölçütlerinden doğruluk, kesinlik, duyarlılık, F_1-Skor ve AUC değerleri eğitim ve test veri setleri için hesaplanmıştır. AHP ile ağırlıklandırılmış veri setine ait sınıflandırma performans ölçüt değerleri çizelge 5.16'da verilmiştir.

Çizelge 5.16 AHP ile ağırlıklandırılmış veri setine ilişkin performans değerleri

	Yöntemler	*Doğruluk*	*Kesinlik*	*Duyarlılık*	*F_1-Skor*	*AUC*
Eğitim veri seti	**LR**	0.8553	0.8965	0.9222	0.9092	0.9077
	***k*NN**	0.8730	0.9099	0.9304	0.9200	0.9232
	SVM	0.8618	0.9148	0.9085	0.9117	0.8977
	RF	0.8580	0.9000	0.9215	0.9106	0.9022
Test veri seti	**LR**	0.8522	0.8906	0.9226	0.9063	0.9077
	***k*NN**	0.8308	0.8713	0.9171	0.8936	0.8540
	SVM	0.8522	0.9013	0.9088	0.9050	0.8936
	RF	0.8650	0.8986	0.9309	0.9145	0.9140

Çizelge 5.16'ya göre, performans değerleri bakımından sınıflandırma yöntemlerini karşılaştırmak zordur. Objektif karşılaştırma yapılabilmesi için karar verme yöntemlerinin kullanılması gerekir.

- ***Çok ölçütlü karar verme yöntemleri***

Çizelge 5.16'da verilen performans değerlerine göre sınıflandırma yöntemi seçiminde MCDM yöntemleri uygulanır. Karar verme yöntemlerinin uygulanabilmesinde oluşturulan karar matrisi

$$D = \begin{bmatrix} 0.8522 & 0.8906 & 0.9226 & 0.9063 & 0.9077 \\ 0.8308 & 0.8713 & 0.9171 & 0.8936 & 0.8540 \\ 0.8522 & 0.9013 & 0.9088 & 0.9050 & 0.8936 \\ 0.8650 & 0.8986 & 0.9309 & 0.9145 & 0.9140 \end{bmatrix} \tag{5.2}$$

biçiminde olur. Bu çalışmada MCDM yöntemlerinden TOPSIS, COPRAS, EDAS, CODAS, GRA ve MABAC kullanılmıştır. Her bir kriter eşit öneme sahip olup performans ölçütleri maksimum olacak biçimde fayda amacına uygun Çizelge 4.2'deki formüller uygulanmıştır. Eşitlik (5.2)'deki karar matrisi her bir MCDM yöntemi için uygun formül kullanılarak normalize edilir.

Adım1: D karar matrisi eşitlik (5.2)'deki biçimde oluşturulmuştur.

Toplama Dayalı normalleştirme yaklaşımı:

TOPSIS

Çizelge 5.17 TOPSIS uygulama adımları

Adım 2	***Adım 3***	***Adım 4***	***Adım 5***
$V=\begin{bmatrix}0.5012 & 0.5000 & 0.5014 & 0.5007 & 0.5084\\ 0.4886 & 0.4892 & 0.4984 & 0.4937 & 0.4783\\ 0.5012 & 0.5060 & 0.4939 & 0.5000 & 0.5005\\ 0.5087 & 0.5045 & 0.5059 & 0.5053 & 0.5119\end{bmatrix}$	$V^{+}=\begin{bmatrix}0.5087 & 0.5060 & 0.5059 & 0.5053 & 0.5119\end{bmatrix}$ $V^{-}=\begin{bmatrix}0.4886 & 0.4892 & 0.4939 & 0.4937 & 0.4783\end{bmatrix}$	$S^{+}=\begin{bmatrix}0.0120\\ 0.0448\\ 0.0189\\ 0.0015\end{bmatrix}$ $S^{-}=\begin{bmatrix}0.0358\\ 0.0045\\ 0.0312\\ 0.0452\end{bmatrix}$	$C=\begin{bmatrix}0.7479\\ 0.0914\\ 0.6221\\ 0.9675\end{bmatrix}$

COPRAS

Çizelge 5.18 COPRAS uygulama adımları

Adım 2	***Adım 3***	***Adım 4***	***Adım 5***
$V=\begin{bmatrix}0.2506 & 0.2500 & 0.2507 & 0.2504 & 0.2543\\ 0.2443 & 0.2446 & 0.2492 & 0.2468 & 0.2392\\ 0.2506 & 0.2530 & 0.2469 & 0.2500 & 0.2503\\ 0.2543 & 0.2522 & 0.2530 & 0.2526 & 0.2560\end{bmatrix}$	$S^{+}=\begin{bmatrix}1.2561\\ 1.2243\\ 1.2510\\ 1.2684\end{bmatrix}$ $S^{-}=\begin{bmatrix}0\\ 0\\ 0\\ 0\end{bmatrix}$	$Q=\begin{bmatrix}1.2561\\ 1.2243\\ 1.2510\\ 1.2684\end{bmatrix}$	$P=\begin{bmatrix}99.03\\ 96.52\\ 98.63\\ 100.0\end{bmatrix}$

Orana dayalı normalleştirme yaklaşımı:

EDAS

Çizelge 5.19 EDAS uygulama adımları

Adım 2	***Adım 3***	***Adım 4***	***Adım 5***	***Adım 6***
$AV=\begin{bmatrix}0.8500 & 0.8904 & 0.9198 & 0.9048 & 0.8923\end{bmatrix}$	$PDA=\begin{bmatrix}0.00252 & 0.0001 & 0.0029 & 0.0016 & 0.0172\\ 0 & 0 & 0 & 0 & 0\\ 0.0025 & 0.0121 & 0 & 0.0001 & 0.0014\\ 0.0175 & 0.00915 & 0.0120 & 0.0106 & 0.0242\end{bmatrix}$ $NDA=\begin{bmatrix}0 & 0 & 0 & 0 & 0\\ 0.0226 & 0.0215 & 0.0029 & 0.0124 & 0.0429\\ 0 & 0 & 0.0120 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\end{bmatrix}$	$SP=\begin{bmatrix}0.0245\\ 0\\ 0.0163\\ 0.0737\end{bmatrix}$ $SN=\begin{bmatrix}0\\ 0.1025\\ 0.0120\\ 0\end{bmatrix}$	$NSP=\begin{bmatrix}0.3326\\ 0\\ 0.2212\\ 1\end{bmatrix}$ $NSN=\begin{bmatrix}1\\ 0\\ 0.8828\\ 1\end{bmatrix}$	$AS=\begin{bmatrix}0.6663\\ 0.0000\\ 0.5520\\ 1.0000\end{bmatrix}$

CODAS

Çizelge 5.20 CODAS uygulama adımları

Adım 2	***Adım 3***	***Adım 4***	***Adım 5***	***Adım 6***	***Adım7***
$V=\begin{bmatrix}0.9852 & 0.9881 & 0.9910 & 0.9910 & 0.9931\\ 0.9604 & 0.9667 & 0.9851 & 0.9771 & 0.9343\\ 0.9852 & 1 & 0.9762 & 0.9896 & 0.9776\\ 1 & 0.9970 & 1 & 1 & 1\end{bmatrix}$	$V^{-}=\begin{bmatrix}0.9604 & 0.9667 & 0.9762 & 0.9771 & 0.9343\end{bmatrix}$	$E=\begin{bmatrix}0.0702\\ 0.0089\\ 0.0612\\ 0.0887\end{bmatrix}$	$T=\begin{bmatrix}0.1336\\ 0.0089\\ 0.1138\\ 0.1820\end{bmatrix}$	$R=\begin{bmatrix}0 & 0.0689 & 0.0089 & -0.0184\\ -0.0613 & 0 & -0.0523 & -0.0798\\ -0.0089 & 0.0578 & 0 & -0.0274\\ 0.0184 & 0.0936 & 0.0293 & 0\end{bmatrix}$	$H=\begin{bmatrix}0.0594\\ -0.1935\\ 0.0213\\ 0.1415\end{bmatrix}$

Min-max değerine dayalı normalleştirme yaklaşımı:

GRA

Çizelge 5.21 GRA uygulama adımları

Adım 2	***Adım 3***	***Adım 4***	***Adım 5***	***Adım 6***
$d_0 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}$	$V = \begin{bmatrix} 0.6257 & 0.6433 & 0.6244 & 0.6076 & 0.8950 \\ 0 & 0 & 0.3755 & 0 & 0 \\ 0.6257 & 1 & 0 & 0.5454 & 0.6600 \\ 1 & 0.9100 & 1 & 1 & 1 \end{bmatrix}$	$\Delta = \begin{bmatrix} 0.3742 & 0.3566 & 0.3755 & 0.3923 & 0.1050 \\ 1 & 1 & 0.6244 & 1 & 1 \\ 0.3742 & 0 & 1 & 0.4545 & 0.3400 \\ 0 & 0.090 & 0 & 0 & 0 \end{bmatrix}$	$\gamma = \begin{bmatrix} 0.5719 & 0.5836 & 0.5710 & 0.5603 & 0.8264 \\ 0.3333 & 0.3333 & 0.4466 & 0.3333 & 0.3333 \\ 0.5719 & 1 & 0.3333 & 0.5238 & 0.5952 \\ 1 & 0.8474 & 1 & 1 & 1 \end{bmatrix}$	$\Gamma = \begin{bmatrix} 0.6226 \\ 0.3556 \\ 0.6048 \\ 0.9694 \end{bmatrix}$

MABAC

Çizelge 5.22 MABAC uygulama adımları

Adım 2	***Adım 3***	***Adım 4***	***Adım 5***
$V = \begin{bmatrix} 1.6257 & 1.6433 & 1.6244 & 1.6076 & 1.8950 \\ 1 & 1 & 1.3755 & 1 & 1 \\ 1.6257 & 2 & 1 & 1.5454 & 1.6600 \\ 2 & 1.91 & 2 & 2 & 2 \end{bmatrix}$	$G = \begin{bmatrix} 1.5162 & 1.5828 & 1.4539 & 1.4930 & 1.5837 \end{bmatrix}$	$Q = \begin{bmatrix} 0.1094 & 0.0604 & 0.1704 & 0.1146 & 0.3112 \\ -0.5162 & -0.5828 & -0.0783 & -0.4930 & -0.5837 \\ 0.1094 & 0.4171 & -0.4539 & 0.0524 & 0.0762 \\ 0.4837 & 0.3271 & 0.5460 & 0.5069 & 0.4162 \end{bmatrix}$	$S = \begin{bmatrix} 0.7662 \\ -2.2543 \\ 0.2012 \\ 2.2800 \end{bmatrix}$

Çizelge 5.17-5.22'de TOPSIS, COPRAS, EDAS, CODAS, GRA ve MABAC için uygulama adımları verilmiştir. Çizelge 5.17-5.18'de C_{TOPSIS} ve P_{COPRAS} sıralama değerlerine göre sınıflandırma yöntemlerinin sıralaması benzer olup toplama dayalı normalleştirme yaklaşımına göre RF öncelikli tercih edilir. Çizelge 5.19-5.20'de AS_{EDAS} ve H_{CODAS} sıralama değerlerine göre sınıflandırma yöntemlerinin sıralaması benzer olup orana dayalı normalleştirme yaklaşımına göre RF öncelikli tercih edilir. Çizelge 5.21-5.22'de Γ_{GRA} ve S_{MABAC} sıralama değerlerine göre sınıflandırma yöntemlerinin sıralaması benzer olup min-max değerine dayalı normalleştirme yaklaşımına göre RF öncelikli tercih edilir. Elde edilen tüm sıralama değerleri çizelge 5.23'te verilmiştir.

Çizelge 5.23 AHP ağırlıklı test veri seti için MCDM ile sınıflandırma yöntemlerinin karşılaştırılması

Yöntemler	C_{TOPSIS}	P_{COPRAS}	AS_{EDAS}	H_{CODAS}	Γ_{GRA}	S_{MABAC}
LR	0.7479	99.03	0.660	0.0594	0.6226	0.7662
***k*NN**	0.0914	96.52	0.0001	-0.1935	0.3556	-2.2543
SVM	0.6221	98.63	0.555	0.0213	0.6048	0.2012
RF	**0.9675**	**100.00**	**1.000**	**0.1415**	**0.9694**	**2.2800**

*Her bir MCDM için öncelikli tercih edilen sınıflandırma yöntemi koyu renk ile belirtilmiştir.

Çizelge 5.23'te AHP ağırlıklı test veri seti için hesaplanan performans değerleri MCDM yöntemleri ile karşılaştırılmıştır.

Çizelge 5.24 AHP ağırlıklı test veri seti için MCDM ile sınıflandırma yöntemlerinin sıralanması

	Yöntemler	***Sıralama***
Toplama dayalı	***TOPSIS***	RF ≫ LR ≫ SVM ≫ *k*NN
	COPRAS	RF ≫ LR ≫ SVM ≫ *k*NN
Orana dayalı	***EDAS***	RF ≫ LR ≫ SVM ≫ *k*NN
	CODAS	RF ≫ LR ≫ SVM ≫ *k*NN
Min-max değerine dayalı	***GRA***	RF ≫ LR ≫ SVM ≫ *k*NN
	MABAC	RF ≫ LR ≫ SVM ≫ *k*NN

Çizelge 5.24'e göre, AHP ile ağırlıklandırılmış veri seti için TOPSIS, COPRAS, EDAS, CODAS, GRA ve MABAC yöntemlerine göre RF öncelikli tercih edilir.

AHP ağırlıklı veri seti için elde edilen sınıflandırma performansı, objektif ağırlık belirleme yöntemi ile elde edilerek ağırlıklandırılmış veri setinin sınıflandırma performansı ile karşılaştırılmak istenmiştir. AHP sonucu belirlenen nitel değişkenlerin ağırlıklarının düşük olması sebebiyle sadece nicel değişkenleri kullanan ve objektif ağırlık belirleyen MCDM yöntemleri seçilmiştir. Bu amaçla, veri setine ait objektif ağırlıkların belirlenmesinde ENTROPY ve CRITIC yöntemleri kullanılmıştır. Eşitlik (2.11) ve eşitlik (2.15) ile değişken ağırlıkları hesaplanmıştır. Hesaplanan objektif değişken ağırlıkları çizelge 5.25'te verilmiştir.

Çizelge 5.25 ENTROPY ve CRITIC yöntemleri ile değişken ağırlıkları

Değişkenler	***ENTROPY***	***CRITIC***
Araştırmacı-Tasarımcı Sayısı (X_1)	0.0841	0.3607
Toplam Tamamlanan Proje Sayısı (X_2)	0.0761	0.1975
Toplam Devam eden Proje Sayısı (X_3)	0.0446	0.0646
Patent Başvuru Sayısı (X_4)	0.2853	0.2692
Patent Tescil Sayısı (X_5)	0.2807	0.1072
Merkezdeki Doktora Mezunu Personelin Oranı (X_6)	0.1866	0.0001
Merkezdeki Y. Lisans Mezunu Personelin Oranı (X_7)	0.0423	0.0004

Çizelge 5.25'e göre, ENTROPY ve CRITIC yöntemleri ile veriye dayalı objektif değişken ağırlıkları hesaplanmıştır. ENTROPY yöntemine göre Patent Başvuru Sayısı (X_4) en yüksek değişken ağırlığına $(w_4 = 0.2853)$ ve Merkezdeki Y. Lisans Mezunu Personelin Oranı (X_7) en düşük değişken ağırlığına $(w_7 = 0.0423)$ sahiptir. CRITIC yöntemine göre ise Araştırmacı-Tasarımcı Sayısı (X_1) en yüksek değişken ağırlığına $(w_1 = 0.3607)$ ve Merkezdeki Doktora Mezunu Personelin Oranı (X_6) en düşük değişken ağırlığına $(w_6 = 0.0001)$ sahiptir.

Çizelge 5.25'te verilen objektif değişken ağırlıkları kullanılarak veri seti çizelge 2.3'e göre ağırlıklandırılmıştır. ENTROPY ve CRITIC ile ağırlıklandırılmış veri setlerine, çizelge 5.13'te verilen ayarlanabilir parametreler kullanılarak sınıflandırma yöntemleri uygulanmıştır. ENTROPY ile ağırlıklandırılmış test veri setine ait performans değerleri çizelge 5.26'da verilmiştir.

Çizelge 5.26 ENTROPY ile ağırlıklandırılmış test veri setine ilişkin performans değerleri

	Performans Ölçütleri				
Yöntemler	***Doğruluk***	***Kesinlik***	***Duyarlılık***	***F_1-Skor***	***AUC***
LR	0.8543	0.8909	0.9254	0.9078	0.9083
***k*NN**	0.8458	0.9073	0.8922	0.8997	0.8711
SVM	0.8608	0.9002	0.9226	0.9113	0.9101
RF	0.8650	0.8986	0.9309	0.9145	0.9140

Çizelge 5.26'de ENTROPY ile ağırlıklandırılmış test veri seti için hesaplanan performans değerleri MCDM ile karşılaştırılmıştır.

Çizelge 5.27 ENTROPY ile ağırlıklandırılmış test veri seti için MCDM ile sınıflandırma yöntemlerinin karşılaştırılması

Yöntemler	C_{TOPSIS}	P_{COPRAS}	AS_{EDAS}	H_{CODAS}	Γ_{GRA}	S_{MABAC}
LR	0.6952	99.19	0.6078	0.0244	0.5810	-0.0220
***k*NN**	0.2076	97.64	0.0978	-0.1294	0.4666	-1.7448
SVM	0.8093	99.60	0.7736	0.0377	0.6957	1.0856
RF	**0.8783**	**100.00**	**0.9954**	**0.0802**	**0.8968**	**1.7230**

*Her bir MCDM için öncelikli tercih edilen sınıflandırma yöntemi koyu renk ile belirtilmiştir.

Çizelge 5.27'ye göre, ENTROPY ile ağırlıklandırılmış veri seti için TOPSIS, COPRAS, EDAS, CODAS, GRA ve MABAC yöntemlerine göre RF öncelikli tercih edilir. CRITIC ile ağırlıklandırılmış test veri setine ait performans değerleri çizelge 5.28'te verilmiştir.

Çizelge 5.28 CRITIC ile ağırlıklandırılmış test veri setine ilişkin performans değerleri

	Performans Ölçütleri				
Yöntemler	***Doğruluk***	***Kesinlik***	***Duyarlılık***	***F_1-Skor***	***AUC***
LR	0.8458	0.8918	0.9116	0.9016	0.9056
***k*NN**	0.8286	0.8790	0.9033	0.8910	0.8849
SVM	0.8586	0.9088	0.9088	0.9088	0.8987
RF	0.8650	0.8986	0.9309	0.9145	0.9140

Çizelge 5.29'da CRITIC ile ağırlıklandırılmış test veri seti için hesaplanan performans değerleri MCDM ile karşılaştırılmıştır.

Çizelge 5.29 CRITIC ile ağırlıklandırılmış veri seti için MCDM ile sınıflandırma yöntemlerinin karşılaştırılması

Yöntemler	C_{TOPSIS}	P_{COPRAS}	AS_{EDAS}	H_{CODAS}	Γ_{GRA}	S_{MABAC}
LR	0.4792	98.52	0.4600	-0.0114	0.4962	0.0366
***k*NN**	0.0001	96.97	0.0001	-0.1595	0.3333	-2.3285
SVM	0.6380	99.13	0.6900	0.0623	0.6570	0.9266
RF	**0.8605**	**100.00**	**1.0000**	**0.1300**	**0.9187**	**2.3292**

*Her bir MCDM için öncelikli tercih edilen sınıflandırma yöntemi koyu renk ile belirtilmiştir

Çizelge 5.29'a göre, CRITIC ile ağırlıklandırılmış veri seti için TOPSIS, COPRAS, EDAS, CODAS, GRA ve MABAC yöntemlerine göre RF öncelikli tercih edilir. Çizelge 5.30'ta ENTROPY ve CRITIC ile ağırlıklandırılmış veri setleri için hesaplanan performans değerlerinin MCDM ile sıralama sonuçları verilmiştir.

Çizelge 5.30 ENTROPY ve CRITIC ile ağırlıklandırılmış test veri seti için MCDM ile sınıflandırma yöntemlerinin sıralanması

	Yöntemler	***ENTROPY***	***CRITIC***
Toplama dayalı	***TOPSIS***	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$
	COPRAS	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$
Orana dayalı	***EDAS***	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$
	CODAS	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$
Min-max değerine dayalı	***GRA***	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$
	MABAC	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$	$\mathrm{RF} \gg \mathrm{SVM} \gg \mathrm{LR} \gg k\mathrm{NN}$

Çizelge 5.30'a göre, ENTROPY ve CRITIC ile ağırlıklandırılmış veri setlerinin sınıflandırılmasında RF öncelikli tercih edilir. Buna göre hem subjektif olarak değişken ağırlıklarının belirlendiği AHP hem de objektif olarak değişken ağırlıklarının belirlendiği ENTROPY ve CRITIC ile ağırlıklandırılmış veri setlerinin sınıflandırılmasında RF'nin öncelikli tercih edilir olduğu sonucuna ulaşılmıştır.

5.2 Simülasyon Çalışması

Gerçek veri setleri, kişisel ve gizli bilgi içerdiğinden bir çalışma kapsamında temin edilmesi her zaman mümkün olmayabilir. Simülasyon, gerçek veri setlerinin özelliklerine benzetilerek bilgi gizliliğini koruyarak anonimlik sağlar. Simülasyon çalışması, gerçek dünyada bulunan çeşitli senaryoları göz önünde bulundurarak daha kapsayıcı çalışmaların yapılmasına imkan verir. Simülasyon ile yapılan analiz çalışmaları, çeşitli olası senaryoların modellenmesi ya da gerçek veri seti ile elde edilen sonuçların genelleştirilmesi bakımından oldukça değerlidir.

Bu çalışmada, gerçek veri seti ile yapılan sınıflandırma çalışmasının karşılaştırılabilmesi amacıyla simülasyon çalışması yapılmıştır. Simülasyon ile veri setlerinin üretilmesinde olası farklı senaryolar kurgulanmıştır. Bu amaçla, $n = \{150, 250, 2000\}$ olmak üzere üç farklı örneklem çapına göre senaryolar oluşturulmuştur. Uygulanan senaryolarda bağımlı

değişken, Ar-Ge ve Tasarım Merkezi belgesi alan firmaların faaliyetlerine devam edip etmeme durumunu anlatan kategorik bir değişkendir. Bu değişken, sınıflandırma çalışmasında ele alınan veri setinin dengeli (faaliyette ve faaliyette olmayan firmaların dağılımının dengeli olduğu) veya dengesiz (faaliyette ve faaliyette olmayan firmaların dağılımının dengesiz olduğu) dağılıma sahip bir veri seti olduğunu belirtmektedir. Bağımlı değişken, $p = \{0.50, 0.80\}$ olacak biçimde Binom dağılımı kullanılarak üretilmiştir. Bağımsız değişkenler, kesikli nicel değişkenler için Poisson dağılımı, sürekli nicel değişkenler için Üstel dağılım kullanılarak farklı parametreler ile oluşturulmuştur. Rastgele üretilen yedi bağımsız değişken için kullanılan dağılım adı ve parametre değeri çizelge 5.31'de verilmiştir.

Çizelge 5.31 Simülasyon için bağımsız değişken bilgisi

Değişken	***Dağılım Adı (Parametre Değeri)***
X_1	$Poisson(\lambda_1 = 0.5)$
X_2	$Poisson(\lambda_2 = 0.3)$
X_3	$Poisson(\lambda_3 = 0.2)$
X_4	$Poisson(\lambda_4 = 0.1)$
X_5	$Poisson(\lambda_5 = 0.1)$
X_6	$Üstel(\lambda_6 = 0.0083)$
X_7	$Üstel(\lambda_7 = 0.095)$

Gerçek veri seti analizine benzer olarak veri madenciliği sınıflandırma yöntemlerinden LR, *k*NN, SVM ve RF simülasyon çalışmasında üretilen veri setlerine uygulanmıştır. Sınıflandırma yöntemlerinin performansının ölçülebilmesi amacıyla veri seti, %80 eğitim seti ile %20 test seti olmak üzere ayrılmıştır. Simülasyon çalışmasında, veri setlerinin oluşturulması, sınıflandırma yöntemlerinin algoritmaları ve performans ölçütlerinin hesaplanması bakımından 1000 tekrar yapılmıştır. Araştırmacı ile belirlenen ayarlanabilir parametreleri uygulanarak Python 3.11.3 sürümünde Scikit-learn kütüphanesinde sınıflandırma yöntemlerinin performansı 5-kat CV kullanılarak test edilmiş, ızgara arama yöntemi ile optimal parametreler elde edilmiştir. Araştırmacı tarafından belirlenen ve ızgara arama ile seçilen ayarlanabilir parametreleri çizelge 5.32'de verilmiştir.

Çizelge 5.32 Simülasyon çalışmasında kullanılan ayarlanabilir parametreleri

	Yöntemler	*Ayarlanabilir Parametreleri*
Araştırmacı ile belirlenen	**LR**	$c = \{1e-05,\ 3,\ 0.1\},\ Penalty = \{l_1, l_2,\ None\}$
	***k*NN**	$Metric = \{Minkowski,\ Euclidean\},\ Neighbors(k) = \{3,4,\ldots,25\}$
	SVM	$c = \{0,\ 0.1,\ 1,\ 10,\ 100\},\ Gamma = \{0.001,\ 0.01,\ 1,\ 10\},$ $Kernel = \{Polynomial,\ Rbf,\ Sigmoid,\ Linear\}$
	RF	$Bootstrap = True, Criterion = \{Gini,\ Entropy\},$ $Max\_Depth = \{3,\ 5\}, Max\_Features = \{2,\ 3\},$ $Min\_Samples\_Leaf = \{5,\ 8,\ 10\},$ $Min\_Samples\_Split = \{3,\ 4,\ 5,\ 6\},$ $n\_Estimator = \{100,\ 200,\ 250,\ 500\}$
Izgara arama ile seçilen	**LR**	$c = 1e-05,\ Penalty = l_2 - 05$
	***k*NN**	$Metric = Euclidean,\ Neighbors(k) = 9$
	SVM	$c = 0.1,\ Gamma = 0.001,\ Kernel = Polynomial$
	RF	$Bootstrap = True, Criterion = Gini, Max\_Depth = 3,$ $Max\_Features = 2, Min\_Samples\_Leaf = 5,$ $Min\_Samples\_Split = 3,\ n\_Estimator = 100$

Simülasyon çalışması için oluşturulan farklı senaryolar çizelge 5.33'te verilmiştir.

Çizelge 5.33 Simülasyon çalışması için oluşturulan senaryolar

Senaryolar	*Bağımlı Değişken* $(Binom(n, p))$
Senaryo1	$Binom(n = 150,\ p = 0.50)$
Senaryo2	$Binom(n = 250,\ p = 0.50)$
Senaryo3	$Binom(n = 2000,\ p = 0.50)$
Senaryo4	$Binom(n = 150,\ p = 0.80)$
Senaryo5	$Binom(n = 250,\ p = 0.80)$
Senaryo6	$Binom(n = 2000,\ p = 0.80)$

Simülasyon çalışmasında uygulanan sınıflandırma yöntemlerinin değerlendirilmesinde performans ölçütlerinden doğruluk, kesinlik, duyarlılık, F_1-Skor ve AUC değerleri test veri seti için hesaplanarak çizelge 5.34'te verilmiştir.

Çizelge 5.34 1000 tekrar ile çalıştırılmış test veri seti için performans değerleri ve karar matrisi biçiminde gösterimi

	Performans Değerleri ve Karar Matrisi Biçimi						
	Yöntemler	***Doğruluk***	***Kesinlik***	***Duyarlılık***	***F_1-Skor***	***AUC***	***Karar Matrisi***
Senaryo1	LR	0.4965	0.3917	0.4922	0.3964	0.4919	$D_{Senaryo1}=\begin{bmatrix} 0.4965 & 0.3917 & 0.4922 & 0.3964 & 0.4919 \\ 0.4998 & 0.4979 & 0.4277 & 0.4339 & 0.4983 \\ 0.4948 & 0.4555 & 0.4951 & 0.4386 & 0.4960 \\ 0.4971 & 0.4904 & 0.5071 & 0.4664 & 0.4965 \end{bmatrix}$
	*k*NN	0.4998	0.4979	0.4277	0.4339	0.4983	
	SVM	0.4948	0.4555	0.4951	0.4386	0.4960	
	RF	0.4971	0.4904	0.5071	0.4664	0.4965	
Senaryo2	LR	0.5012	0.3882	0.4932	0.3986	0.4984	$D_{Senaryo2}=\begin{bmatrix} 0.5012 & 0.3882 & 0.4932 & 0.3986 & 0.4984 \\ 0.5008 & 0.4932 & 0.4298 & 0.4406 & 0.4978 \\ 0.4967 & 0.4525 & 0.4860 & 0.4378 & 0.5028 \\ 0.5004 & 0.4879 & 0.4952 & 0.4685 & 0.4950 \end{bmatrix}$
	*k*NN	0.5008	0.4932	0.4298	0.4406	0.4978	
	SVM	0.4967	0.4525	0.4860	0.4378	0.5028	
	RF	0.5004	0.4879	0.4952	0.4685	0.4950	
Senaryo3	LR	0.4992	0.3769	0.4941	0.3996	0.5002	$D_{Senaryo3}=\begin{bmatrix} 0.4992 & 0.3769 & 0.4941 & 0.3996 & 0.5002 \\ 0.5003 & 0.5006 & 0.4187 & 0.4472 & 0.5006 \\ 0.4991 & 0.4799 & 0.4943 & 0.4457 & 0.4999 \\ 0.5010 & 0.5008 & 0.4957 & 0.4744 & 0.5014 \end{bmatrix}$
	*k*NN	0.5003	0.5006	0.4187	0.4472	0.5006	
	SVM	0.4991	0.4799	0.4943	0.4457	0.4999	
	RF	0.5010	0.5008	0.4957	0.4744	0.5014	
Senaryo4	LR	0.7999	0.8000	0.9998	0.8870	0.8085	$D_{Senaryo4}=\begin{bmatrix} 0.7999 & 0.8000 & 0.9998 & 0.8870 & 0.8085 \\ 0.7811 & 0.7994 & 0.9698 & 0.8740 & 0.8183 \\ 0.7933 & 0.8002 & 0.9858 & 0.8823 & 0.8111 \\ 0.7996 & 0.8001 & 0.9991 & 0.8868 & 0.8059 \end{bmatrix}$
	*k*NN	0.7811	0.7994	0.9698	0.8740	0.8183	
	SVM	0.7933	0.8002	0.9858	0.8823	0.8111	
	RF	0.7996	0.8001	0.9991	0.8868	0.8059	
Senaryo5	LR	0.8020	0.8026	0.9999	0.8889	0.8110	$D_{Senaryo5}=\begin{bmatrix} 0.8020 & 0.8026 & 0.9999 & 0.8889 & 0.8110 \\ 0.7862 & 0.8034 & 0.9709 & 0.8778 & 0.8213 \\ 0.7976 & 0.8023 & 0.9923 & 0.8860 & 0.8101 \\ 0.8023 & 0.8026 & 0.9994 & 0.8891 & 0.8107 \end{bmatrix}$
	*k*NN	0.7862	0.8034	0.9709	0.8778	0.8213	
	SVM	0.7976	0.8023	0.9923	0.8860	0.8101	
	RF	0.8023	0.8026	0.9994	0.8891	0.8107	
Senaryo6	LR	0.8003	0.8003	1.0000	0.8892	0.8013	$D_{Senaryo6}=\begin{bmatrix} 0.8003 & 0.8003 & 1.0000 & 0.8892 & 0.8013 \\ 0.7848 & 0.8002 & 0.9804 & 0.8810 & 0.8135 \\ 0.7990 & 0.8003 & 0.9993 & 0.8886 & 0.8015 \\ 0.8004 & 0.8003 & 0.9999 & 0.8893 & 0.8021 \end{bmatrix}$
	*k*NN	0.7848	0.8002	0.9804	0.8810	0.8135	
	SVM	0.7990	0.8003	0.9993	0.8886	0.8015	
	RF	0.8004	0.8003	0.9999	0.8893	0.8021	

Çizelge 5.34'e göre performans değerleri bakımından sınıflandırma yöntemlerini karşılaştırmak zordur. Karşılaştırma yapılabilmesi için MCDM yöntemleri uygulanmıştır. Sınıflandırma yöntemleri alternatifler ve performans ölçütleri kriterler olmak üzere bir karar verme problemine dönüştürülmüştür. Buna göre, çizelge 5.34'te her bir senaryo için karar matrisleri verilmiştir. Çizelge 5.34'te verilen senaryolar ayrı ayrı değerlendirilmiştir.

Senaryo1:

Senaryo1'de, $Y \sim Binom(n=150, p=0.50)$ dağılımlı değişkene sahip veri seti 1000 tekrar ile rastgele üretilmiştir. Her bir sınıflandırma yöntemi için 1000 tekrarla hesaplanan performans değerlerinin standart sapmaları hesaplanmıştır. 1000 tekrarla üretilmiş test veri seti için performans değerlerinin ortalaması ve standart sapmaları çizelge 5.35'te verilmiştir.

Çizelge 5.35 Sınıflandırma performans değerleri ve standart sapmaları (Senaryo1)

Yöntemler	***Doğruluk***	***Kesinlik***	***Duyarlılık***	***F_1-Skor***	***AUC***
LR	0.4960 (0.0921)	0.3904 (0.2513)	0.4718 (0.3881)	0.3817 (0.2682)	0.4922 (0.1124)
*k*NN	0.4996 (0.0892)	0.4955 (0.1642)	0.4201 (0.2055)	0.4291 (0.1590)	0.4984 (0.1017)
SVM	0.5013 (0.0947)	0.4623 (0.2099)	0.4905 (0.3032)	0.4409 (0.2127)	0.4991 (0.1113)
RF	0.4970 (0.0944)	0.4882 (0.1659)	0.4900 (0.2529)	0.4563 (0.1712)	0.4965 (0.1127)

*Her bir performans ölçüt değerinin yanında parantez içerisinde standart sapma değeri verilmiştir.

Çizelge 5.35'de görülen değerlere göre performans değerleri bakımından sınıflandırma yöntemlerini karşılaştırmak zordur. Objektif karşılaştırma yapılabilmesi için karar verme yöntemlerinin kullanılması gerekir. Bu amaçla çizelge 5.36'da verilen performans değerlerine göre sınıflandırma yöntemi seçiminde MCDM yöntemleri uygulanır. Karar verme yöntemlerinin uygulanabilmesi için karar matrisine ihtiyaç duyulur. Karar matrisinin oluşturulmasında test veri setinin performans değerleri kullanılır. Buna göre, çizelge 5.35'te verilen performans değerleri kullanılarak eşitlik (4.6)'ya göre oluşturulan karar matrisi

$$D_{Senaryo1} = \begin{bmatrix} 0.4965 & 0.3917 & 0.4922 & 0.3964 & 0.4919 \\ 0.4998 & 0.4979 & 0.4277 & 0.4339 & 0.4983 \\ 0.4948 & 0.4555 & 0.4951 & 0.4386 & 0.4960 \\ 0.4971 & 0.4904 & 0.5071 & 0.4664 & 0.4965 \end{bmatrix} \tag{5.3}$$

biçimindedir. Her bir kriter eşit öneme sahip olacak biçimde değerlendirilmiştir. Performans değerleri maksimum olacak biçimde fayda amacına uygun olan formüller çizelge 4.2'deki gibi kullanılarak eşitlik (5.3)'teki karar matrisi normalleştirilir. Gerçek veri setinde kullanılan yöntemlere benzer olarak MCDM yöntemleri uygulanır. Senaryo1 için MCDM ile sınıflandırma yöntemlerinin karşılaştırılması çizelge 5.36'da verilmiştir.

Çizelge 5.36 MCDM ile sınıflandırma yöntemlerinin karşılaştırılması (Senaryo1)

Yöntemler	C_{TOPSIS}	P_{COPRAS}	AS_{EDAS}	H_{CODAS}	Γ_{GRA}	S_{MABAC}
LR	0.3211	91.96	0.0604	-0.3465	0.4316	-1.4701
*k*NN	0.5766	95.87	0.5113	0.0764	0.7703	0.9132
SVM	0.6557	98.71	0.5799	-0.0144	0.5592	0.0706
RF	**0.9472**	**100.00**	**1.0000**	**0.4125**	**0.7994**	**1.4856**

*Her bir MCDM için öncelikli tercih edilen sınıflandırma yöntemi koyu renk ile belirtilmiştir.

Çizelge 5.36'daki sonuçlara göre MCDM yöntemlerinin sıralaması çizelge 5.37'de verilmiştir.

Çizelge 5.37 MCDM ile sınıflandırma yöntemlerinin sıralanması (Senaryo1)

	Yöntemler	***Sıralama***
Toplama dayalı	***TOPSIS***	RF ≫ SVM ≫ *k*NN ≫ LR
	COPRAS	RF ≫ SVM ≫ *k*NN ≫ LR
Orana dayalı	***EDAS***	RF ≫ SVM ≫ *k*NN ≫ LR
	CODAS	RF ≫ *k*NN ≫ SVM ≫ LR
Min-max değerine dayalı	***GRA***	RF ≫ *k*NN ≫ SVM ≫ LR
	MABAC	RF ≫ *k*NN ≫ SVM ≫ LR

Çizelge 5.37'de verilen sonuçlara göre Senaryo1 için RF öncelikli tercih edilir.

Senaryo2:

Senaryo2'de, $Y \sim Binom(n = 250,\ p = 0.50)$ dağılımlı değişkene sahip veri seti için 1000 tekrar ile rastgele üretilmiştir. Her bir sınıflandırma yöntemi için 1000 tekrarla hesaplanan performans değerlerinin standart sapmaları hesaplanmıştır. 1000 tekrarla üretilmiş test veri seti için performans değerlerinin ortalaması ve standart sapmaları çizelge 5.38'te verilmiştir.

Çizelge 5.38 Sınıflandırma performans değerleri ve standart sapmaları (Senaryo2)

Yöntemler	***Doğruluk***	***Kesinlik***	***Duyarlılık***	***F_1-Skor***	***AUC***
LR	0.5012 (0.0680)	0.3882 (0.2318)	0.4932 (0.3835)	0.3986 (0.2631)	0.4984 (0.0826)
*k*NN	0.5008 (0.0702)	0.4932 (0.1139)	0.4298 (0.1594)	0.4406 (0.1239)	0.4978 (0.0790)
SVM	0.4967 (0.0690)	0.4525 (0.1727)	0.4860 (0.2856)	0.4378 (0.1976)	0.5028 (0.0808)
RF	0.5004 (0.0717)	0.4879 (0.1174)	0.4952 (0.2220)	0.4685 (0.1480)	0.4950 (0.0829)

*Her bir performans ölçüt değerinin yanında parantez içerisinde standart sapma değeri verilmiştir.

Çizelge 5.38'te verilen performans değerleri kullanılarak eşitlik (4.6)'ya göre oluşturulan karar matrisi

$$D_{Senaryo2} = \begin{bmatrix} 0.5012 & 0.3882 & 0.4932 & 0.3986 & 0.4984 \\ 0.5008 & 0.4932 & 0.4298 & 0.4406 & 0.4978 \\ 0.4967 & 0.4525 & 0.4860 & 0.4378 & 0.5028 \\ 0.5004 & 0.4879 & 0.4952 & 0.4685 & 0.4950 \end{bmatrix} \tag{5.4}$$

biçimindedir. Eşitlik (5.4)'deki karar matrisi kullanılarak MCDM yöntemleri uygulanır. Senaryo2 için MCDM ile sınıflandırma yöntemlerinin karşılaştırılması çizelge 5.39'da verilmiştir.

Çizelge 5.39 MCDM ile sınıflandırma yöntemlerinin karşılaştırılması (Senaryo2)

Yöntemler	C_{TOPSIS}	P_{COPRAS}	AS_{EDAS}	H_{CODAS}	Γ_{GRA}	S_{MABAC}
LR	0.3228	92.76	0.1043	-0.3222	0.6157	-0.3484
*k*NN	0.6216	96.46	0.5436	0.1052	0.6353	0.1172
SVM	0.6397	96.92	0.5610	-0.0529	0.6418	0.2787
RF	**0.9394**	**100.00**	**0.9850**	**0.3760**	**0.7958**	**1.0180**

*Her bir MCDM için öncelikli tercih edilen sınıflandırma yöntemi koyu renk ile belirtilmiştir.

Çizelge 5.39'daki sonuçlara göre MCDM yöntemlerinin sıralaması çizelge 5.40'ta verilmiştir.

Çizelge 5.40 MCDM ile sınıflandırma yöntemlerinin sıralanması (Senaryo2)

	Yöntemler	***Sıralama***
Toplama dayalı	***TOPSIS***	$\text{RF} \gg \text{SVM} \gg k\text{NN} \gg \text{LR}$
	COPRAS	$\text{RF} \gg \text{SVM} \gg k\text{NN} \gg \text{LR}$
Orana dayalı	***EDAS***	$\text{RF} \gg \text{SVM} \gg k\text{NN} \gg \text{LR}$
	CODAS	$\text{RF} \gg k\text{NN} \gg \text{SVM} \gg \text{LR}$
Min-max değerine dayalı	***GRA***	$\text{RF} \gg \text{SVM} \gg k\text{NN} \gg \text{LR}$
	MABAC	$\text{RF} \gg \text{SVM} \gg k\text{NN} \gg \text{LR}$

Çizelge 5.40'a göre Senaryo2 için RF öncelikli tercih edilir.

Senaryo3:

Senaryo3'te, $Y \sim Binom(n = 2000, p = 0.50)$ dağılımlı değişkene sahip veri seti 1000 tekrar ile rastgele üretilmiştir. Her bir sınıflandırma yöntemi için 1000 tekrarla hesaplanan performans değerlerinin standart sapmaları hesaplanmıştır. 1000 tekrarla üretilmiş test

veri seti için performans değerlerinin ortalaması ve standart sapmaları çizelge 5.41'de verilmiştir.

Çizelge 5.41 Sınıflandırma performans değerleri ve standart sapmaları (Senaryo3)

Yöntemler	***Doğruluk***	***Kesinlik***	***Duyarlılık***	***F_1-Skor***	***AUC***
LR	0.4992 (0.0243)	0.3769 (0.2166)	0.4941 (0.3850)	0.3996 (0.2626)	0.5002 (0.0306)
*k*NN	0.5003 (0.0252)	0.5006 (0.0401)	0.4187 (0.1134)	0.4472 (0.0768)	0.5006 (0.0268)
SVM	0.4991 (0.0246)	0.4799 (0.1174)	0.4943 (0.2852)	0.4457 (0.1885)	0.4999 (0.0268)
RF	0.5010 (0.0256)	0.5008 (0.0415)	0.4957 (0.2046)	0.4744 (0.1239)	0.5014 (0.0296)

*Her bir performans ölçüt değerinin yanında parantez içerisinde standart sapma değeri verilmiştir.

Çizelge 5.41'de verilen performans ölçüt değerleri kullanılarak eşitlik (4.6)'ya göre oluşturulan karar matrisi

$$D_{Senaryo3} = \begin{bmatrix} 0.4992 & 0.3769 & 0.4941 & 0.3996 & 0.5002 \\ 0.5003 & 0.5006 & 0.4187 & 0.4472 & 0.5006 \\ 0.4991 & 0.4799 & 0.4943 & 0.4457 & 0.4999 \\ 0.5010 & 0.5008 & 0.4957 & 0.4744 & 0.5014 \end{bmatrix} \tag{5.5}$$

biçimindedir. Eşitlik (5.5)'teki karar matrisi kullanılarak MCDM yöntemleri uygulanır. Senaryo3 için MCDM ile sınıflandırma yöntemlerinin karşılaştırılması çizelge 5.42'de verilmiştir.

Çizelge 5.42 MCDM ile sınıflandırma yöntemlerinin karşılaştırılması (Senaryo3)

Yöntemler	C_{TOPSIS}	P_{COPRAS}	AS_{EDAS}	H_{CODAS}	Γ_{GRA}	S_{MABAC}
LR	0.3346	91.43	0.0976	-0.4161	0.4713	1.1385
*k*NN	0.6231	95.63	0.5202	0.0645	0.5937	-2.8567
SVM	0.7864	97.96	0.6999	0.1077	0.5890	1.0376
RF	**1.0000**	**100.00**	**1.0000**	**0.3985**	**1.0000**	**1.9760**

*Her bir MCDM için öncelikli tercih edilen sınıflandırma yöntemi koyu renk ile belirtilmiştir.

Çizelge 5.42'deki sonuçlara göre MCDM yöntemlerinin sıralaması çizelge 5.43'te verilmiştir.

Çizelge 5.43 MCDM ile sınıflandırma yöntemlerinin sıralanması (Senaryo3)

	Yöntemler	***Sıralama***
Toplama dayalı	***TOPSIS***	RF ≫ SVM ≫ *k*NN ≫ LR
	COPRAS	RF ≫ SVM ≫ *k*NN ≫ LR
Orana dayalı	***EDAS***	RF ≫ SVM ≫ *k*NN ≫ LR
	CODAS	RF ≫ SVM ≫ *k*NN ≫ LR
Min-max değerine dayalı	***GRA***	RF ≫ *k*NN ≫ SVM ≫ LR
	MABAC	RF ≫ SVM ≫ *k*NN ≫ LR

Çizelge 5.43'e göre Senaryo3 için RF öncelikli tercih edilir. Dengeli veri setlerinde RF algoritmasının sınıflandırma performansı bakımından öncelikli olarak tercih edildiği görülmektedir.

Senaryo4:

Senaryo4'te, $Y \sim Binom(n = 150, p = 0.80)$ dağılımlı değişkene sahip veri seti için 1000 tekrar ile rastgele üretilmiştir. Her bir sınıflandırma yöntemi için 1000 tekrarla hesaplanan performans değerlerinin standart sapmaları hesaplanmıştır. 1000 tekrarla üretilmiş test veri seti için performans değerlerinin ortalaması ve standart sapmaları çizelge 5.44'te verilmiştir.

Çizelge 5.44 Sınıflandırma performans değerleri ve standart sapmaları (Senaryo4)

Yöntemler	***Doğruluk***	***Kesinlik***	***Duyarlılık***	***F_1-Skor***	***AUC***
LR	0.7999 (0.0726)	0.8000 (0.0725)	0.9998 (0.0038)	0.8870 (0.0456)	0.8085 (0.0965)
*k*NN	0.7811 (0.0774)	0.7994 (0.0733)	0.9698 (0.0504)	0.8740 (0.0507)	0.8183 (0.0920)
SVM	0.7933 (0.0750)	0.8002 (0.0730)	0.9858 (0.0346)	0.8823 (0.0478)	0.8111 (0.0976)
RF	0.7996 (0.0726)	0.8001 (0.0726)	0.9991 (0.0065)	0.8868 (0.0455)	0.8059 (0.0977)

*Her bir performans ölçüt değerinin yanında parantez içerisinde standart sapma değeri verilmiştir.

Çizelge 5.44'te verilen performans ölçüt değerleri kullanılarak eşitlik (4.6)'ya göre oluşturulan karar matrisi

$$D_{Senaryo4} = \begin{bmatrix} 0.7999 & 0.8000 & 0.9998 & 0.8870 & 0.8085 \\ 0.7811 & 0.7994 & 0.9698 & 0.8740 & 0.8183 \\ 0.7933 & 0.8002 & 0.9858 & 0.8823 & 0.8111 \\ 0.7996 & 0.8001 & 0.9991 & 0.8868 & 0.8059 \end{bmatrix} \quad (5.6)$$

biçimindedir. Eşitlik (5.6)'daki karar matrisi kullanılarak MCDM yöntemleri uygulanır. Senaryo4 için MCDM ile sınıflandırma yöntemlerinin karşılaştırılması çizelge 5.45'te verilmiştir.

Çizelge 5.45 MCDM ile sınıflandırma yöntemlerinin karşılaştırılması (Senaryo4)

Yöntemler	C_{TOPSIS}	P_{COPRAS}	AS_{EDAS}	H_{CODAS}	Γ_{GRA}	S_{MABAC}
LR	**0.7738**	**100.00**	**0.9563**	**0.0443**	**0.8108**	**1.2324**
*k*NN	0.2705	98.85	0.1844	-0.0603	0.4666	-1.7272
SVM	0.5636	99.51	0.4736	-0.0215	0.6295	0.5128
RF	0.7225	**99.91**	0.9065	**0.0404**	**0.8055**	1.0931

*Her bir MCDM için öncelikli tercih edilen sınıflandırma yöntemi koyu renk ile belirtilmiştir.

Çizelge 5.45'teki sonuçlara göre MCDM yöntemlerinin sıralaması çizelge 5.46'da verilmiştir.

Çizelge 5.46 MCDM ile sınıflandırma yöntemlerinin sıralanması (Senaryo4)

	Yöntemler	***Sıralama***
Toplama dayalı	***TOPSIS***	LR ≫ RF ≫ SVM ≫ *k*NN
	COPRAS	LR ≃ RF ≫ SVM ≫ *k*NN
Orana dayalı	***EDAS***	LR ≫ RF ≫ SVM ≫ *k*NN
	CODAS	LR ≃ RF ≫ SVM ≫ *k*NN
Min-max değerine dayalı	***GRA***	LR ≃ RF ≫ SVM ≫ *k*NN
	MABAC	LR ≫ RF ≫ SVM ≫ *k*NN

Çizelge 5.46'ya göre örneklem çapının küçük ve dengesiz dağılıma sahip veri setinin kurgulandığı Senaryo4 için TOPSIS ve MABAC dışında LR ve RF öncelikli tercih edilir. TOPSIS ve MABAC yöntemleri için RF ikinci sırada tercih edilse de çizelge 5.44'e göre,

LR ve RF yöntemleri arasında performans değerlerinin ortalaması ve standart sapmaları bakımından önemli bir fark bulunmamaktadır.

Senaryo5:

Senaryo5'te, $Y \sim Binom(n = 250, p = 0.80)$ dağılımlı değişkene sahip veri seti için 1000 tekrar ile rastgele üretilmiştir. Her bir sınıflandırma yöntemi için 1000 tekrarla hesaplanan performans değerlerinin standart sapmaları hesaplanmıştır. 1000 tekrarla üretilmiş test veri seti için performans değerlerinin ortalaması ve standart sapmaları çizelge 5.47'de verilmiştir.

Çizelge 5.47 Sınıflandırma performans değerleri ve standart sapmaları (Senaryo5)

Yöntemler	***Doğruluk***	***Kesinlik***	***Duyarlılık***	***F_1-Skor***	***AUC***
LR	0.8020 (0.0568)	0.8026 (0.0565)	0.9999 (0.0069)	0.8889 (0.0353)	0.8110 (0.0769)
*k*NN	0.7862 (0.0626)	0.8034 (0.0573)	0.9709 (0.0379)	0.8778 (0.0383)	0.8213 (0.0733)
SVM	0.7976 (0.0590)	0.8023 (0.0567)	0.9923 (0.0273)	0.8860 (0.0375)	0.8101 (0.0792)
RF	0.8023 (0.0566)	0.8026 (0.0564)	0.9994 (0.0047)	0.8891 (0.0353)	0.8107 (0.0740)

*Her bir performans ölçüt değerinin yanında parantez içerisinde standart sapma değeri verilmiştir.

Çizelge 5.47'de verilen performans ölçüt değerleri kullanılarak eşitlik (4.6)'ya göre oluşturulan karar matrisi

$$D_{Senaryo5} = \begin{bmatrix} 0.8020 & 0.8026 & 0.9999 & 0.8889 & 0.8110 \\ 0.7862 & 0.8034 & 0.9709 & 0.8778 & 0.8213 \\ 0.7976 & 0.8023 & 0.9923 & 0.8860 & 0.8101 \\ 0.8023 & 0.8026 & 0.9994 & 0.8891 & 0.8107 \end{bmatrix} \quad (5.7)$$

biçimindedir. Eşitlik (5.7)'deki karar matrisi kullanılarak MCDM yöntemleri uygulanır. Senaryo5 için MCDM ile sınıflandırma yöntemlerinin karşılaştırılması çizelge 5.48'te verilmiştir.

Çizelge 5.48 MCDM ile sınıflandırma yöntemlerinin karşılaştırılması (Senaryo5)

Yöntemler	C_{TOPSIS}	P_{COPRAS}	AS_{EDAS}	H_{CODAS}	Γ_{GRA}	S_{MABAC}
LR	**0.7469**	**100.00**	**0.9624**	**0.0347**	**0.7378**	**0.8962**
*k*NN	0.2680	99.04	0.2731	-0.0605	0.6000	-0.4204
SVM	0.6146	99.64	0.5247	-0.0061	0.5199	0.2487
RF	**0.7408**	**99.99**	**0.9605**	**0.0342**	**0.7439**	**0.8886**

*Her bir MCDM için öncelikli tercih edilen sınıflandırma yöntemi koyu renk ile belirtilmiştir.

Çizelge 5.48'teki sonuçlara göre MCDM yöntemlerinin sıralaması çizelge 5.49'da verilmiştir.

Çizelge 5.49 MCDM ile sınıflandırma yöntemlerinin sıralanması (Senaryo5)

	Yöntemler	***Sıralama***
Toplama dayalı	***TOPSIS***	LR ≃ RF ≫ SVM ≫ *k*NN
	COPRAS	LR ≃ RF ≫ SVM ≫ *k*NN
Orana dayalı	***EDAS***	LR ≃ RF ≫ SVM ≫ *k*NN
	CODAS	LR ≃ RF ≫ SVM ≫ *k*NN
Min-max değerine dayalı	***GRA***	LR ≃ RF ≫ *k*NN ≫ SVM
	MABAC	LR ≃ RF ≫ SVM ≫ *k*NN

Çizelge 5.49'a göre Senaryo5 için LR ve RF öncelikli tercih edilir. Çizelge 5.47'ye göre, LR ve RF yöntemleri arasında performans değerlerinin ortalaması ve standart sapmaları bakımından önemli bir fark bulunmamaktadır.

Senaryo6:

Senaryo6'da, $Y \sim Binom(n = 2000, p = 0.80)$ dağılımlı değişkene sahip veri seti için 1000 tekrar ile rastgele üretilmiştir. Gerçek veri setine benzer bir dağılım seçilmiştir. Her bir sınıflandırma yöntemi için 1000 tekrarla hesaplanan performans değerlerinin standart sapmaları hesaplanmıştır. 1000 tekrarla üretilmiş test veri seti için performans değerlerinin ortalaması ve standart sapmaları çizelge 5.50'de verilmiştir.

Çizelge 5.50 Sınıflandırma performans değerleri ve standart sapmaları (Senaryo6)

Yöntemler	***Doğruluk***	***Kesinlik***	***Duyarlılık***	***F_1-Skor***	***AUC***
LR	0.8003 (0.0213)	0.8003 (0.0213)	1.0000 (0.0001)	0.8892 (0.0131)	0.8013 (0.0290)
*k*NN	0.7848 (0.0214)	0.8002 (0.0212)	0.9804 (0.0108)	0.8810 (0.0135)	0.8135 (0.0288)
SVM	0.7990 (0.0214)	0.8003 (0.0213)	0.9993 (0.0025)	0.8886 (0.0132)	0.8015 (0.0288)
RF	0.8004 (0.0213)	0.8003 (0.0213	0.9999 (0.0002)	0.8893 (0.0131)	0.8021 (0.0287)

*Her bir performans ölçüt değerinin yanında parantez içerisinde standart sapma değeri verilmiştir.

Çizelge 5.50'de verilen performans ölçüt değerleri kullanılarak eşitlik (4.6)'ya göre oluşturulan karar matrisi

$$D_{Senaryo6} = \begin{bmatrix} 0.8003 & 0.8003 & 1.0000 & 0.8892 & 0.8013 \\ 0.7848 & 0.8002 & 0.9804 & 0.8810 & 0.8135 \\ 0.7990 & 0.8003 & 0.9993 & 0.8886 & 0.8015 \\ 0.8004 & 0.8003 & 0.9999 & 0.8893 & 0.8021 \end{bmatrix} \tag{5.8}$$

biçimindedir. Eşitlik (5.8)'deki karar matrisi kullanılarak MCDM yöntemleri uygulanır. Senaryo6 için MCDM ile sınıflandırma yöntemlerinin karşılaştırılması çizelge 5.51'de verilmiştir.

Çizelge 5.51 MCDM ile sınıflandırma yöntemlerinin karşılaştırılması(Senaryo6)

Yöntemler	C_{TOPSIS}	P_{COPRAS}	AS_{EDAS}	H_{CODAS}	Γ_{GRA}	S_{MABAC}
LR	0.6584	99.98	0.9336	0.0150	0.8550	1.1177
*k*NN	0.3405	99.31	0.4349	-0.0413	0.4666	-1.8822
SVM	0.6530	99.95	0.8685	0.0107	0.7742	0.8886
RF	**0.6745**	**100.00**	**0.9568**	**0.0165**	**0.8694**	**1.1561**

*Her bir MCDM için öncelikli tercih edilen sınıflandırma yöntemi koyu renk ile belirtilmiştir.

Çizelge 5.51'deki sonuçlara göre MCDM yöntemlerinin sıralaması Çizelge 5.52'de verilmiştir.

Çizelge 5.52 MCDM ile sınıflandırma yöntemlerinin sıralanması (Senaryo6)

	Yöntemler	***Sıralama***
Toplama dayalı	***TOPSIS***	RF ≫ LR ≫ SVM ≫ *k*NN
	COPRAS	RF ≫ LR ≫ SVM ≫ *k*NN
Orana dayalı	***EDAS***	RF ≫ LR ≫ SVM ≫ *k*NN
	CODAS	RF ≫ LR ≫ SVM ≫ *k*NN
Min-max değerine dayalı	***GRA***	RF ≫ LR ≫ SVM ≫ *k*NN
	MABAC	RF ≫ LR ≫ SVM ≫ *k*NN

Çizelge 5.52'ye göre Senaryo6 için RF öncelikli tercih edilir. Senaryo6, gerçek veri seti ile benzer örneklem çapı ve dağılıma sahiptir. Buna göre, örneklem çapı büyük ve dengesiz dağılıma sahip hem gerçek veri seti hem de simülasyon çalışmasında, RF öncelikli tercih edilir.

6. SONUÇ VE ÖNERİLER

Veri madenciliği ve karar destek sistemleri, veriden bilgi elde etme ve stratejik karar verme süreçlerinde kritik bir role sahiptir. Bilgi keşfinin en önemli aşaması olan veri ön işlemede, betimsel istatistikler ve görselleştirme ile veri setinin özetlenmesi, çok değişkenli istatistiksel yöntemler ile değişkenler arasındaki örtülü ilişkilerin ortaya çıkarılması ve değişkenlerin gruplaması yapılır. Veriden bilgi elde etmeye yönelik subjektif değerlendirmenin önemli olduğu durumlarda, MCDM yöntemi ile uzman görüşleri dikkate alınıp değişkenlerin ağırlıkları belirlenerek uzman bilgisinden yararlanılır. Uzman görüşünün elde edilemediği durumlarda ise objektif olarak veriye dayalı ağırlıklandırma yapan MCDM yöntemleri tercih edilir.

Veri madenciliği sınıflandırma yöntemleri, veriden anlamlı bilginin elde edilmesine imkân sağlar. Gelecek çalışmalar hakkında öngörü oluşturulması ve hedeflenen başarıya ulaşılması bakımından gözlemlerin doğru bir biçimde belirlenmesinde sınıflandırma çalışması oldukça önemlidir. Veri setine amacına uygun bir biçimde veri madenciliği sınıflandırma yöntemleri uygulanır. Sınıflandırma yöntemlerinin başarısının belirlenmesinde performans ölçütleri kullanılır. Literatürde bu konu ile ilgili pek çok çalışma bulunmaktadır. Performans ölçütlerine göre sınıflandırma yöntemlerinin seçiminde MCDM yöntemlerinin kullanılması daha güvenilir sonuçlar elde edilmesini sağlar. Buna göre, sınıflandırma yöntemlerinin performanslarının objektif olarak değerlendirilmesinde, sınıflandırma yöntemleri alternatifler, performans ölçütleri kriterler olacak biçimde bir karar verme problemine dönüştürülür. Bu problemin çözümü aşamasında MCDM yöntemlerinden yararlanılır.

Bu tez çalışmasında, subjektif değerlendirmeler dikkate alınarak en iyi performansı gösteren sınıflandırma yönteminin seçilmesinin hedeflendiği hibrit karar verme yaklaşımı ile ilgilenilmiştir. Bu amaçla, uzman görüşünün karar verme sürecine dahil edilmesine imkân sağlayan AHP ile değişken ağırlıkları elde edilmiştir. Literatürde kullanılan LR, *k*NN, SVM ve RF sınıflandırma yöntemleri, AHP ile ağırlıklandırılmış veri seti üzerinde uygulanmıştır. Çalışmada, ağırlıklandırılmış veri setinin sınıflandırılması sonucu elde edilen performans ölçütleri de bir karar verme problemi olarak ele alınmıştır. Hibrit karar

verme amacıyla, bu problem, farklı normalleştirme yaklaşımı kullanan MCDM yöntemleri ile çözülmüştür. Bu aşamada, toplama, orana ve min-max değerine dayalı normalleştirme yaklaşımı uygulayan MCDM yöntemleri kullanılmıştır. Gerçek veri seti ve simülasyon çalışmaları ile uygulamalar yapılmıştır.

Gerçek veri seti uygulamasında, Sanayi ve Teknoloji Bakanlığı'nın Ar-Ge Teşvikleri Genel Müdürlüğü'ne bağlı Ar-Ge ve Tasarım Merkezleri Dairesi Başkanlığı'ndan uygun biçimde temin edilen veri seti kullanılmıştır. Veri seti, 2008-2021 yılları arasında faaliyeti devam eden 2334 adet Ar-Ge ve Tasarım merkezlerini kapsamaktadır. Ar-Ge ve Tasarım merkezlerine ait veri seti titizlikle veri ön işleme aşamasından geçirilmiştir. Çok değişkenli istatistiksel yöntemlerden faktör analizi ile değişkenler arasındaki örtülü ilişki ortaya çıkarılarak birbiriyle ilişkili değişkenlerin gruplaması yapılmıştır. Veri setindeki değişkenler için Ar-Ge ve Tasarım Merkezleri Dairesi'nde çalışan uzmanların görüşleri alınarak AHP ile değişken ağırlıkları belirlenmiştir. Subjektif olarak ağırlıklandırılmış veri setine, LR, *k*NN, SVM ve RF sınıflandırma yöntemleri uygulanarak yöntemlerin performans ölçütlerinden doğruluk, kesinlik, duyarlılık, F_1-Skor ve AUC hesaplanmıştır. Elde edilen sınıflandırma performans değerleri bir karar matrisi olarak ele alınmıştır. Karar verme probleminin çözümünde, toplama dayalı normalleştirme yaklaşımı uygulayan TOPSIS ve COPRAS, orana dayalı normalleştirme yaklaşımı uygulayan EDAS ve CODAS, min-max değerine dayalı normalleştirme yaklaşımı uygulayan GRA ve MABAC yöntemleri uygulanmıştır. Uygulanan tüm MCDM yöntemlerinin sonuçlarına göre sınıflandırma yöntemlerinden RF'nin öncelikli tercih edilebileceği kararı elde edilmiştir. Ayrıca, objektif ağırlık belirleyen MCDM yöntemlerinden ENTROPY ve CRITIC ile belirlenen değişken ağırlıkları kullanılarak veri seti ağırlıklandırılmıştır. Objektif olarak ağırlıklandırılmış veri setine, subjektif olarak ağırlıklandırılmış veri setine benzer analizler uygulanmıştır. Buna göre, RF'nin öncelikli tercih edilebileceği kararı elde edilmiştir. Gerçek veri setinde subjektif ve objektif olarak belirlenen değişken ağırlığı ile yapılan sınıflandırma çalışmaları sonucunda, Ar-Ge ve Tasarım merkezlerinin faaliyetlerinin değerlendirilmesinde RF'nin LR, *k*NN ve SVM'ye göre daha iyi sınıflandırma performansı gösterdiği sonucuna ulaşılmıştır.

Gerçek veri seti ile elde edilen sınıflandırma performans sonuçlarının genelleştirilebilmesi için simülasyon çalışması yapılmıştır. Bu amaçla, simülasyon çalışması üretilmiş veri setlerinin oluşturulmasında olası farklı senaryolar kurgulanmıştır. Simülasyon çalışmasında, üç farklı örneklem çapı $n = \{150, 250, 2000\}$ ve p oranı $p = \{0.50, 0.80\}$ için altı senaryo kurgulanmıştır. p oranı, sınıflandırma çalışmasında ele alınan veri setinin dengeli (faaliyette ve faaliyette olmayan firma sayılarının yaklaşık eşit olduğu) veya dengesiz (faaliyette ve faaliyette olmayan firma sayılarının yaklaşık eşit olmadığı) dağılıma sahip bir veri seti olduğu belirtilmiştir. $Senaryo1, Senaryo2, \cdots, Senaryo6$ sınıflandırma yöntemlerinin algoritmaları, rastgele veri setinin oluşturulması ve performans ölçütlerinin hesaplanması bakımından 1000 tekrar yapılmıştır. Sınıflandırma yöntemleri bakımından elde edilen doğruluk, kesinlik, duyarlılık, F_1-Skor ve AUC performans değerleri bir karar matrisi olarak ele alınmıştır. Gerçek veri seti uygulamasına benzer olarak, TOPSIS, COPRAS, EDAS, CODAS, GRA ve MABAC uygulanmıştır Sonuçlar incelendiğinde, performans değerleri bakımından dengeli ve dengesiz veri setleri için RF'nin öncelikli tercih edildiği görülmüştür.

Bu tezin literatüre sağladığı katkılar ve özgün değeri açısından en önemli noktalarından birisi MCDM yönteminin veri ön işleme sürecinde uygulanmasıdır. MCDM yönteminin veri ön işlemede kullanılması, veriden bilgi elde etme sürecinde deneyimin ve subjektif değerlendirmelerin önemini vurgulamaktadır. Bu tez çalışmasında, AHP yöntemi ile uzman görüşlerinin karar verme sürecine dahil edilerek farklı sınıflandırma yöntemlerinde uygulanması incelenmiştir.

Tezin yenilikçi yönlerinden biri olan simülasyon çalışması ile yapılan uygulamalar, gerçek veri seti sonuçlarının karşılaştırmalı analizini sunarak özgün bir yaklaşım ortaya koymaktadır. Ayrıca, veri madenciliğinde sınıflandırma yöntemlerinin seçiminde ve performanslarının değerlendirilmesinde MCDM yöntemlerinin kullanılması karar verme aşamasını kolaylaştırmıştır.

Bu çalışma, ilk defa Ar-Ge ve Tasarım merkezlerinin faaliyetlerinin değerlendirilmesinde veri madenciliği yöntemleri ile hibrit karar verme yaklaşımının uygulanmasını ele

almaktadır. Hibrit yaklaşım, bilimsel ölçütlere göre yapılan değerlendirme çalışmaları ile hem kamu kaynağının etkin kullanılması hem de hedeflenen başarının sağlanması bakımından büyük önem taşımaktadır.

KAYNAKLAR

Aghdaie, M. H. and Alimardani, M. 2015. Target Market Selection Based on Market Segment Evaluation: A Multiple Attribute Decision Making Approach. International Journal of Operational Research, 24(3), 262-278.

Altunkaynak B. 2022. Veri Madenciliği Yöntemleri ve R Uygulamaları, Seçkin Yayıncılık, 304 s., Ankara.

Anonim. 2016. T. C. Resmî Gazete, Araştırma ve Geliştirme Faaliyetlerinin Desteklenmesi Hakkında Kanun ile Bazı Kanun Hükmünde Kararnamelerde Değişiklik Yapılmasına Dair Kanun, No: 29636. Ankara.

Arlı, N. B., Gürsakal S. ve Engin, M. 2022. Denetimli Makine Öğrenmesi Algoritmaları R ve Python Uygulamaları. Nobel Akademik Yayıncılık, 294 s., Ankara.

Arslan, S. ve Belgin, Ö. 2020. Yüksek ve Orta-Yüksek Teknoloji Alanındaki Sektörlerin Çok Kriterli Karar Verme Teknikleri ile Önceliklendirilmesi. Verimlilik Dergisi, 4, 7-23.

Atakan, C. 2022. Çok Değişkenli İstatistik Analiz Yöntemleri Dersi. Ankara Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Açık Ders Malzemeleri. https://acikders.ankara.edu.tr/course/view.php?id=3627&lang=en/Erişim Tarihi: 24.06.2024.

Ayhan, S. ve Erdoğmuş, Ş. 2014. Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü için Çekirdek Fonksiyonu Seçimi. Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi, 9(1), 175-201.

Başer, B. Ö., Yangın, M. ve Sarıdaş, E. S. 2021. Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 25 (1), 112-120.

Bilalli, B., Abelló, A., Aluja-Banet, T. and Wrembel, R. 2019. Presistant: Learning Based Assistant for Data Pre-processing. Data & Knowledge Engineering, 123, 101727.

Breiman, L. 2001. Random Forests. Machine Learning. 45, 5–32.

Burkov A. 2021. The Hundred-Page Machine Learning Book kitabından çeviri, Çeviren: Okatan A., Karatekin T. ve K. Okatan, K. 2021. 100 Sayfada Makine Öğrenmesi Kitabı, Papatya Yayıncılık Eğitim, 134 s., İstanbul.

Cebeci, C., Kamasak and R., Soyaltin, T.E. 2021. Team Perception and Learning Orientation as the Predictors of R&D Performance. Journal of Management, Marketing and Logistics (JMML), 8(4), 203-217.

Comrey, A. L. 1973. A First Courses in Factor Analysis. Academic Press, New York,

Cortes, C. and Vapnik, V. 1995. Support Vector Networks. Machine Learning, 20(3), 273-297.

Çelen, A. 2014. Comparative Analysis of Normalization Procedures in TOPSIS Method: With an Application to Turkish Deposit Banking Market, Informatica, 25(2), 185–208.

Çetin, B. and Kuvat, Ö. 2022. Ranking of Level 2 Regions in Terms of Economic Indicators in Turkey with Improved Entropy and Critic-Based Copras Method. Aksaray Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14(1), 11-36.

Çetin, V. and Yıldız, O. A. 2022. A Comprehensive Review on Data Preprocessing Techniques in Data Analysis. Pamukkale University Journal of Engineering Sciences, 28(2), 299-312.

Çınar, A. ve Silahtaroğlu, G. 2012. Veri Madenciliği Teknikleri ile Müşteri Memnuniyetine Etki Eden Gizli Nedenlerin Keşfi. Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi, 33(2), 309-330.

Çokluk, Ö. 2010. Lojistik Regresyon Analizi: Kavram ve Uygulama. Kuram ve Uygulamada Eğitim Bilimleri, 10(3), 1357-1407.

Demir, G., Özyalçın A.T. ve Bircan H. 2021. Çok Kriterli Karar Verme Yöntemleri ve ÇKKV Yazılımı ile Problem Çözümü. Nobel Yayıcılık, 392 s., Ankara.

Deng, J. 1982. Control Problems of Grey System. Systems and Control Letters, 5, 288-294.

Diakoulaki, D., Mavrotas, G. and Papayannakis, L. 1995. Determining Objective Weights in Multiple Criteria Problems: The Critic Method. Computers & Operations Research, 22(7), 763-770.

Dirican, E. 2019. Makine Öğrenimi Yöntemlerini Kullanarak Evre III İnvaziv Duktal Karsinomlu Hasta Verilerinin Sınıflandırılması. Doktora Tezi. Dicle Üniversitesi, Sağlık Bilimleri Enstitüsü, 98 s., Diyarbakır.

Ekinci, G. 2021. Gaziantep İli Ar-Ge Merkezleri ve İnovasyon Performans Göstergeleri. Adıyaman Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, (39), 395-430.

Erbaş, S. O. 2016. Olasılık ve İstatistik Problemler ve Çözümleri ile. Gazi Kitabevi, 637 s., Ankara.

Erden, C. 2021. Python ile Veri Madenciliği. Kodlab Yayınevi, 290 s., İstanbul.

Eşiyok, S., Ariş, E., ve Antmen, F. 2023. G7 Ülkeleri ve Türkiye'nin Entropy, CRITIC ve EDAS Yöntemleriyle GGGI Göstergelerine Göre Sıralaması ve

Değerlendirilmesi. Çukurova Üniversitesi Mühendislik Fakültesi Dergisi, 38(3), 647-660.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37-37.

Fletcher, T. 2009. Support Vector Machines Explained. Tutorial Paper, 1-19.

Frawley, W. J., Piatetsky-Shapiro, G. and Matheus, C. J. 1992. Knowledge Discovery in Databases: An Overview. AI Magazine, 13(3), 57-57.

Fu G. H., Yi L. Z. and Pan J. 2019. Tuning model parameters in class-imbalanced learning with precision-recall curve. Biometrical Journal, 61, 652-664.

Gao, H., Gajjar, S., Kulahci, M., Zhu, Q. and Palazoglu, A. 2016. Process Knowledge Discovery Using Sparse Principal Component Analysis. Industrial & Engineering Chemistry Research, 55(46), 12046-12059.

Geron, A. 2017. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc., 1005, 564.

Geron A. 2021. Scikit-Learn Keras ve TensorFlow ile Uygulamalı Makine Öğrenmesi. Çeviri. Buzdağı Yayınevi, 862 s., Ankara.

Ghorabaee, M.K, Zavadskas, E. K., Olfat, L. and Turskis, Z. 2015. Multi-Criteria Inventory Classification Using a New Method of Evaluation Based on Distance from Average Solution (EDAS). Informatica, 26(3), 435-451.

Ghorabaee, M.K., Zavadskas, E.K., Turskis, Z. and Antucheviciene, J. 2016. A New Combinative Distance-Based Assessment (CODAS) Method for Multi-Criteria Decision-Making. Economic Computation and Economic Cybernetics Studies and Research, 50, 25–44.

Ghosh, S., Saha, S. and Bera, B. 2022. Flood Susceptibility Zonation Using Advanced Ensemble Machine Learning Models within Himalayan Foreland Basin. Natural Hazards Research, 2(4), 363-374.

Gülyaz, S. and Ertürk, A. 2020. Investigation of the Relationship Between Knowledge Management and Innovation Capability: An Empirical Study on R&D Centers in Eastern Marmara Region. Research Journal of Business and Management (RJBM), 7(4), 322-335.

Harman, H. H. 1976. Modern Factor Analysis. University of Chicago Press, 486 p., Chicago.

Hossin M. and Sulaiman M. N. 2015. A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process, 5, 01-11.

Hwang, C. L. and Yoon, K. 1981. Multiple Attribute Decision Making: Methods and Applications. Springer- Verlag, 228 p., New York.

Joshi, M. V. 2002. On Evaluating Performance of Classifiers for Rare C lasses. IEEE International Conference on Data Mining. Proceedings, 641-644, IEEE.

Kabak, M. ve Çınar, Y. 2020. Çok Kriterli Karar Verme Yöntemleri: Ms Excel Çözümlü Uygulamalar. Nobel Akademik Yayıncılık, 127 s., Ankara.

Karagöz, Y. 2019. SPSS-AMOS-META Uygulamalı İstatistiksel Analizler. Nobel Akademik Yayıncılık, 1316 s., Ankara.

Keleş, K. ve Tunca, P. Z. 2015. Hiyerarşik Electre Yönteminin Teknokent Seçiminde Kullanımı Üzerine Bir Çalışma. Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 20 (1), 199-223.

Korkmaz, D., Çelik, H. E. ve Kapar, M. 2018. Sınıflandırma ve Regresyon Ağaçları ile Rastgele Orman Algoritması Kullanarak Botnet Tespiti: Van Yüzüncü Yıl Üniversitesi Örneği. Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 23(3), 297-307.

Kumar, A., Singh, S. S., Singh, K., and Biswas, B. 2020. Link Prediction Techniques, Applications, and Performance: A Survey. Physica A: Statistical Mechanics and its Applications, 553, 124289.

Li, T. and Ruan, D. 2007. An Extended Process Model of Knowledge Discovery in Database. Journal of Enterprise Information Management, 20(2), 169-177.

Mete, M. H. ve Dağdeviren, M. 2017. Ar-Ge Merkezleri için Bilgi Yönetimi Modellemesi ve Bilgi Yönetiminin Ar-Ge Performansı ile İlişkisi. Verimlilik Dergisi, 2, 75-108.

Metlek, S. ve Kayaalp, K. 2020. Makine Öğrenmesinde, Teoriden Örnek MATLAB Uygulamalarına Kadar Destek Vektör Makineleri. İksad Yayınevi, 98 s., Ankara.

Oğuzlar, A. 2003. Veri Önişleme. Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 21, 67-76.

Osmanoğlu, U. Ö. 2021. Makine Öğrenmesi Sınıflama Algoritmalarıyla Kalp Yetersizliği Mortalitesinin Tahminlenmesi. Doktora Tezi. Eskişehir Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü, 58 s., Eskişehir.

Özbek, A. 2017. Çok Kriterli Karar Verme Yöntemleri ve Excel ile Problem Çözümü. Seçkin Yayıncılık, 382 s., Ankara.

Özçalıcı, M. 2017. Matlab ile Çok Kriterli Karar Verme Teknikleri. Nobel Akademik Yayıncılık, 566 s., Ankara.

Özkan, Y. 2016. Veri madenciliği yöntemleri. Papatya Yayıncılık Eğitim, 240 s., İstanbul.

Pamučar, D. and Ćirović, G. 2015. The Selection of Transport and Handling Resources in Logistics Centers Using Multi-Attributive Border Approximation Area Comparison (MABAC). Expert Systems with Applications, 42 (6), 3016-3028.

Patel, J., Shah, S., Thakkar, P. and Kotecha, K. 2015. Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques. Expert Systems with Applications, 42(1), 259-268.

Pehlivanoğlu, Y. V. 2017. Optimizasyon: Temel Kavramlar & Yöntemler. 740 s., Ankara.

Rafiei-Sardooi, E., Azareh, A., Choubin, B., Mosavi, A. H. and Clague, J. J. 2021. Evaluating Urban Flood Risk Using Hybrid Method of TOPSIS and Machine Learning. International Journal of Disaster Risk Reduction, 66, 102614.

Refaeilzadeh, P., Tang, L. and Liu, H. 2009. Cross-validation. Encyclopedia of database systems, 5, 532-538.

Romero, C., Romero, J. and Ventura, S. 2013. A Survey on Pre-processing Educational Data. Educational Data Mining: Applications and Trends, 29-64.

Saaty, T. L. 1977. A Scaling Method for Priorities in Hierarchical Structures. Journal of Mathematical Psychology, 15, 234-281.

Salman, S. ve Peker, İ. 2021. Savunma Sanayi Ar-Ge Merkezlerinin Performanslarının ENTROPY ve ARAS Yöntemleri ile Değerlendirilmesi. Uluslararası Ekonomi ve Yenilik Dergisi, 7 (1), 51-73.

Shannon, C. E. 1948. A Mathematical Theory of Communication. The Bell System Technical Journal, 27(3), 379-423.

Sıkı, N. ve Acartürk, F. 2019. Türkiye'deki İlaç Ar-Ge Merkezlerinin Faaliyetlerinin Patent ve Yenilik Açısından Değerlendirilmesi. Journal of Faculty of Pharmacy of Ankara University, 43(1), 44-63.

Sıtkı, Y. H. 2020. Tıp Bilişiminde Veri Madenciliği Yöntemleri Kullanılarak Hastalıkların Tahmin Edilmesi. Yüksek Lisans Tezi. Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, 86 s., Ankara.

Silahtaroğlu, G. 2008. Veri madenciliği. Papatya Yayıncılık Eğitim, 300 s., İstanbul.

Sorhun, E. 2021. Python ile Makine Öğrenmesi. 464 s., Abaküs Kitap, İstanbul.

Soyal, B., Soysal, M. ve Ömürgönülşen, M. 2022. Ar-Ge Projelerindeki Kritik Başarı Faktörlerinin Algılanan Proje Performansı Üzerindeki Etkileri: Havacılık Sektöründe Bir Araştırma. Verimlilik Dergisi, 2, 147-164.

Spearman, C. 1904. General Intelligence, Objectively Determined and Measured. American Journal of Psychology. 15, 201-293.

Tabachnick, B. G., Fidell, L. S. and Ullman, J. B. 2007. Using Multivariate Statistics. 5, 481-498. Boston, MA: Pearson.

Tatlıdil, H. 2002. Uygulamalı Çok Değişkenli İstatistik Analiz. Ziraat Matbaacılık A.Ş., 424 s., Ankara.

Torkayesh, A. E., Tirkolaee, E. B., Bahrini, A., Pamucar, D. and Khakbaz, A. 2023. A Systematic Literature Review of MABAC Method and Applications: An Outlook for Sustainability and Circularity. Informatica, 34(2), 415-448.

Türkşen, Ö. 2023. Optimizasyon Yöntemleri ve Matlab, Python, R Uygulamaları. Nobel Akademik Yayıncılık, 468 s., Ankara.

Uğuz, S. 2021. Makine Öğrenmesi Teorik Yönleri ve Python Uygulamaları ile Yapay Zekâ Ekolü. Nobel Akademik Yayıncılık, 298 s., Ankara.

Vafaei, N., Ribeiro, R. A. and Camarinha-Matos, L. M. 2022. Assessing Normalization Techniques for Simple Additive Weighting Method. Procedia Computer Science, 199, 1229-1236.

Xie, S. and Zhang, J. 2023. TOPSIS-based Comprehensive Measure of Variable Importance in Predictive Modelling. Expert Systems with Applications, 120682.

Yeşildal, E. S. ve Kamasak, R. 2021. Bilgi İşleme Yeteneklerinin İnovasyon Performansına Etkisi: Ar-Ge Merkezleri Üzerine Ampirik Bir Araştırma. Research Journal of Business and Management, 8(4), 291-299.

Yıldırım, B. F. ve Önder, E. 2018. Operasyonel, Yönetsel ve Stratejik Problemlerin Çözümünde Çok Kriterli Karar Verme Yöntemleri. Dora Basım Yayın, 339 s., Bursa.

Yıldız, A. 2014. İş Zekası ve Veri Madenciliğine Dayalı Karar Verme. Gazi Kitapevi, 118 s., Ankara.

Yong, A. G. and Pearce, S. 2013. A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. Tutorials in Quantitative Methods for Psychology, 9(2), 79-94.

Yu, X., Shi, Y., Zhang, L., Nie, G. and Huang, A. 2014. Intelligent Knowledge Beyond Data Mining: Influences of Habitual Domains. Communications of the Association for Information Systems, 985-1000.

Zafar, S., Alamgir, Z. and Rehman, M.H. 2021. An Effective Blockchain Evaluation System Based on ENTROPY-CRITIC Weight Method and MCDM Techniques. Peer-to-Peer Networking and Applications. 14(2), 3110–3123.

Zan, Ç. Ö. 2021. Prediction of Soil Radon Gas Using Meteorological Parameters with Machine Learning Algorithms. Master Thesis. Dokuz Eylül University, Graduate School of Natural and Applied Sciences, 78 p., İzmir.

Zavadskas, E. K. and Kaklauskas, A. 1996. Systemotechnical Evaluation of Buildings (Pastatų sistemotechninis įvertinimas). Vilnius: Technika, 280 p. (in Lithuanian).

Zengin, G. 2022. Finansal Teknoloji Alanında Kullanıcı Deneyimlerinin Makine Öğrenmesi Yöntemleri ile İncelenmesi. Yüksek Lisans Tezi. Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, 117 s., İstanbul.

EKLER

EK 1 Uzman Görüşlerinin Karşılaştırma Matrisi

$$U_1 = \begin{bmatrix} 1.00 & 5.00 & 5.00 & 7.00 & 7.00 & 5.00 & 5.00 & 5.00 & 1.00 & 7.00 & 1.00 & 3.00 \\ 0.20 & 1.00 & 0.33 & 0.20 & 0.20 & 0.20 & 0.20 & 3.00 & 3.00 & 0.20 & 3.00 & 3.00 \\ 0.20 & 0.14 & 1.00 & 0.20 & 0.20 & 0.20 & 0.20 & 3.00 & 1.00 & 0.20 & 5.00 & 3.00 \\ 0.14 & 5.00 & 5.00 & 1.00 & 0.14 & 0.14 & 0.14 & 3.00 & 5.00 & 0.20 & 5.00 & 3.00 \\ 0.14 & 5.00 & 5.00 & 7.00 & 1.00 & 5.00 & 5.00 & 5.00 & 5.00 & 5.00 & 5.00 & 5.00 \\ 0.20 & 5.00 & 5.00 & 7.00 & 0.20 & 1.00 & 7.00 & 7.00 & 7.00 & 5.00 & 5.00 & 5.00 \\ 0.20 & 5.00 & 5.00 & 7.00 & 0.20 & 0.14 & 1.00 & 5.00 & 5.00 & 5.00 & 3.00 & 3.00 \\ 0.20 & 0.33 & 0.33 & 0.33 & 0.20 & 0.14 & 0.20 & 1.00 & 3.00 & 0.14 & 3.00 & 3.00 \\ 1.00 & 0.33 & 1.00 & 0.20 & 0.20 & 0.14 & 0.20 & 0.33 & 1.00 & 0.14 & 3.00 & 3.00 \\ 0.14 & 5.00 & 5.00 & 5.00 & 0.20 & 0.20 & 0.20 & 7.00 & 7.00 & 1.00 & 0.20 & 5.00 \\ 1.00 & 0.33 & 0.20 & 0.20 & 0.20 & 0.33 & 0.33 & 0.33 & 0.33 & 0.20 & 1.00 & 3.00 \\ 0.33 & 0.33 & 0.33 & 0.33 & 0.20 & 0.33 & 0.33 & 0.33 & 0.33 & 0.20 & 0.33 & 1.00 \end{bmatrix}$$

$$U_2 = \begin{bmatrix} 1.00 & 5.00 & 5,00 & 5.00 & 5.00 & 3.00 & 3.00 & 7.00 & 1.00 & 0.20 & 7.00 & 5.00 \\ 0.20 & 1.00 & 5,00 & 0.20 & 0.20 & 0.14 & 0.14 & 0.20 & 0.20 & 0.20 & 0.14 & 0.20 \\ 0.20 & 0.20 & 1.00 & 0.20 & 0.20 & 0.14 & 0.14 & 0.20 & 0.20 & 0.20 & 0.14 & 0.20 \\ 0.20 & 5.00 & 5.00 & 1.00 & 9.00 & 0.11 & 0.11 & 1.00 & 0.20 & 0.14 & 0.14 & 0.20 \\ 0.20 & 5.00 & 5.00 & 0.11 & 1.00 & 0.11 & 0.11 & 1.00 & 0.20 & 0.14 & 0.14 & 0.20 \\ 0.33 & 7.00 & 7.00 & 9.00 & 9.00 & 1.00 & 9.00 & 3.00 & 0.20 & 0.14 & 0.14 & 0.20 \\ 0.33 & 7.00 & 7.00 & 9.00 & 9.00 & 0.11 & 1.00 & 3.00 & 0.20 & 0.14 & 0.14 & 0.20 \\ 0.14 & 5.00 & 5.00 & 1.00 & 1.00 & 0.33 & 0.33 & 1.00 & 0.20 & 5.00 & 0.33 & 0.33 \\ 1.00 & 5.00 & 5.00 & 5.00 & 5.00 & 5.00 & 5.00 & 5.00 & 1.00 & 0.20 & 0.20 & 0.20 \\ 5.00 & 5.00 & 5.00 & 7.00 & 7.00 & 7.00 & 7.00 & 0.20 & 5.00 & 1.00 & 0.14 & 0.14 \\ 0.14 & 7.00 & 7.00 & 7.00 & 7.00 & 7.00 & 7.00 & 3.00 & 5.00 & 7.00 & 1.00 & 5.00 \\ 0.20 & 5.00 & 5.00 & 5.00 & 5.00 & 5.00 & 5.00 & 3.00 & 5.00 & 7.00 & 0.20 & 1.00 \end{bmatrix}$$

$$U_3 = \begin{bmatrix} 1.00 & 0.14 & 0.14 & 3.00 & 5,00 & 0.14 & 0.11 & 7.00 & 7.00 & 9.00 & 7.00 & 7.00 \\ 7.00 & 1.00 & 0.20 & 9.00 & 9,00 & 5.00 & 5.00 & 7.00 & 7.00 & 5.00 & 5.00 & 5.00 \\ 7.00 & 5.00 & 1.00 & 7.00 & 7,00 & 5.00 & 5.00 & 7.00 & 7.00 & 5.00 & 5.00 & 5.00 \\ 0.33 & 0.11 & 0.14 & 1.00 & 3,00 & 1.00 & 0.20 & 3.00 & 3.00 & 3.00 & 3.00 & 0.33 \\ 0.20 & 0.11 & 0.14 & 0.33 & 1,00 & 0.20 & 1.00 & 3.00 & 3.00 & 3.00 & 3.00 & 0.33 \\ 7.00 & 0.20 & 0.20 & 1.00 & 5,00 & 1.00 & 7.00 & 7.00 & 7.00 & 7.00 & 5.00 & 5.00 \\ 9.00 & 0.20 & 0.20 & 5.00 & 1,00 & 0.14 & 1.00 & 7.00 & 7.00 & 7.00 & 5.00 & 5.00 \\ 0.14 & 0.14 & 0.14 & 0.33 & 0,33 & 0.14 & 0.14 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \\ 0.14 & 0.14 & 0.14 & 0.33 & 0,33 & 0.14 & 0.14 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \\ 0.11 & 0.20 & 0.20 & 0.33 & 0,33 & 0.14 & 0.14 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \\ 0.14 & 0.20 & 0.20 & 0.33 & 0,33 & 0.20 & 0.20 & 1.00 & 1.00 & 1.00 & 1.00 & 0.33 \\ 0.14 & 0.20 & 0.20 & 3.00 & 3,00 & 0.20 & 0.20 & 1.00 & 1.00 & 1.00 & 3.00 & 1.00 \end{bmatrix}$$

$$U_4 = \begin{bmatrix}
1.00 & 0.33 & 0.20 & 5.00 & 3.00 & 0.20 & 0.33 & 9.00 & 0.14 & 5.00 & 9.00 & 3.00 \\
3.00 & 1.00 & 1.00 & 3.00 & 1.00 & 5.00 & 3.00 & 9.00 & 3.00 & 5.00 & 9.00 & 5.00 \\
5.00 & 1.00 & 1.00 & 3.00 & 1.00 & 5.00 & 3.00 & 9.00 & 3.00 & 5.00 & 9.00 & 5.00 \\
0.20 & 0.33 & 0.33 & 1.00 & 1.00 & 0.11 & 0.14 & 5.00 & 0.14 & 5.00 & 9.00 & 3.00 \\
0.33 & 1.00 & 1.00 & 1.00 & 1.00 & 0.11 & 0.14 & 1.00 & 0.14 & 5.00 & 9.00 & 3.00 \\
5.00 & 0.20 & 0.20 & 9.00 & 9.00 & 1.00 & 7.00 & 9.00 & 5.00 & 5.00 & 9.00 & 5.00 \\
3.00 & 0.33 & 0.33 & 7.00 & 7.00 & 0.14 & 1.00 & 7.00 & 0.20 & 5.00 & 9.00 & 5.00 \\
0.11 & 0.11 & 0.11 & 0.20 & 1.00 & 0.11 & 0.14 & 1.00 & 0.11 & 5.00 & 9.00 & 0.33 \\
7.00 & 0.33 & 0.33 & 7.00 & 7.00 & 0.20 & 5.00 & 9.00 & 1.00 & 7.00 & 9.00 & 5.00 \\
0.20 & 0.20 & 0.20 & 0.20 & 0.20 & 0.20 & 0.20 & 0.20 & 0.14 & 1.00 & 9.00 & 5.00 \\
0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 1.00 & 9.00 \\
0.33 & 0.33 & 0.33 & 0.33 & 0.33 & 0.20 & 0.20 & 3.00 & 0.20 & 0.20 & 0.11 & 1.00
\end{bmatrix}$$

$$U_5 = \begin{bmatrix}
1.00 & 5.00 & 0.20 & 0.20 & 0.14 & 0.20 & 0.20 & 5.00 & 0.20 & 0.14 & 0.11 & 0.20 \\
0.20 & 1.00 & 0.20 & 0.14 & 0.20 & 1.00 & 0.14 & 5.00 & 0.14 & 0.14 & 0.20 & 0.14 \\
5.00 & 5.00 & 1.00 & 0.14 & 0.20 & 0.20 & 0.20 & 5.00 & 0.20 & 0.14 & 0.20 & 0.14 \\
5.00 & 7.00 & 7.00 & 1.00 & 0.20 & 0.20 & 0.14 & 5.00 & 0.20 & 0.14 & 0.20 & 0.20 \\
7.00 & 5.00 & 5.00 & 5.00 & 1.00 & 0.14 & 0.14 & 5.00 & 0.20 & 0.14 & 5.00 & 0.20 \\
5.00 & 1.00 & 5.00 & 5.00 & 7.00 & 1.00 & 0.20 & 5.00 & 1.00 & 0.20 & 5.00 & 5.00 \\
5.00 & 7.00 & 5.00 & 7.00 & 7.00 & 5.00 & 1.00 & 5.00 & 1.00 & 0.20 & 5.00 & 0.14 \\
0.20 & 0.20 & 0.20 & 0.20 & 0.20 & 0.20 & 0.20 & 1.00 & 0.20 & 0.20 & 0.20 & 5.00 \\
5.00 & 5.00 & 5.00 & 5.00 & 5.00 & 1.00 & 1.00 & 5.00 & 1.00 & 0.20 & 5.00 & 5.00 \\
7.00 & 7.00 & 7.00 & 7.00 & 7.00 & 5.00 & 5.00 & 5.00 & 5.00 & 1.00 & 0.14 & 0.14 \\
5.00 & 5.00 & 5.00 & 5.00 & 0.20 & 0.20 & 0.20 & 5.00 & 0.20 & 7.00 & 1.00 & 5.00 \\
5.00 & 7.00 & 5.00 & 5.00 & 5.00 & 0.20 & 7.00 & 0.20 & 0.20 & 7.00 & 0.20 & 1.00
\end{bmatrix}$$

$$U_6 = \begin{bmatrix}
1.00 & 1.00 & 1.00 & 0.20 & 0.14 & 0.20 & 0.33 & 9.00 & 5.00 & 0.33 & 5.00 & 3.00 \\
1.00 & 1.00 & 0.33 & 5.00 & 7.00 & 0.20 & 0.33 & 7.00 & 5.00 & 0.14 & 7.00 & 0.33 \\
1.00 & 3.00 & 1.00 & 0.20 & 0.14 & 0.33 & 0.33 & 7.00 & 5.00 & 0.14 & 5.00 & 0.33 \\
5.00 & 0.20 & 5.00 & 1.00 & 0.33 & 3.00 & 5.00 & 9.00 & 7.00 & 1.00 & 9.00 & 7.00 \\
7.00 & 0.14 & 7.00 & 3.00 & 1.00 & 3.00 & 7.00 & 9.00 & 9.00 & 3.00 & 9.00 & 5.00 \\
5.00 & 5.00 & 3.00 & 0.33 & 0.33 & 1.00 & 5.00 & 7.00 & 7.00 & 3.00 & 5.00 & 3.00 \\
3.00 & 3.00 & 3.00 & 0.20 & 0.14 & 0.20 & 1.00 & 5.00 & 5.00 & 1.00 & 5.00 & 1.00 \\
0.11 & 0.14 & 0.14 & 0.11 & 0.11 & 0.14 & 0.20 & 1.00 & 0.20 & 0.11 & 0.33 & 0.14 \\
0.20 & 0.20 & 0.20 & 0.14 & 0.11 & 0.14 & 0.20 & 5.00 & 1.00 & 0.33 & 5.00 & 0.33 \\
3.00 & 7.00 & 7.00 & 1.00 & 0.33 & 0.33 & 1.00 & 9.00 & 3.00 & 1.00 & 3.00 & 1.00 \\
0.20 & 0.14 & 0.20 & 0.11 & 0.11 & 0.20 & 0.20 & 3.00 & 0.20 & 0.33 & 1.00 & 0.14 \\
0.33 & 3.00 & 3.00 & 0.14 & 0.20 & 0.33 & 1.00 & 7.00 & 3.00 & 1.00 & 7.00 & 1.00
\end{bmatrix}$$

$$U_7 = \begin{bmatrix} 1.00 & 5.00 & 3.00 & 0.33 & 0.20 & 0.14 & 0.20 & 9.00 & 9.00 & 9.00 & 9.00 & 5.00 \\ 0.20 & 1.00 & 3.00 & 0.33 & 0.20 & 0.14 & 0.20 & 9.00 & 9.00 & 9.00 & 9.00 & 5.00 \\ 0.33 & 0.33 & 1.00 & 0.33 & 0.20 & 0.14 & 0.20 & 9.00 & 9.00 & 9.00 & 9.00 & 5.00 \\ 3.00 & 3.00 & 3.00 & 1.00 & 0.33 & 0.20 & 0.33 & 9.00 & 9.00 & 9.00 & 9.00 & 5.00 \\ 5.00 & 5.00 & 5.00 & 3.00 & 1.00 & 0.14 & 0.20 & 9.00 & 9.00 & 9.00 & 9.00 & 5.00 \\ 7.00 & 7.00 & 7.00 & 5.00 & 7.00 & 1.00 & 5.00 & 9.00 & 9.00 & 9.00 & 9.00 & 9.00 \\ 5.00 & 5.00 & 5.00 & 3.00 & 5.00 & 0.20 & 1.00 & 9.00 & 9.00 & 9.00 & 9.00 & 9.00 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 1.00 & 1.00 & 1.00 & 1.00 & 0.33 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 1.00 & 1.00 & 1.00 & 1.00 & 0.33 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 1.00 & 1.00 & 1.00 & 1.00 & 0.33 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 1.00 & 1.00 & 1.00 & 1.00 & 0.33 \\ 0.20 & 0.20 & 0.20 & 0.20 & 0.20 & 0.11 & 0.11 & 3.00 & 3.00 & 3.00 & 3.00 & 1.00 \end{bmatrix}$$

$$U_8 = \begin{bmatrix} 1.00 & 7.00 & 5.00 & 3.00 & 3.00 & 3.00 & 3.00 & 9.00 & 7.00 & 7.00 & 9.00 & 5.00 \\ 0.14 & 1.00 & 0.33 & 3.00 & 1.00 & 0.33 & 3.00 & 5.00 & 3.00 & 5.00 & 7.00 & 3.00 \\ 0.20 & 3.00 & 1.00 & 5.00 & 3.00 & 3.00 & 5.00 & 3.00 & 3.00 & 5.00 & 7.00 & 3.00 \\ 0.33 & 0.33 & 0.20 & 1.00 & 0.20 & 0.33 & 1.00 & 5.00 & 3.00 & 3.00 & 5.00 & 3.00 \\ 0.33 & 1.00 & 0.33 & 5.00 & 1.00 & 3.00 & 3.00 & 7.00 & 7.00 & 7.00 & 7.00 & 5.00 \\ 0.33 & 3.00 & 0.33 & 3.00 & 0.33 & 1.00 & 7.00 & 7.00 & 5.00 & 7.00 & 7.00 & 3.00 \\ 0.33 & 0.33 & 0.20 & 1.00 & 0.33 & 0.14 & 1.00 & 7.00 & 3.00 & 1.00 & 3.00 & 1.00 \\ 0.11 & 0.20 & 0.33 & 0.20 & 0.14 & 0.14 & 0.14 & 1.00 & 1.00 & 3.00 & 3.00 & 1.00 \\ 0.14 & 0.33 & 0.33 & 0.33 & 0.14 & 0.20 & 0.33 & 1.00 & 1.00 & 3.00 & 3.00 & 1.00 \\ 0.14 & 0.20 & 0.20 & 0.33 & 0.14 & 0.14 & 1.00 & 0.33 & 0.33 & 1.00 & 0.33 & 0.20 \\ 0.11 & 0.14 & 0.14 & 0.20 & 0.14 & 0.14 & 0.33 & 0.33 & 0.33 & 3.00 & 1.00 & 0.33 \\ 0.20 & 0.33 & 0.33 & 0.33 & 0.20 & 0.33 & 1.00 & 1.00 & 1.00 & 5.00 & 3.00 & 1.00 \end{bmatrix}$$

$$U_9 = \begin{bmatrix} 1.00 & 7.00 & 7.00 & 5.00 & 5.00 & 7.00 & 7.00 & 0.20 & 0.14 & 5.00 & 0.14 & 5.00 \\ 0.14 & 1.00 & 5.00 & 7.00 & 7.00 & 5.00 & 5.00 & 0.33 & 0.20 & 7.00 & 3.00 & 5.00 \\ 0.14 & 0.20 & 1.00 & 5.00 & 5.00 & 3.00 & 3.00 & 0.14 & 0.20 & 7.00 & 0.33 & 0.33 \\ 0.20 & 0.14 & 0.20 & 1.00 & 3.00 & 0.20 & 0.20 & 0.14 & 0.20 & 5.00 & 0.20 & 0.20 \\ 0.20 & 0.14 & 0.20 & 0.33 & 1.00 & 7.00 & 5.00 & 0.33 & 0.33 & 5.00 & 0.20 & 0.33 \\ 0.14 & 0.20 & 0.33 & 5.00 & 0.14 & 1.00 & 0.20 & 0.14 & 0.14 & 7.00 & 0.20 & 0.20 \\ 0.14 & 0.20 & 0.33 & 5.00 & 0.20 & 5.00 & 1.00 & 0.20 & 0.11 & 5.00 & 0.33 & 0.20 \\ 5.00 & 3.00 & 7.00 & 7.00 & 3.00 & 7.00 & 5.00 & 1.00 & 3.00 & 7.00 & 5.00 & 5.00 \\ 7.00 & 5.00 & 5.00 & 5.00 & 3.00 & 7.00 & 9.00 & 0.33 & 1.00 & 5.00 & 7.00 & 7.00 \\ 0.20 & 0.14 & 0.14 & 0.20 & 0.20 & 0.14 & 0.20 & 0.14 & 0.20 & 1.00 & 0.20 & 0.33 \\ 7.00 & 0.33 & 3.00 & 5.00 & 5.00 & 5.00 & 3.00 & 0.20 & 0.14 & 5.00 & 1.00 & 5.00 \\ 0.20 & 0.20 & 3.00 & 5.00 & 3.00 & 5.00 & 5.00 & 0.20 & 0.14 & 3.00 & 0.20 & 1.00 \end{bmatrix}$$

$$U_{10} = \begin{bmatrix}
1.00 & 3.00 & 1.00 & 5.00 & 5.00 & 0.33 & 1.00 & 9.00 & 9.00 & 7.00 & 7.00 & 5.00 \\
0.33 & 1.00 & 0.33 & 3.00 & 3.00 & 0.20 & 0.33 & 7.00 & 7.00 & 5.00 & 5.00 & 3.00 \\
1.00 & 3.00 & 1.00 & 5.00 & 5.00 & 0.33 & 1.00 & 9.00 & 9.00 & 7.00 & 7.00 & 5.00 \\
0.20 & 0.33 & 0.20 & 1.00 & 3.00 & 0.20 & 0.14 & 7.00 & 5.00 & 5.00 & 5.00 & 3.00 \\
0.20 & 0.33 & 0.20 & 0.33 & 1.00 & 0.20 & 0.14 & 7.00 & 5.00 & 5.00 & 5.00 & 3.00 \\
3.00 & 5.00 & 3.00 & 5.00 & 5.00 & 1.00 & 3.00 & 9.00 & 9.00 & 7.00 & 7.00 & 5.00 \\
1.00 & 3.00 & 1.00 & 7.00 & 7.00 & 0.33 & 1.00 & 9.00 & 9.00 & 7.00 & 7.00 & 5.00 \\
0.11 & 0.14 & 0.11 & 0.14 & 0.14 & 0.11 & 0.11 & 1.00 & 1.00 & 0.33 & 0.33 & 0.33 \\
0.11 & 0.14 & 0.11 & 0.20 & 0.20 & 0.11 & 0.11 & 1.00 & 1.00 & 0.33 & 0.33 & 0.33 \\
0.14 & 0.20 & 0.14 & 0.20 & 0.20 & 0.14 & 0.14 & 3.00 & 3.00 & 1.00 & 1.00 & 0.33 \\
0.14 & 0.20 & 0.14 & 0.20 & 0.20 & 0.29 & 0.14 & 3.00 & 3.00 & 3.00 & 1.00 & 1.00 \\
0.20 & 0.33 & 0.20 & 0.33 & 0.33 & 0.40 & 0.20 & 3.00 & 3.00 & 3.00 & 1.00 & 1.00
\end{bmatrix}$$

$$U_{11} = \begin{bmatrix}
1.00 & 0.20 & 0.33 & 0.33 & 0.20 & 0.20 & 0.33 & 3.00 & 3.00 & 3.00 & 5.00 & 3.00 \\
5.00 & 1.00 & 0.33 & 0.33 & 0.11 & 0.20 & 0.33 & 3.00 & 3.00 & 3.00 & 7.00 & 5.00 \\
3.00 & 3.00 & 1.00 & 0.33 & 0.14 & 3.00 & 5.00 & 3.00 & 3.00 & 3.00 & 7.00 & 7.00 \\
5.00 & 3.00 & 3.00 & 1.00 & 0.11 & 1.00 & 3.00 & 3.00 & 3.00 & 5.00 & 7.00 & 7.00 \\
5.00 & 9.00 & 7.00 & 9.00 & 1.00 & 9.00 & 9.00 & 9.00 & 9.00 & 9.00 & 9.00 & 9.00 \\
5.00 & 5.00 & 0.33 & 1.00 & 0.11 & 1.00 & 9.00 & 7.00 & 7.00 & 9.00 & 9.00 & 9.00 \\
3.00 & 3.00 & 0.20 & 0.33 & 0.11 & 0.11 & 1.00 & 7.00 & 5.00 & 7.00 & 9.00 & 9.00 \\
0.33 & 0.33 & 0.33 & 0.33 & 0.11 & 0.14 & 0.14 & 1.00 & 0.11 & 0.33 & 1.00 & 1.00 \\
0.33 & 0.33 & 0.33 & 0.33 & 0.11 & 0.14 & 0.20 & 9.00 & 1.00 & 5.00 & 7.00 & 9.00 \\
0.33 & 0.33 & 0.33 & 0.20 & 0.11 & 0.11 & 0.14 & 3.00 & 0.20 & 1.00 & 5.00 & 7.00 \\
0.20 & 0.14 & 0.14 & 0.14 & 0.11 & 0.11 & 0.11 & 1.00 & 0.14 & 0.20 & 1.00 & 5.00 \\
0.33 & 0.20 & 0.14 & 0.14 & 0.11 & 0.11 & 0.11 & 1.00 & 0.11 & 0.14 & 0.20 & 1.00
\end{bmatrix}$$

$$U_{12} = \begin{bmatrix}
1.00 & 5.00 & 3.00 & 0.33 & 0.20 & 0.33 & 0.33 & 9.00 & 7.00 & 0.33 & 3.00 & 0.20 \\
0.20 & 1.00 & 1.00 & 0.20 & 0.14 & 0.33 & 0.33 & 7.00 & 7.00 & 0.20 & 0.33 & 0.33 \\
0.33 & 1.00 & 1.00 & 0.33 & 0.20 & 0.33 & 0.33 & 7.00 & 7.00 & 0.33 & 0.33 & 0.20 \\
3.00 & 5.00 & 3.00 & 1.00 & 0.20 & 0.33 & 1.00 & 7.00 & 7.00 & 0.33 & 3.00 & 0.33 \\
5.00 & 7.00 & 5.00 & 5.00 & 1.00 & 3.00 & 3.00 & 9.00 & 9.00 & 1.00 & 5.00 & 3.00 \\
3.00 & 3.00 & 3.00 & 3.00 & 0.33 & 1.00 & 9.00 & 9.00 & 9.00 & 5.00 & 5.00 & 3.00 \\
3.00 & 3.00 & 3.00 & 1.00 & 0.33 & 0.11 & 1.00 & 9.00 & 9.00 & 1.00 & 3.00 & 1.00 \\
0.11 & 0.14 & 0.14 & 0.14 & 0.11 & 0.11 & 0.11 & 1.00 & 1.00 & 0.11 & 0.11 & 0.11 \\
0.14 & 0.14 & 0.14 & 0.14 & 0.11 & 0.11 & 0.11 & 1.00 & 1.00 & 0.11 & 0.11 & 0.11 \\
3.00 & 5.00 & 3.00 & 3.00 & 0.11 & 0.20 & 1.00 & 9.00 & 9.00 & 1.00 & 3.00 & 0.33 \\
0.33 & 3.00 & 3.00 & 0.33 & 0.20 & 0.20 & 0.33 & 9.00 & 9.00 & 0.33 & 1.00 & 0.20 \\
5.00 & 3.00 & 5.00 & 3.00 & 0.20 & 0.33 & 1.00 & 9.00 & 9.00 & 3.00 & 5.00 & 1.00
\end{bmatrix}$$

$$U_{13} = \begin{bmatrix} 1.00 & 0.14 & 0.14 & 3.00 & 3.00 & 0.14 & 0.14 & 9.00 & 7.00 & 9.00 & 9.00 & 9.00 \\ 7.00 & 1.00 & 1.00 & 7.00 & 7.00 & 7.00 & 7.00 & 9.00 & 7.00 & 9.00 & 9.00 & 9.00 \\ 7.00 & 0.14 & 1.00 & 7.00 & 7.00 & 7.00 & 7.00 & 9.00 & 7.00 & 9.00 & 9.00 & 9.00 \\ 0.33 & 0.14 & 0.14 & 1.00 & 1.00 & 0.14 & 0.14 & 9.00 & 7.00 & 9.00 & 9.00 & 9.00 \\ 0.33 & 0.14 & 0.14 & 1.00 & 1.00 & 0.14 & 0.14 & 9.00 & 7.00 & 9.00 & 9.00 & 9.00 \\ 7.00 & 0.14 & 0.14 & 7.00 & 7.00 & 1.00 & 3.00 & 9.00 & 7.00 & 9.00 & 9.00 & 9.00 \\ 7.00 & 0.11 & 0.14 & 7.00 & 7.00 & 0.33 & 1.00 & 9.00 & 9.00 & 9.00 & 9.00 & 9.00 \\ 0.11 & 0.14 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 1.00 & 0.20 & 0.20 & 0.20 & 0.14 \\ 0.14 & 0.11 & 0.14 & 0.14 & 0.14 & 0.14 & 0.11 & 5.00 & 1.00 & 1.00 & 0.14 & 0.14 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 5.00 & 1.00 & 1.00 & 0.14 & 0.14 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 5.00 & 7.00 & 7.00 & 1.00 & 0.20 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 7.00 & 7.00 & 5.00 & 7.00 & 1.00 \end{bmatrix}$$

$$U_{14} = \begin{bmatrix} 1.00 & 3.00 & 0.33 & 0.20 & 0.14 & 0.14 & 0.20 & 7.00 & 7.00 & 7.00 & 9.00 & 7.00 \\ 0.33 & 1.00 & 0.33 & 0.20 & 0.14 & 0.14 & 0.20 & 7.00 & 7.00 & 7.00 & 9.00 & 7.00 \\ 3.00 & 3.00 & 1.00 & 0.20 & 0.14 & 0.14 & 0.20 & 7.00 & 7.00 & 7.00 & 9.00 & 7.00 \\ 5.00 & 5.00 & 5.00 & 1.00 & 0.20 & 3.00 & 5.00 & 9.00 & 7.00 & 7.00 & 9.00 & 9.00 \\ 7.00 & 7.00 & 7.00 & 5.00 & 1.00 & 3.00 & 5.00 & 9.00 & 7.00 & 7.00 & 9.00 & 9.00 \\ 7.00 & 7.00 & 7.00 & 0.33 & 0.33 & 1.00 & 5.00 & 9.00 & 7.00 & 7.00 & 9.00 & 9.00 \\ 5.00 & 5.00 & 5.00 & 0.20 & 0.20 & 0.20 & 1.00 & 7.00 & 7.00 & 7.00 & 9.00 & 7.00 \\ 0.14 & 0.14 & 0.14 & 0.11 & 0.11 & 0.11 & 0.14 & 1.00 & 0.20 & 0.33 & 3.00 & 0.33 \\ 0.14 & 0.14 & 0.14 & 0.14 & 0.14 & 0.14 & 0.14 & 5.00 & 1.00 & 7.00 & 9.00 & 7.00 \\ 0.14 & 0.14 & 0.14 & 0.14 & 0.14 & 0.14 & 0.14 & 3.00 & 0.14 & 1.00 & 3.00 & 0.14 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.33 & 0.11 & 0.33 & 1.00 & 0.14 \\ 0.14 & 0.14 & 0.11 & 0.11 & 0.11 & 0.11 & 0.14 & 3.00 & 0.14 & 7.00 & 7.00 & 1.00 \end{bmatrix}$$

$$U_{15} = \begin{bmatrix} 1.00 & 7.00 & 7.00 & 9.00 & 9.00 & 3.00 & 3.00 & 9.00 & 9.00 & 9.00 & 9.00 & 9.00 \\ 0.14 & 1.00 & 3.00 & 3.00 & 3.00 & 5.00 & 5.00 & 9.00 & 9.00 & 0.14 & 5.00 & 5.00 \\ 0.14 & 0.33 & 1.00 & 7.00 & 7.00 & 7.00 & 7.00 & 9.00 & 9.00 & 0.14 & 7.00 & 7.00 \\ 0.11 & 0.33 & 0.14 & 1.00 & 0.14 & 3.00 & 3.00 & 9.00 & 9.00 & 0.14 & 7.00 & 9.00 \\ 0.11 & 0.33 & 0.14 & 7.00 & 1.00 & 5.00 & 5.00 & 9.00 & 9.00 & 0.14 & 7.00 & 9.00 \\ 0.33 & 0.20 & 0.14 & 0.33 & 0.20 & 1.00 & 3.00 & 9.00 & 9.00 & 0.14 & 9.00 & 9.00 \\ 0.33 & 0.20 & 0.14 & 0.33 & 0.20 & 0.33 & 1.00 & 9.00 & 9.00 & 0.14 & 9.00 & 9.00 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 1.00 & 1.00 & 0.14 & 1.00 & 9.00 \\ 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 1.00 & 1.00 & 0.14 & 1.00 & 9.00 \\ 0.11 & 7.00 & 7.00 & 7.00 & 7.00 & 7.00 & 7.00 & 7.00 & 7.00 & 1.00 & 7.00 & 7.00 \\ 0.11 & 0.20 & 0.14 & 0.14 & 0.14 & 0.11 & 0.11 & 1.00 & 1.00 & 0.14 & 1.00 & 1.00 \\ 0.11 & 0.20 & 0.14 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.11 & 0.14 & 1.00 & 1.00 \end{bmatrix}$$