VERİ MADENCİLİĞİ VE SEYREK YÖNTEMLERLE BİLGİ KEŞFİ:

Sağlık, Mutluluk ve Ekonomi Üzerine Uygulamalar

Editörler
Prof. Dr. Bahaddin SİNSOYSAL
Dr. Öğr. Üyesi Ayhan GÖLCÜKCÜ

1

VERİ MADENCİLİĞİ VE SEYREK YÖNTEMLERLE BİLGİ KEŞFİ: Sağlık, Mutluluk ve Ekonomi Üzerine Uygulamalar

YAZARLAR

Prof. Dr. Öznur İşçi GÜNERİ

Dr. Öğretim Üyesi Doğan YILDIZ

Öğr. Gör. Dr. Burcu DURMUŞ

Arş. Gör. Muzaffer GÖZTAŞ

Arş. Gör. Nida ORUÇ

Ece Nur GENÇER

EDİTÖRLER

Prof. Dr. Bahaddin SİNSOYSAL

Dr. Öğr. Üyesi Ayhan GÖLCÜKCÜ

ISBN: 978-625-5923-59-2

DOI: https://doi.org/10.5281/zenodo.15599950

Publishing House®

(The Licence Number of Publicator: 2018/42945)

E mail: ubakyayinevi@gmail.com

www.ubakyayinevi.org

It is responsibility of the author to abide by the publishing ethics rules.

UBAK Publishing House – 2025©

ISBN: 978-625-5923-59-2

Cover Design: Dr.Öğr. Üyesi Cengiz ÖZBEK

June / 2025

Ankara / Turkey

İÇİNDEKİLER

ÖNSÖZ

Veri, günümüz dünyasının en değerli kaynaklarından biri haline gelmiş; sosyal, ekonomik ve çevresel sistemlerin anlaşılmasında vazgeçilmez bir rol oynamaya başlamıştır. Gelişen bilgi teknolojileriyle birlikte, büyük veri ortamlarında anlamlı desenlerin keşfi, yalnızca klasik yöntemlerle değil, aynı zamanda yüksek boyutlu ve seyrek yapılarla başa çıkabilen modern algoritmalarla da mümkün hale gelmiştir. Matematiğin istatistiğe, istatistiğin veri bilimine evirildiği bir düzlemde, çok boyutlu veri setlerini analiz etmek için geliştirilen çağdaş yöntemler, hem teorik hem de uygulamalı yönleriyle öne çıkmaktadır. Bu bağlamda, veri madenciliğine dayalı seyrek istatistiksel yöntemlerin çok boyutlu veri kümeleri üzerindeki uygulamalarını disiplinlerarası bir perspektifle ele almak bir gereklilik olarak karşımıza çıkmaktadır. Bu kapsamda bu kitap, dört farklı veri setini kullanarak modern analiz yöntemlerini ele almakta ve veri madenciliği algoritmalarının karşılaştırılmasından, zaman serilerinde tekrar eden yapıların keşfine, seyrek bileşen analizinden ekonomik veri sınıflandırmalarına kadar geniş bir yelpazede yöntemsel ve uygulamalı çözümler sunmaktadır.

İlk bölümünde temel ve meta algoritmaların farklı veriler üzerindeki başarıları incelenmiş, ilk adımda kayıp veri tamamlama ve normalizayon temelinde veri önişleme analizi, ikinci aşamada ise WEKA programı ekseninde veriler temel ve meta sınıflandırıcılar ile modellenmiş ve elde edilen sonuçlar çeşitli performans metrikleri aracılığı ile karşılaştırılmıştır.

İkinci bölümde; 2023 yılı Dünya Mutluluk Raporu veri setinde yer alan altı temel göstergeye Seyrek Temel Bileşenler Analizi (STBA) uygulanmış ve ülkelerin refah düzeylerini yansıtan ve Sosyo-Ekonomik Gelişmişlik Göstergeleri (SeyrekBileşen1) ile İstihdam Yapısı (SeyrekBileşen2) bileşenlerinden oluşan iki boyutlu bileşen düzlemi elde edilmiştir. Bu bileşenler temelinde, K-Ortalamalar algoritması kullanılarak ülkeler benzer gösterge profillerine göre de gruplandırılmış ve analiz bulguları; ülkelerin benzer sosyo-ekonomik özellikler çerçevesinde kümelendiğini ve özellikle Kişi Başına GSYİH (Log), Sosyal Destek, Sağlıklı Yaşam Beklentisi, Yaşam Seçimi Özgürlüğü,

Cömertlik ve Yolsuzluk Algısı değişkenleri açısından da anlamlı biçimde ayrıştığını göstermiştir.

Üçüncü bölümde; Seyrek K-Ortalamalar, Seyrek Temel Bileşenler Analizi (Sparse PCA) ve Ufak Örneklem K-Ortalamalar (Mini-Batch K-Means) yöntemleri kullanılarak Avrupa Birliği'ne üye 27 ülkenin 2019-2023 yılları arasındaki madencilik ve taş ocakçılığı ürünlerine ilişkin ithalat ve ihracat miktarları ve değerlerine ait veriler üzerinden ülkelerin ithalat ve ihracat dengeleri, ekonomik entegrasyon düzeyleri ve sektörel bağımlılıkları açısından anlamlı kümelere ayrıldığı gösterilmiştir. Elde edilen bulgular, Avrupa Birliği içinde madencilik ve taş ocakçılığı sektörlerinin ülkeler arası ekonomik yapılarla olan ilişkisini ortaya koymak açısından da önemlidir.

Son bölümde; değişkenin kendi içinde trendsiz olması, aynı dereceden durağanlığa sahip olması ve kesit kümeleri arasında yüksek korelasyon olması varsayımları altında, zaman serilerinin klasik doğrusal regresyon modellerinde uygulanabilirliğini kolaylaştırmak için epsilon uzaklıklı kümeler kullanılarak zaman serisi verilerinden kesit veri üretimi önerilmiş ve önerilen yöntemin etkinliği, iklim verileri kullanılarak sıcaklık tahmini üzerinden test edilmiştir. Buradaki yaklaşımın amacı, varsayımlara dayalı klasik zaman serisi analizlerinde karşılaşılan zorlukları azaltarak, regresyon modellerinin genelleştirilebilirliğini ve uygulanabilirliğini artırmaktır.

Netice olarak, bu kitapta ele alınan yöntemler, yüksek boyutlu ve karmaşık veri yapılarının analizinde karşılaşılan birtakım zorluklara çözüm üretmeyi amaçlayan çağdaş yaklaşımları temsil etmektedir.

Verinin gücünü keşfetmek ve onu anlamlı bilgiye dönüştürmek isteyen herkese faydalı olması dileğiyle…

19/06/2025

Prof. Dr. Bahaddin SİNSOYSAL

Dr. Öğr. Üyesi Ayhan GÖLCÜKCÜ

BÖLÜM 1:

VERİ MADENCİLİĞİNDE TEMEL VE META (TOPLULUK) ALGORİTMALARIN WEKA İLE KARŞILAŞTIRMASI: SAĞLIK ALANI UYGULAMASI

Burcu Durmuş[1]
Öznur İşçi Güneri[2]

Özet

Son yıllarda teknolojide yaşanan gelişmeler verilerin hızla depolanmasını kolaylaştırmıştır. Bu durum karmaşık ve farklı yapılarda verilerin ortaya çıkmasına sebep olmuştur. Bunun sonucu olarak manuel olarak verilerin analiz edilmesi oldukça zor bir hal almıştır. Bu noktada veri madenciliği çalışmaları hız kazanmış ve farklı yapılardaki karmaşık verilerden anlamlı ve doğru bilgilerin ortaya çıkarılması işi kolaylaşmıştır. Bu kapsamda araştırmacılar öncelikle temel sınıflandırıcılar ile çalışmıştır. İlerleyen çalışmalarda ise, meta (topluluk) sınıflandırıcıları önerilmiştir. Bu çalışmada, temel ve meta algoritmaların farklı veriler üzerindeki başarıları incelenmiş olup genel çıkarımlar üzerine tartışılmıştır. Çalışmanın ilk adımında veri setleri veri önişleme analizine tabi tutulmuştur. Bu aşamada kayıp veri tamamlama ve normalizasyon üzerinde incelemelerde bulunulmuştur. İkinci aşamada ise veriler temel ve meta sınıflandırıcılar ile modellenmiş ve elde edilen sonuçlar çeşitli performans metrikleri aracılığı ile karşılaştırılmıştır. Çalışma, veri seti için uygun algoritma seçimi için çeşitli önbilgiler sunarken aynı zamanda veri önişleme analizinin önemini vurgulamaktadır.

[1] Öğr. Gör. Dr., Tekirdağ Namık Kemal Üniversitesi, Rektörlük, Tekirdağ, Türkiye, bdurmus@nku.edu.tr
[2] Prof. Dr., Muğla Sıtkı Koçman Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Muğla, Türkiye, oznur.isci@mu.edu.tr

COMPARISON OF BASIC AND META (ENSEMBLE) ALGORITHMS WITH WEKA IN DATA MINING: HEALTH FIELD APPLICATION

Abstract

In recent years, technological developments have facilitated the rapid storage of data. This has led to the emergence of complex and diverse data. As a result, manual analysis of data has become quite difficult. At this point, data mining studies have accelerated and the task of extracting meaningful and accurate information from complex data with different structures has become easier. In this context, researchers have primarily worked with basic classifiers. In subsequent studies, meta (ensemble) classifiers have been proposed. In this study, the success of basic and meta algorithms on different data has been examined and general inferences have been discussed. In the first step of the study, data sets were subjected to data preprocessing analysis. In this stage, examinations were made on missing data completion and normalization. In the second stage, data were modeled with basic and meta classifiers and the obtained results were compared through various performance metrics. While the study provides various preliminary information for the selection of the appropriate algorithm for the data set, it also emphasizes the importance of data preprocessing analysis.

1.1. GİRİŞ

Gelişen teknoloji ile birlikte verilerin depolanması artmış ve verilerde çeşitlilik ortaya çıkmıştır. Bunun sonucu olarak verilerin insan eli ile analiz edilmesi oldukça zor bir hal almıştır. Bu karmaşık verilerin analizinde bilgisayarlar ve özellikle makine öğrenmesi algoritmaları büyük rol oynamaktadır. Verilerin makine öğrenmesi algoritmaları ile modellenerek analiz edilmesi araştırmacılara kolaylık sağlamanın yanı sıra ileriye dönük güçlü tahminlerde bulunulmasına yardımcı olmaktadır.

Makine öğrenmesi yöntemleri, araştırmacıların bilgiyi en uygun şekilde kullanmalarına ve elde etmek istedikleri bulguya kısa yoldan ulaşmalarına olanak sağlamaktadır. Farklı bilgi kaynaklarından farklı yöntemlerle depolanan ve veri tabanları aracılığıyla ulaşılabilen büyük veriler, makine öğrenmesi yöntemleri ve teknolojik gelişmeler yardımıyla analiz edilmektedir. Elde edilen bulgular ekonomi, teknoloji, eğitim, sağlık gibi pek çok alanda kullanılmaktadır. Bu açıdan makine öğrenmesi yöntemleri ile yapılan çalışmalar günümüzde oldukça yaygınlaşmıştır.

Makine öğrenmesi yöntemleri günümüzde pek çok sağlık araştırmasında kullanılmasına rağmen, sağlık alanında uygulamalarının başlaması diğer alanlara göre daha uzun yıllar almıştır. Bunun en temel nedeni, bazı sağlık uzmanlarının makine öğrenmesi yöntemlerine güven duymamaları ve hastalık teşhisinde bu yöntemleri yanıltıcı bulmalarıdır (Durmuş vd., 2023). Ancak 1990'lı yıllara gelindiğinde hastalık ve tedavi maliyetlerini tahmin etmek için sinir ağlarının

kullanımı yaygınlaşmıştır (Yıldırım vd., 2008). Böylece makine öğrenmesi yöntemlerinin sağlık alanında yaygın olarak kullanılmasının önü açılmıştır. Günümüzde sağlık alanında kullanılan Sağlık Bilişim Sistemleri, hastalık teşhisinde uzmanlara kolaylık sağlamaktadır. Makine öğrenmesi algoritmasına dayalı bu sistemlerden erken tanı ve teşhis, uygun tedavi, iyileşme süresi tahmini, bütçe belirleme gibi konularda sıklıkla yararlanılmaktadır.

Makine öğrenmesi algoritmaları ile kurulan modellerin başarısı büyük ölçüde toplanan verilere bağlıdır. Ancak, verinin doğrudan analiz edilmesi uygun bir aksiyon değildir. Verilerden doğru bilgi çıkarımı yapılabilmesi için analiz sürecinde verinin analize hazır hale getirilmesi amacıyla veri ön işleme aşaması etkin bir rol oynamaktadır.

Çoğu makine öğrenmesi çalışmasında veriler için önişleme analizi yapmak göz ardı edilmektedir. Veri ön işleme teknikleri bazen bilgi kaybına neden olarak modelin yanlılığının artmasına etki edebilir. Ancak uygun teknik yardımıyla bu durum ile başa çıkıldığında modelin daha güvenilir sonuçlar verebileceği unutulmamalıdır. Kayıp gözlem içeren ya da gözlem değerleri arasında aşırı fark olan veri setleri için veri önişleme tekniklerinin uygulanması model yanlılığı ve güvenirliliği açısından daha etkin sonuçlar verecektir.

Veri ön işleme aşaması temelde kayıp verilerin tamamlanması ve normalizasyon işlemi olarak gerçekleştirilir. Veri setlerinde kayıp (eksik) veri varlığı sıklıkla karşılaşılan bir durumdur. Kayıp veri varlığı çeşitli sebepler ile ortaya çıkabilir. Veri kaynağından doğan problemler, verinin saklanması ve korunması sırasında ortaya çıkan problemler,

veriler arasındaki ilişkisel durumlar, veriyi kayıt eden personelin dikkatsizliği, verinin çeşitli sebeplerden dolayı silinmesi bu sebepler arasında sayılabilir.

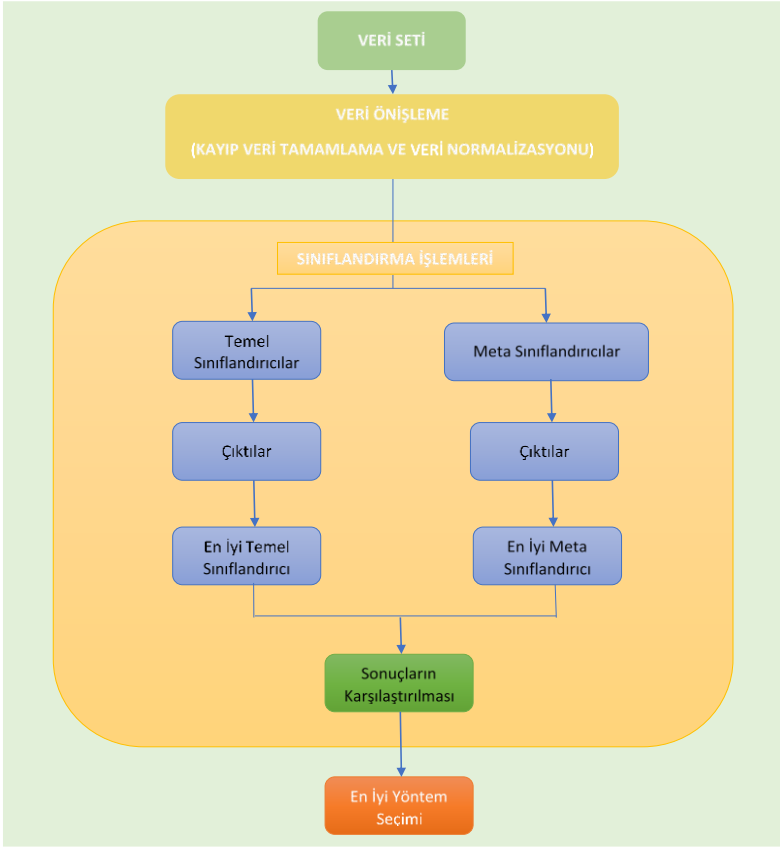
Kayıp verilerin varlığı istatistiksel analiz sonuçlarının yanlı olmasına ve model etkinliğinin azalmasına sebep olmaktadır. Ayrıca kayıp verilerin bulunduğu bir veri seti ile çalışmak negatif olarak kümülatif bir etki doğurmaktadır. Bu nedenle analiz sonuçları elde edilmeden önce verilerdeki kayıplar tespit edilmeli ve giderilmelidir. Bu noktada asıl üzerinde durulması gereken konu, kayıp verilerin veri setinin karakteristik unsurlarını taşıyacak şekilde tamamlanmasıdır. Veri setinin yapısına uygun olmayan bir yöntem seçimi, analiz sonuçlarının güvenirliğinin azalmasına sebep olacaktır. Dolayısıyla sağlıklı bir analiz sonucu elde etmek için kayıp verileri tamamlama yöntem seçiminde titiz davranılması gerekmektedir. Literatürde, kayıp veri problemi için geliştirilmiş pek çok yöntem bulunmaktadır. En yaygın kullanılan yöntemler kayıp veri yerine verilerin ortalama ya da medyanın kullanılması, eksik veri içeren gözlemin veri setinden çıkarılması, kayıp verinin önceki ya da sonraki değer ile tamamlanması ve karar ağacı, komşuluk derecesi gibi teoriler ile yeni bir değer atamak olarak sayılabilir.

Verilerin analize hazır hale getirilmesindeki bir diğer önemli husus ise veri normalizasyonudur. Veri normalizasyonu, farklı değer aralıklarında bulunan verilerin aynı değer aralığına ölçeklendirilmesi işlemidir. Bu işlem, özellikle farklı özniteliklere sahip veri setlerinde daha da önem kazanmaktadır. Bunun en temel sebebi özniteliklerin

farklı değer aralıklarında olmasıdır. Bu noktada veri normalizasyonuna ihtiyaç duyulmaktadır. Bu konuda literatürde farklı yaklaşımlar bulunmaktadır. Bu yaklaşımların çok tercih edilenleri min-max, medyan ve z-score gibi yaklaşımlardır.

Literatürde yer alan makine öğrenmesi algoritmalarının etkinliğinin incelendiği çalışmalarda veri normalizasyonu ve kayıp veri tamamlama işlemlerinin birlikte değerlendirdiği çalışmaların kısıtlı olduğu görülmektedir. Özellikle sağlık gibi önemli bir alanda veri seti üzerinde hem veri ölçekleme hem de eksik veri tamamlama işlemlerinin algoritmalar üzerindeki etkisinin incelenmesi önem teşkil etmektedir.

Bu çalışma kapsamında, sağlık alanından farklı veri yapılarına sahip sekiz veri seti kullanılmıştır. Veri setlerindeki kayıp gözlemler ortalama, medyan, kNN, Rpart ve Mice yöntemleri ile tamamlanmıştır. Her bir veri seti için en uygun yöntem RMSE değeri hesaplanarak tespit edilmiştir. Çalışmada kullanılan veriler için normalizasyon tekniğinin sınıflandırma analizi kapsamında uygun olup olmayacağının araştırılması için min-max, sigmoid, unit, decimal, softnorm yöntemleri kullanılmıştır. Sonuçlar ham veriler ile birlikte Random Forest sınıflandırma analizi yardımıyla yorumlanmıştır. Veri ön işleme aşaması tamamlanan veri setleri, WEKA paket programı kullanılarak temel ve meta algoritmalar ile sınıflandırılmıştır. Sınıflandırma sonuçları performans değerlendirme kriterleri esas alınarak değerlendirilmiştir. Çalışma özeti Şekil 1.1'de verilen akış diyagramı ile açıklanmaktadır.



Şekil 1.1. Çalışma adımları

1.2. MATERYAL VE METOD

Bu çalışmada, WEKA programı ile pratik bir şekilde sonuç elde edilebilen temel ve meta algoritmaların farklı veri setleri üzerindeki sınıflandırma başarıları incelenmiştir. Bu kapsamda, çeşitli performans metrikleri yardımıyla sonuçlar ölçülmüş ve okuyucuya sunulmuştur. Bu amaç doğrultusunda çalışmada 8 farklı veri seti materyal olarak; WEKA programında yer alan sınıflandırma algoritmaları, performans metrikleri, kayıp veri tamamlama yöntemleri ve normalizasyon yöntemleri ise metot olarak kullanılmıştır.

1.2.1. Sınıflandırma Algoritmaları

Verileri önceden kategorize edilmiş bir eğitim setine göre kategorilere ayırma süreci olarak tanımlanan sınıflandırma işlemi, denetimli bir öğrenme yöntemidir. Sınıflandırma algoritmalarının en iyi bilinen örnekleri iki sınıflı tahminlerdir: "var" veya "yok", "spam" veya "spam değil", "hasta" veya "hasta değil". Sınıflandırma algoritmaları, eğitim setinden öğrendiği bilgiler ile çeşitli kurallara dayalı modeller kurar ve test verileri ile modelin performansını ölçer.

Çalışmamızın bir diğer konusunu meta (topluluk) algoritmalar oluşturmaktadır. Meta algoritmalar, daha iyi performans ile sınıflandırma yapabilmek için iki veya daha fazla makine öğrenmesi algoritmasının tahminlerini oylama veya sonuçların ortalamasını alma yoluyla birleştirir. Bu çalışmada sınıflandırma analiz sonuçları için WEKA programı kullanılmıştır. Bölümün devamında çalışmada ele alınan algoritmalar açıklanmaktadır.

1.2.1.1. *Çalışmada Kullanılan Temel Algoritmalar*

Temel ve meta algoritmaların sınıflandırma performansına etkisinin incelendiği bu çalışmada, WEKA programının sınıflandırma analizi penceresi altında yer alan ve farklı matematiksel modellere dayanan J48, LMT, Random Forest, Naive Bayes, Bayes Net, kNN, destek vektör makineleri ve yapay sinir ağları algoritmaları kullanılmıştır.

a. J48

ID3 ve C4.5 algoritmalarının WEKA'ya uyarlanmış bir türevidir. Bu algoritma, öğrenilmesi ve uygulanması bakımından en yaygın ve en basit sınıflandırma algoritmalarından biridir. J48 algoritması temelde

bir karar ağacıdır. Ağaç üç bileşenden oluşur: kök düğüm, dal (bağlantı) ve yaprak düğüm. Kök, farklı nitelikler için test koşulunu temsil eder. Dal testte bulunabilecek tüm olası sonuçları temsil eder. Yaprak düğümler ise ait olduğu sınıfın etiketini içerir (Salzberg, 1994; Mustakim vd., 2024). Bu algoritmada amaç, karar ağacını optimize etmektir (Coşkun ve Baykal, 2011).

b. Logistic Model Tree (LMT)

Bu algoritma, lojistik regresyon modelini ve karar ağaçlarını bir arada kullanır. Teknik, karar ağacının yapraklarının parçalı doğrusal regresyon modeline sahip bir karar ağacı yapısı olmasına dayanır (Landwehr vd., 2005).

c. Random Forest (RF)

Random Forest (RF), özellik verileri arasındaki korelasyonu azaltmak için çok sayıda karar ağacını birleştiren bir tekniktir. O hata fonksiyonu olmak üzere, RF'nin hesaplama maliyeti *O(n)*'dir. Bu entegrasyon teknik olarak paralel yürütmeyi sağlar ve bu da artan hız sağlar. RF'de örnekler ve özellikler rastgele seçilir. Böylece karar ağaçları arasındaki korelasyonun azaltılması sağlanır (Cutler vd., 2011; Salman vd., 2024).

RF algoritmasında birden fazla karar ağacı oluşturulur ve bu ağaçlar birleştirilerek karar ormanları elde edilir. Ağaç yapısındaki her bir veri seti, gerçek veri setinden yer değiştirmeli olarak türetilir.

d. Naive Bayes (NB)

Olasılık teorisinin temellerine dayanan Naive Bayes (NB), istatistiksel bir sınıflandırma modeli sunar (Başarslan, 2017). Tekniğin amacı, olasılık teorisi yardımıyla belirsizlik durumlarını tespit etmektir. Bu

algoritma için verilerdeki tüm nitelikler aynı derecede önemlidir. Başka bir deyim ile, niteliklerin birbirinden bağımsız oldukları kabul edilir.

e. Bayes Net (BN)

Bayes ağları olarak da bilinen bu algoritma, Naive Bayes algoritmasında olduğu gibi Bayes teoremine dayanır. Veri modelleme için kullanılan Bayes ağları, istatistiksel ağlardır ve düğümler arası geçiş yapan kolları istatistiksel kararlara göre seçer. Bayes ağlarının daha geniş hali belirsiz karar ağaçları olarak bilinir (Van Harmelen vd., 2008).

f. En Yakın Komşu (kNN)

En yakın komşu algoritması, gözlemler arası uzaklığa dayalı bir tekniktir (Trevor vd., 2011). Verilerin birbirlerine olan uzaklıkları genel olarak Öklid uzaklık ölçüsü ile hesaplanır. Algoritma, beş adımda özetlenebilir:

1. k komşuluk parametresi seçilir.
2. Mesafe ölçüm uzayı belirlenir.
3. Birbirine en yakın k nokta belirlenir.
4. Bu gözlemlerin en sık atandığı sınıf belirlenir.
5. Yeni verinin bu sınıfa ait olduğu kabul edilir.

g. Destek Vektör Makineleri (DVM)

Destek vektör makinaları sınır düzlemlerine dayalı olarak tasarlanmış bir tekniktir. Bu yöntemde sınıflandırma için ideal sınır belirlenir. Algoritma, sınıflandırma sonuçlarını en iyilemek için ideal sınırı (doğru ya da düzlem) esas alarak model kurar. Algoritma farklı çekirdek

fonksiyonları ile çalışır. Bu fonksiyonlar arasında lineer, polinomyal, radyal ve sigmoid çekirdekler en temelleridir (Li vd., 2014).

h. Yapay Sinir Ağları (YSA)

Yapay Sinir Ağları (YSA), insan sinir sistemini oluşturan ağların biyolojik sinir ağlarına benzetilmesiyle tasarlanan bir bilgisayar programıdır (Elmas, 2003). İnsan beyninde birbirine bağlı çok sayıda nöron (sinir hücresi) bulunur. Yapay sinir ağları da birbirine bağlanmış yapay nöronlardan oluşur. Yapay sinir ağlarındaki her bir nöron, kendinden önce gelen nöronlar tarafından kendisine iletilen bilgileri işler ve bulduğu sonucu diğer nöronlara iletir.

1.2.1.2. *Çalışmada Kullanılan Meta Algoritmalar*

Çalışmada WEKA programının 'meta' klasörü altında yer alan algoritmalar ele alınmıştır. Meta algoritmalar için temel öğrenici olarak Random Forest algoritması seçilmiş olup, çalışma sonuçlarında Random Forest algoritması doğruluk değeri ayrı olarak gösterilmiştir.

a. Ada Boost

AdaBoost algoritması, her bir öznitelikten zayıf sınıflandırıcıların bir araya getirilerek güçlü bir sınıflandırıcı topluluğu elde edilmesi mantığına dayanır. Her aşamada 'temel öğrenici' adı verilen bir öğrenme algoritması çağırır ve boosting sınıflandırıcısını oluşturur. Bu sınıflandırıcıya ağırlık katsayısı atanır. Burada hata oranı en düşük zayıf sınıflandırıcılar yeni bir güçlü sınıflandırıcı oluşturmada rol oynar. Böylece, güçlü sınıflandırıcı olarak nitelendirilmeyen zayıf sınıflandırıcılara ilişkin öznitelikler çıkarılır. Zayıf sınıflandırıcıların hatası ne kadar düşük olursa, son aşamada ağırlığı o kadar yüksek olur.

Son sınıflandırma kararı ise her bir öznitelik için negatif ve pozitif örneklerin ağırlıklı ortalaması alınarak ağırlıklı oylaması ile elde edilir (Kégl, 2009).

b. Bagging

Breiman (1996) tarafından önerilen Bagging yöntemi, önceden belirlenmiş bir eğitim setinden yeni eğitim setleri türeterek temel öğreniciyi yeniden eğitme mantığına dayanır. Bu yöntem ile tahmin edicinin farklı versiyonları oluşturularak birleştirilmiş bir tahmin edici elde edilir (Tu vd., 2009). Algoritma adımları, başlangıçta yer alan eğitim setinin girdi olarak sisteme alınmasıyla başlar. Yeni eğitim setlerinin oluşturulması ve temel sınıflandırıcının eğitilmesi ile devam eder. Bu yöntemde birbirinden farklı eğitim setleri elde edilir. Her eğitim seti, temel öğrenici ile eğitilir ve sonuçlar ağırlıklı oylama (voting) yöntemiyle birleştirilir. Bagging yönteminde temel amaç, yeni veri setleri yardımıyla farklılıkları belirleyerek nihai sınıflandırma başarısını artırmaktır.

c. Stacking

Stacking algoritmasında, eğitim veri setinin *n* farklı alt kümesi oluşturulur. Her bir alt örneklem seti seçilen bir öğrenici yardımıyla eğitilir. Bu yöntemde, alt modellerin çıktıları girdi olarak sisteme alınır. Algoritma, iyi bir model tahmini için alt küme performanslarını birleştirir. Model sonunda elde edilen çıktılar meta sınıflandırıcı için girdi olarak alınır ve meta sınıflandırıcı aracılığıyla birleştirme işlemi yapılır (Akram ve Taser, 2020). Birleştirme işlemi, her bir modelin performansıyla orantılı olarak bir ağırlık atanmasıyla yapılır (Sikora ve

Al-laymoun, 2014). Yöntem için seçilecek meta sınıflandırıcı için herhangi bir kural bulunmamaktadır.

d. Logit Boost

Logit Boost yöntemi, Friedman ve arkadaşları tarafından 1998'de önerilen bir yaklaşımdır (Friedman vd., 2000; Sun vd., 2014). Yöntem hem iki hem de çok sınıflı öğrenmeyi desteklemektedir. Logit Boost, ilave bir lojistik model oluşturmak için uyarlanabilir Newton adımlarını kullanır. Logit Boost denklem olarak lojistik kayıp fonksiyonu (negatif koşullu log-olasılık) ile bir boosting yaklaşımı sunar (Tanha vd., 2020):

$$L_{LogitBoost}(y, f) = -\sum_{k=1}^{K} ḃ(k = y) log P_k$$

$$P_k(x) = \frac{\exp(F_k(x))}{\sum_{j=1}^{K} \exp(F_j(x))} ve \ \sum_{k=1}^{K} F_k(x) = 0$$

$$P_k = \frac{1}{K}, k = 1,2, \dots, K$$

$$y_i = (y_1, \dots, y_K)^T, i = 1,2, \dots, N$$

ḃ: doğru olduğunda değeri 1' e eşit aksi takdirde sıfıra olan gösterge fonksiyonudur.

e. Attribute Selected Classifier

Attribute Selected Classifier yöntemi, sınıflandırma analizine başlanmadan önce eğitim ve test verilerinin boyutunun öznitelik seçimi yoluyla azaltılması ile başlar. Devamında temel eğitim setinin bir alt kümesi ve yeni bir aday alt küme oluşturulur. Aday küme değerlendirme sonucunda daha iyi sonuçlar verirse, en iyi alt küme olarak seçilir. Sonlandırma koşulu sağlanıncaya kadar işlemler devam eder. Çalışma adımlarının tamamlanması sonucu elde edilen öznitelik sayısı azalmış yeni alt küme, önceden belirlenmiş bir sınıflandırıcı yardımıyla modellenir ve sınıflandırma işlemi tamamlanır (Malik vd., 2019).

f. Classification Via Regression

Classification Via Regression yönteminde problemler, regresyon fonksiyonlarına dönüştürülür (Ruan vd., 2014). Yöntemin iki ana adımı vardır (Arora ve Dhir, 2017):

1. Karar ağacı yapısı oluşturulur.
2. Karar ağacının olası dalları üzerinde budama yapılır ve regresyon fonksiyonu ile yapraklar üzerinde birleştirilir.

g. Random Committee

Random Committee yönteminde, temel sınıflandırıcılardan bir topluluk oluşturulur ve tahminler ortalama olasılık tahminleriyle birleştirilir. Bu yöntemde temel sınıflandırıcılar aynı verilere dayanmakta ve farklı rastgele sayı çekirdeği kullanılmaktadır. Bu durum, sadece temel sınıflandırıcı rastgele seçildiğinde anlamlı olmaktadır. Aksi durumda, tüm sınıflandırıcılar aynı olur (Witten ve Frank, 2005).

h. Cost Sensitive Classifier

Cost Sensitive Classifier, temel sınıflandırıcıyı maliyete duyarlı hale getiren bir yöntemdir. Yöntem, temel sınıflandırıcının örnek ağırlıklarını işleyemediği durumlarda (örnek ağırlıkları tek tip değilse) verilerin temel sınıflandırıcıya iletilmeden önce yeniden ağırlıklandırılması fikrine dayanır. Burada temel amaç, toplam maliyetin en aza indirilmesidir. CostSensitive sınıflandırıcısı için iki algoritma kullanılmaktadır. Birincisi, toplam maliyete göre eğitim örneklerini yeniden ağırlıklandırır. Diğeri ise, minimum seviyede beklenen yanlış sınıflandırma maliyetine ilişkin sınıfı tahmin eder (Kim vd., 2012).

i. Meta Cost

Sınıflandırıcıyı maliyete duyarlı hale getirmek için geliştirilen yöntem, eğitim setinde yer alan gözlemlerin maliyete duyarlı bir sınıflandırıcı ile yeniden etiketlenmesi mantığına dayanır (Domingos, 1999). Hataya dayalı bir sınıflandırıcıyı maliyete duyarlı bir sınıflandırıcıya dönüştürmek için izlenecek en basit yol, minimum beklenen maliyet kriteri eklemektir (Michie vd., 1994). Bu kriter yalnızca sınıflandırma sırasında uygulanmakta ve uygulanan bu prosedür maliyet değiştiğinde aynı modelin tekrar kullanılmasına olanak vermektedir.

j. CV Parameter Selection

CV Parameter Selection yöntemi, seçilen sınıflandırıcı için çapraz doğrulama yoluyla parametre seçimi yapar. Ardından sınıflama işlemi yapılır (Reddy vd., 2021).

k. Filtered Classifier

Filtered Classifier yöntemi, eğitim verilerinin rastgele bir filtreden geçirilerek sınıflandırma işlemi yapılması mantığına dayanır. Test verileri, yapıları değiştirilmeden filtre edilir. Filtre, eşit olmayan örnek ağırlıkları (veya öznitelik ağırlıkları) durumu ile başa çıkamıyorsa, örnekler filtreye iletilmeden önce ağırlıklara göre yeniden örneklenir (Kawade ve Oza, 2015).

l. Randomizable Filtered Classifier

Filtered Classifier yönteminden yola çıkılarak geliştirilen bu yöntem, modeli seçilen bir filtre ve kNN temel sınıflandırıcısı ile başlatır. Çalışma mantığı FilteredClassifier ile çok benzerdir. Filtreden geçirilen

eğitim verileri, seçilen bir sınıflandırıcı ile modellenir. Test verileri, yapıları değiştirilmeden filtrelenir (Hall vd., 2009).

m. Multi Class Classifier

Çok sınıflı verileri işlemek için önerilen bu yöntem, artan doğruluk için hata düzeltme çıkış kodlarını kullanır. Temel sınıflandırıcı örnek ağırlıklarını işleyemiyorsa ya da örnek ağırlıkları tek tip değilse, veriler temel sınıflandırıcıya iletilmeden önce ağırlıklara göre yeniden örneklenir (Makhtar vd., 2011).

n. Multi Scheme

MultiScheme, eğitim verileri üzerinde çapraz doğrulamayı kullanarak sınıflandırma yapan bir yöntemdir. Performans, yüzde doğru (sınıflandırma) veya ortalama kare hatasına (regresyon) göre ölçülür (Balogun vd., 2017).

o. Random Sub Space

Ho (1998) tarafından önerilen bu yöntem, Random Forest algoritmasının genelleştirilmesiyle oluşturulmuş paralel öğrenme algoritmasıdır. Yaygın bir biçimde temel sınıflandırıcı olarak karar ağaçları kullanılır. Her bir karar ağacı bağımsız bir şekilde oluşturulur. Random Sub Space yöntemi ile oluşturulan model, bir varlık uzayı üzerinde farklı değişkenleri kullanır (Skurichina ve Duin 2002). Random Sub Space algoritmasında bir öğrenici ve eğitim setinin farklı öznitelikleriyle oluşturulmuş yeni veri setleri oluşturulur. Yeni eğitim setleri, başlangıç eğitim setinin bazı boyutları silinerek türetilir. Örneğin, m özniteliği bulunan bir veri setinden rastgele n adet özniteliği bulunan ($n<m$) yeni veri setleri seçilir. Daha sonra belirlenen temel

öğrenici, yeni oluşturulan eğitim setleri ile eğitilir. Öğrenicinin test seti için kararı, yeni veri setleri ile elde edilen öğrenicilerin verdikleri kararlar birleştirilerek hesaplanır (Koçali, 2019).

p. Nested Dichotomies

Nested Dichotomies yöntemi, rastgele sınıf dengeli bir ağaç yapısı oluşturarak iki sınıflı sınıflandırıcıların yardımıyla çok sınıflı veri kümelerini sınıflandırmak için önerilmiştir. Nested Dichotomies bir ağaç yapısıdır, ancak karar ağacı değildir. Bu yöntemde düğümler, örnekleri bölmek için kullanılır ve sınıf yalnızca bir yaprağa atanır (Frank ve Kramer, 2004; Dong ve diğerleri, 2005).

q. Ultra Boost

Ultra Boost yöntemi, her biri farklı bir sınıflandırıcıya sahip iki veya üç aşamayı yapılandırarak her aşamada farklı bir sınıflandırıcı kullanma fikrine dayanmaktadır. Yöntemde varsayılan olarak, Naive Bayes ve lojistik regresyon kullanılır (Santosh, 2015).

r. Vote

Vote yönteminde birden fazla sınıflandırma algoritması aynı eğitim seti ile ya da tek bir sınıflandırma algoritması farklı parametre değerleri aracılığıyla aynı veri seti ile eğitilebilir. Böylece farklı sınıflandırma modelleri oluşturulur. Modellerden elde edilen çıktılar oylama tekniğiyle birleştirilerek nihai değer hesaplanır (Kittler vd., 1998; Kuncheva, 2004).

s. Weighted Instances Handler Wrapper

Weighted Instances Handler Wrapper yöntemi, ağırlıklandırma eğitim örnekleri için sarmalayıcı yaklaşımı kullanır (Karagiannopoulos vd.,

2007). Algoritma uygulanırken varsayılan olarak, örnek ağırlıkları kontrol edilir ve eğitim veri seti temel sınıflandırıcıya çaprazlanır. Varsayılan bir seçenek olarak, temel sınıflandırıcı ağırlıkları çalıştırabiliyorsa eğitim verilerini kullanır, ancak ağırlıklarla birlikte yeniden örnekleme yaklaşımlarını da uygulayabilir (Kargar vd., 2021).

Sağlıklı bir analiz yapmanın ilk şartı sağlıklı bir veri seti ile çalışmaktır. Bunun için verinin analize hazır hale getirilmesi gerekmektedir. Bu gereklilik verilerin birtakım problemler içermesinden kaynaklanmaktadır. İstatistiksel analizlerde en sık karşılaşılan durum veri setinin kayıp veriler barındırmasıdır. Kayıp veri durumu farklı sebeplerden ortaya çıkabilir: veri kaynağından doğan problemler, verinin saklanması ve korunması aşamasından kaynaklı problemler, veriler arasındaki bazı ilişkisel durumlar (Erken ve Şenyay, 2023).

1.2.2. Performans Metrikleri

İkili sınıflandırma sorunları için, sınıflandırma eğitimi sırasında en iyi (optimal) çözümün ayrım değerlendirmesi, Tablo 1.1'de gösterildiği gibi karmaşıklık matrisine göre tanımlanabilir. Tablonun satırı tahmin edilen sınıfı temsil ederken, sütun gerçek sınıfı temsil eder.

Tablo 1.1. İkili sınıflandırma için karmaşıklık matrisi

	Gerçek pozitif sınıf	Gerçek negatif sınıf
Tahmin pozitif sınıf	Doğru Pozitif (DP)	Yanlış Negatif (YN)
Tahmin negatif sınır	Yanlış Pozitif (YP)	Doğru Negatif (DN)

Bu karışıklık matrisinden, DP ve DN doğru sınıflandırılan pozitif ve negatif örneklerin sayısını belirtirken, YP ve YN sırasıyla yanlış sınıflandırılan negatif ve pozitif örneklerin sayısını belirtir. Tablo

1.1'den, Tablo 1.2'de gösterildiği gibi, farklı değerlendirme odaklarına sahip sınıflandırıcının performansını değerlendirmek için yaygın olarak kullanılan metrikler oluşturulabilir. Çoklu sınıflandırma değerlendirmeleri için Tablo 1.2'de listelenen ölçümler genişletilir.

Tablo 1.2. Sınıflandırma performansları için ölçüm metrikleri

Metrik	Formül	Açıklama
Doğruluk	$\frac{DP + DN}{DP + DN + YP + YN}$	Değerlendirilen toplam örnek sayısına göre doğru tahminlerin oranını ölçer.
Kesinlik	$\frac{DP}{DP + YP}$	Pozitif bir sınıftaki toplam tahmin edilen örüntülerden doğru şekilde tahmin edilen pozitif örüntüleri ölçmek için kullanılır.
Duyarlılık	$\frac{DP}{DP + DN}$	Doğru şekilde sınıflandırılan pozitif örüntülerin oranını ölçmek için kullanılır.
F-Ölçütü	$\frac{2 * p * r}{p + r}$	Duyarlılık ve kesinlik değerleri arasındaki harmonik ortalamayı temsil eder.
MCC	$\frac{DP * DN - YP * YN}{\sqrt{(DP + YP)(DP + YN)(DN + YP)(DN + YN)}}$	Gerçek veriler ile tahmin edilen veriler arasındaki korelasyon ilişkisine bakarak değerlendirme yapan MCC ölçütü, karmaşıklık matrisinde yer alan tüm parametreleri kullanır.
ROC Area	Modelin farklı kesme noktalarında performansını ve sınıfları ayırt etme yeteneğini görsel olarak sunar.	
PRC Area	PRC eğrisi ikili sınıflandırma algoritmalarının performansını değerlendirmek için kullanılır. Genellikle sınıfların aşırı dengesiz olduğu durumlarda kullanılır. Kesinlik-duyarlılık eğrileri, tek bir değer yerine, bir sınıflandırıcının birçok eşikteki performansının grafiksel bir gösterimini sağlar.	
Kappa	İki değerleyici arasındaki karşılaştırmalı uyuşmanın güvenirliğini ölçer. -1 ile 1 arasında değer alır. Sonuç 1'e yakın ise değerleyiciler arası uyum vardır, -1'e yakınsa uyum yoktur yorumu yapılır.	

1.2.3. Kayıp Veri Tamamlama Yöntemleri

Doğru ve güvenilir bir analiz için veri setinin kayıp veri içermemesi oldukça önemlidir. Kayıp veri problemi ile başa çıkabilmek için çeşitli yöntemler bulunmaktadır. En basit yöntem kayıp veri içeren kaydı yok saymaktır. Bu yöntem, özellikle kayıp verinin çok olduğu ya da veri sayısının az olduğu durumlarda modelde sapmalara neden olacaktır. Bu nedenle kayıp veri içeren gözlemi silmek yerine, kayıp veri için bir değer atamak daha iyi sonuçlar elde edilmesini sağlayacaktır. Bu çalışmada ortalama, medyan, kNN, Rpart ve Mice yöntemlerinden yararlanarak kayıp veri problemi giderilmiştir.

1.2.3.1. *Ortalama ve Medyan*

Kayıp verilerden kurtulmak için kullanılan en yaygın yöntemlerden biri kayıp veri yerine diğer gözlemlerin ortalamasının ya da medyan (ortanca) değerinin atanmasıdır. Bu yöntem veri aralığı düşük olan verilerde daha kullanışlıdır. Diğer durumlarda hata oranını arttırarak varyans değerini düşürmektedir.

1.2.3.2. *k-en yakın komşu (kNN)*

kNN yöntemi, gözlemler arası uzaklıklar üzerine kurulu bir yöntemdir. Kayıp veri içermeyen değişkenler arasındaki mesafe ölçülerek '*k*' en yakın gözlemlerin ortalaması hesaplanarak kayıp veriler tamamlanır (Little ve Rubin, 2014).

1.2.3.3. *Recursive partitioning (Rpart)*

Rpart ile kayıp veri tamamlama yöntemi karar ağaçlarının çalışma adımlarına dayalı olarak geliştirilen bir yöntemdir. Bu yöntemde, kayıp

veri yerine veri setindeki diğer değişkenler ile bir karar ağacı modeli oluşturur ve bu model yardımıyla kayıp değerler tahmin edilir.

1.2.3.4. *Multivariate Imputation by Chained Equations (Mice)*

Mice yöntemi, Buuren ve Groothuis-Oudshoorn (2011) tarafından önerilmiştir. Bu yöntemde, her değişken için koşullu bir model oluşturulur. Bu model, kayıp veriler için tahmin edici olarak kullanılır.

1.2.4. Veri Normalizasyonu Yöntemleri

İstatistiksel modellemelerde ham veriyi kullanmak her zaman kullanışlı olmayabilir. Veri setindeki değişkenlere ilişkin değer aralıklarının anlamlı derecede birbirinden farklı olduğu durumlarda normalizasyon işlemi uygulanır. Normalizasyon işlemi, değişkenlerin ortalama ve varyanslarını birbirlerine yaklaştırarak yüksek ortalamaya (ya da varyansa) sahip değişkenlerin diğer değişkenler üzerindeki baskısını azaltır. Yapılan analizlerde değişkenlerin sonuçlar üzerindeki etkisini standartlaştırmak amacıyla veriler normalize edilir (Özkan, 2008; Larose, 2014). Bu çalışma kapsamında normalizasyon işlemi için sigmoid, min-max, unit, decimal ölçeklendirme, softmax yöntemleri kullanılmıştır.

1.2.4.1. *Sigmoid*

Sigmoid normalizasyonu giriş verilerini doğrusal olmayan bir şekilde 0 ile 1 veya -1 ile 1 arasında sınıflandıran bir tekniktir. Orijinal veriler ilk önce ortalamaya göre merkezlenir ve daha sonra sigmoid doğrusal bölgesiyle eşlenir. Birkaç tane doğrusal olmayan sigmoid fonksiyon çeşidi vardır. Lojistik (Denklem 1) ve hiperbolik tanjant (Denklem 2)

fonksiyonları yaygın olarak kullanılmaktadır. Lojistik sigmoid verileri 0 ile 1 arasına dönüştürür. Hiperbolik tanjant ise verileri -1 ile 1 arasına dönüştürür. Bu teknik, özellikle aykırı değerlerin mevcut olduğu durumlar için oldukça kullanışlıdır.

$$x' = \frac{1}{1 + e^{-x_i}} \tag{1.1}$$

x': Normalize edilmiş veri
x_i: Girdi değeri
e: Doğal logaritma değeri

$$x' = \frac{e^{x_i} - e^{-x_i}}{e^{x_i} + e^{-x_i}} \tag{1.2}$$

x': Normalize edilmiş veri
x_i: Girdi değeri
e: Doğal logaritma değeri

1.2.4.2. *Min-Max*

Min-max normalizasyonu verileri doğrusal olarak normalize eden bir tekniktir. Başka bir ifade ile, normalize edilen iki değer arasındaki fark değişmez. Yöntemin uygulanmasında, bir özelliğin minimum değeri, özelliğin her bir değerinden çıkarılır ve ardından bulunan fark özelliğin aralığına bölünür. Bu yeni değerler, özelliğin yeni aralığı ile çarpılır ve son olarak özelliğin yeni minimum değerine eklenir. Bu yöntemde veriler, Denklem 3 yardımıyla 0-1 aralığında boyutlandırılır.

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1.3}$$

x': Normalize edilmiş veri
x_i: Girdi değeri
x_{max}: Girdideki en büyük sayı
x_{min}: Girdideki en küçük sayı

1.2.4.3. *Unit*

Unit (birim) normalizasyonu uygulandığında, her profildeki tüm veriler, ilgili vektörün uzunluğu 1'e eşit olacak şekilde çarpılır. Vektörün uzunluğu, tüm değerlerin karelerinin toplamının kareköküdür. Vektör normalizasyonu için Denklem 4 ile verilen değişkenlerin norm değeri hesaplanır. Daha sonra her bir değişken norm değerine bölünür ve Denklem 5 ile verilen yeni normalize değer bulunur (Gautam vd., 2015).

$$|x| = \sqrt{x_1^2 + x_2^2 + x_3^2 + \cdots + x_i^2} \tag{1.4}$$

$$x' = \frac{x_i}{|x|} \tag{1.5}$$

x': Normalize edilmiş veri
x_i: Girdi değeri

1.2.4.4. *Decimal Ölçeklendirme*

Decimal (ondalık) ölçeklendirme normalizasyonunda, bir matrise veya veri çerçevesine ondalık ölçekleme uygular. Ondalık ölçeklendirme Denklem 6'da açıklandığı üzere, her özelliğin maksimum değerinin mutlak değerinin 10^k ile bölünmesiyle 1'den küçük veya 1'e eşit olacak şekilde k'yi bularak verileri [-1,1] aralığına dönüştürür.

$$x' = \frac{x_i}{10^k} \tag{1.6}$$

x': Normalize edilmiş veri
x_i: Girdi değeri
k: $Max(|x'|) < 1$ olacak şekilde en küçük tamsayı

1.2.4.5. *Softmax*

Softmax normalizasyonu adını yöntemin işleyiş mantığından almaktadır. Bu yöntemde maksimum ve minimum değere doğru 'yumuşak' bir ilerleme olur, ancak tam olarak o noktalara ulaşılmaz. Dönüşüm orta aralıkta doğrusala yakındır ama uçlarda doğrusal değildir. Normalizasyon işlemi Denklem 7 ile verilen eşitlik ile yapılır ve kapsanan çıkış aralığı [0,1] şeklindedir.

$$x' = \frac{e^{x_i}}{\sum e^{x^i}} \tag{1.7}$$

x': Normalize edilmiş veri
x_i: Girdi değeri
e: Doğal logaritma değeri

1.2.5. Veri Setleri

Bu çalışmada Tablo 1.3'teki veri setlerine ilişkin genel bilgileri yer alan, sağlık alanından toplanan sekiz kategorideki veri seti kullanılmıştır.

Tablo 1.3. Veri setlerine dair bilgiler

Veri Seti	Değişken Sayısı	Nümerik Değişken Sayısı	Kategorik Değişken Sayısı
Erken Saç Beyazlaması	26	6	20
İmmünoterapi	8	4	4
Karaciğer	7	6	1
Dermatoloji	35	1	34
Hindistan Karaciğer Hastası	11	9	2
Lensler	5	5	0
Cilt Segmentasyonu	4	3	1
Lenfografi	19	19	0

Çalışmada kullanılan immünoterapi, karaciğer, dermatoloji, Hindistan karaciğer hastası, lensler, cilt segmentasyonu ve lenfografi veri setleri ücretsiz olarak erişilebilen UCI Machine Learning Respository veri tabanından alınmıştır (URL-1, 2023). Erken saç beyazlaması veri seti ise, Belli ve arkadaşlarının (2015) yaptığı bir çalışmadan elde edilmiştir.

1.3. SONUÇLAR

Bu bölümde çalışmada kullanılan verilere uygulanan veri önişleme (kayıp veri tamamlama ve normalizasyon) tekniklerinin sınıflandırma performansına etkisi analiz edilmiştir. Bu kapsamda literatürden elde edilen sağlık konu alanına ait sekiz veri seti ile çalışılmıştır. İlk adım olarak veri setlerine ilişkin ön inceleme yapılmış ve kayıp veriler tespit edilmiştir. Kayıp veri problemi giderilen veri setleri için normalizasyon işlemi yapılmıştır. Hem kayıp veri tamamlama hem de normalizasyon işlemi için çeşitli teknikler denenmiş ve sonuçlar karşılaştırılmıştır. Çalışmanın devamında iki aşamalı sınıflandırma analizi yapılmıştır. 1. Temel sınıflandırıcılar ile sınıflandırma analizi, 2. Meta sınıflandırıcılar ile sınıflandırma analizi. Çalışmanın son aşamasında sonuçlar karşılaştırılmış ve en iyi yöntem seçimi için yorumlanmıştır.

1.3.1. Kayıp Verilerin Tamamlanması İçin Elde Edilen RMSE Sonuçları

Kök ortalama kare hatası (RMSE- Root Mean Square Error) bir performans ölçüsüdür ve genellikle modelin doğru cevabı vermekten ne kadar uzak olduğunu gösterir. Bazen tahmin edilen değerin kalitesini ölçmek için gerçek değer ile aralarında bir tür fark bulunması gerekir. RMSE, tahmin edilen değerlerin gerçek değerlere ne kadar yakın

olduğunun bir ölçüsünü sağlar. Daha düşük bir RMSE daha iyi model uyumunu gösterir. RMSE, Denklem 8 ile gösterildiği üzere tahmin edilen ve gerçek değerler arasındaki kare farkların ortalamasının karekökü olarak hesaplanır.

$$RMSE = \sqrt{\sum_{t=1}^{n} \frac{(a_i - p_i)^2}{n}} \tag{1.8}$$

a_i: i. gözlem için gerçek değer
p_i: i. gözlem için tahmin değeri

Bu çalışmada sekiz veri seti ele alınmış olup bu veri setlerinden altısı kayıp veri içermemektedir. Erken saç beyazlaması veri setinde bulunan kayıp veri sayısının tüm veri setine oranı %0.014 olarak hesaplanmıştır. Ayrıca kayıp verilerin aynı gözleme ait olduğu görülmüştür. Bu nedenle kayıp veri içeren gözlemlerin silinmesinin bu veri seti için daha uygun olacağı düşünülmüştür. Dermatoloji veri setinde ise kayıp oranı %0.022 olarak hesaplanmıştır. Kayıp veriler farklı gözlemlere ait olduğundan bu veri seti için veri silme işlemi tercih edilmemiştir. Ortalama, medyan, kNN, Rpart ve Mice yöntemleri ile kayıp veri tamamlama işlemi gerçekleştirilmiştir. Veri seti için hangi sonucun daha güvenilir olduğunu test etmek için RMSE değerleri R programı yardımıyla hesaplanmıştır. Tablo 1.4'te kayıp veri içeren gözlemlerin veri setine oranları ve kayıp değerlerin tamamlanması için RMSE sonuçları verilmiştir. Sonuçlar kNN yönteminin daha başarılı olduğunu göstermekte olduğundan çalışmada bu yöntem ile oluşturulan veri seti kullanılmıştır.

Tablo 1.4. RMSE sonuçları

Veri seti	Toplam Kayıp Oranı (%)	Kayıp Değerlerin Tamamlanması için RMSE Sonuçları				
		Ortalama	Medyan	kNN	Rpart	Mice
Erken Saç Beyazlaması	0.014	Kayıp veriler aynı gözlemde olduğundan, kayıp veri içeren gözlemler silinmiştir.				
İmmünoterapi	Kayıp veri bulunmamaktadır.					
Karaciğer	Kayıp veri bulunmamaktadır.					
Dermatoloji	0.022	15.062	15.059	**13.483**	15.314	14.179
Hindistan Karaciğer Hastası	Kayıp veri bulunmamaktadır.					
Lensler	Kayıp veri bulunmamaktadır.					
Cilt Segmentasyonu	Kayıp veri bulunmamaktadır.					
Lenfografi	Kayıp veri bulunmamaktadır.					

1.3.2. Veri Normalizasyonu Sonuçları

Veri setindeki değişkenlere ilişkin değer aralıklarının anlamlı derecede birbirinden farklı olduğu durumlarda normalizasyon işlemi uygulanır. Normalizasyon işleminin, değişkenlerin ortalama ve varyanslarının birbirlerine yaklaştırarak yüksek ortalamaya (ya da varyansa) sahip değişkenlerin diğer değişkenler üzerindeki baskısını azalttığı bilinmektedir. Bu nedenle yapılan analizlerde değişkenlerin sonuçlar üzerindeki etkisini standartlaştırmak amacıyla veriler normalize edilir. Bazı çalışmalarda ise, normalizasyon işleminin bilgi kaybına neden olacağı düşüncesi ile normalizasyon işlemi yapılmadan analize devam edilir (Jaiswal, 2024; URL-2, 2023). Bu çalışmada, ele alınan sekiz veri setine farklı normalizasyon işlemleri (min-max, sigmoid, unit, decimal, softnorm) R programı yardımıyla uygulanmıştır. Normalizasyon işleminin sınıflandırma performansına etkisini araştırmak amacıyla hem orijinal veri setleri hem de normalize edilmiş veri setleri Random

Forest (RF) algoritması ile sınıflandırılmıştır. Tablo 5 normalizasyon ve RF ile sınıflandırma sonuçlarını göstermektedir. Buna göre, erken saç beyazlaması veri seti için sigmoid, dermatoloji veri seti için min-max ve unit, Hindistan karaciğer hastası veri seti için decimal, cilt segmentasyonu veri seti için softnorm ve iki veri seti (karaciğer ve immünoterapi) için orijinal gözlemler ile en iyi sınıflandırma performansına ulaşılmıştır. Lenfografi ve lensler veri setleri kategorik verilerden oluştuğundan, bu veri setlerine normalizasyon işlemi uygulanmamıştır. Çalışmanın devamında, RF ile sınıflandırma sonucu en iyi olan veri seti analize dahil edilmiştir.

Tablo 5. Veri Normalizasyonu sonuçları

Veri seti	**RF için En İyi Sonucu Veren Normalizasyon Yöntemi**	
	Yöntem	**Doğruluk (%)**
Erken Saç Beyazlaması	Sigmoid	94.560
Karaciğer	Orijinal Gözlemler	73.044
İmmünoterapi	Orijinal Gözlemler	84.444
Dermatoloji	Min-Max, Unit	97.486
Hindistan Karaciğer Hastası	Decimal	71.698
Cilt Segmentasyonu	Softnorm	99.963
Lenfografi	Tüm veriler kategorik olduğundan normalizasyon işlemi yapılmamıştır.	
Lensler	Tüm veriler kategorik olduğundan normalizasyon işlemi yapılmamıştır.	

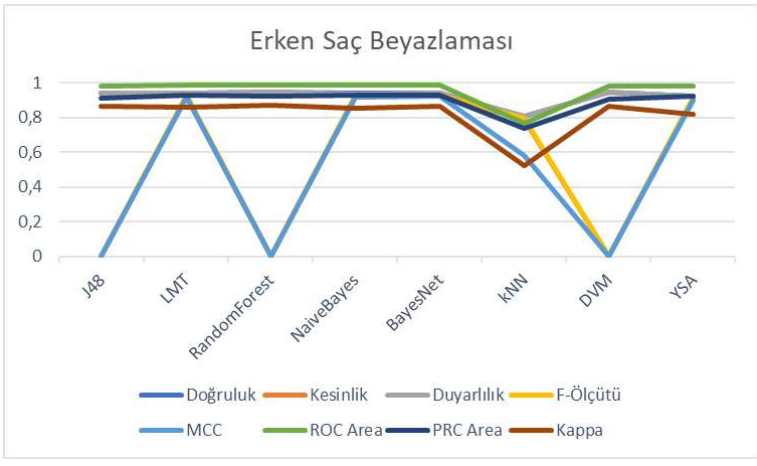
1.3.3. Sınıflandırma Sonuçları

Sınıflandırma analizi, sağlık sektöründe diğer alanlara göre daha geç uygulanmaya başlamasına rağmen son yıllarda sağlık alanında çok çeşitli uygulamalar ile karşımıza çıkmaktadır. Özellikle tanılar ve tedavi seçenekleri konusunda yapılan çalışmaların verimliliğini arttırmakta ve hasta sonuçlarını iyileştirmeye yardımcı olmaktadır.

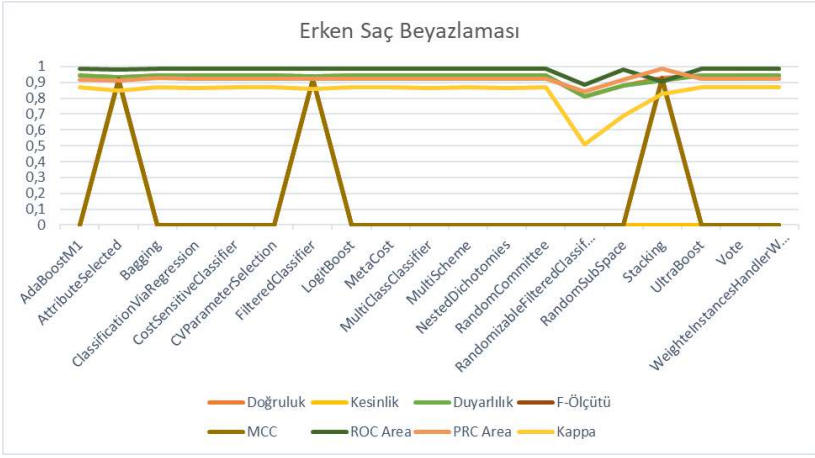
Sınıflandırma analizinin sağlık alanındaki öneminden dolayı bu çalışmada sağlık alanından seçilmiş veri setleri ele alınmıştır. Elde edilen algoritma sonuçları karşılaştırılmış ve yorumlanmıştır.

Bu çalışmada temel ve meta algoritmalar ile sınıflandırma yapılmış olup, makine öğrenmesi ile sınıflaması yapılmak istenen başka veriler için seçilecek sınıflandırma algoritmalarının ne derecede etkin sonuç verdikleri ortaya konmuştur.

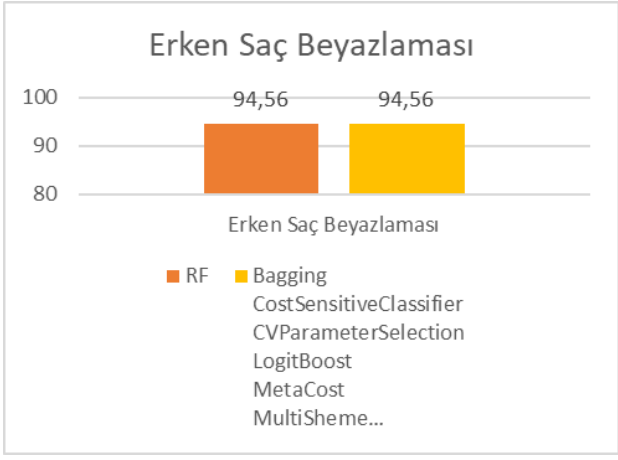
Şekil 1.2-1.4 ile erken saç beyazlaması veri seti için temel ve meta algoritmalar için sınıflandırma sonuçları verilmiştir. Elde edilen sonuçlar, temel sınıflandırıcı olan Random Forest algoritması ile meta sınıflandırıcılardan Bagging, Cost Sensitive Classifier, CV Parameter Selection, LogitBoost, Meta Cost, Multi Sheme yöntemlerinin %94,56 doğruluk oranıyla en yüksek performansı sergilediğini göstermektedir.



Şekil 1.2. Erken Saç Beyazlaması veri seti için temel algoritma sonuçları

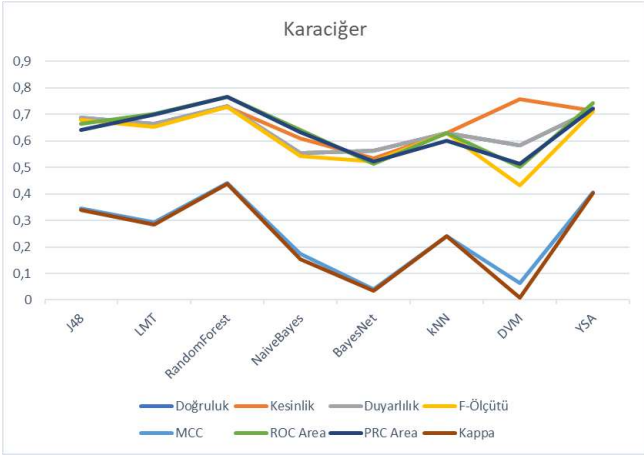


Şekil 1.3. Erken Saç Beyazlaması veri seti için meta algoritma sonuçları

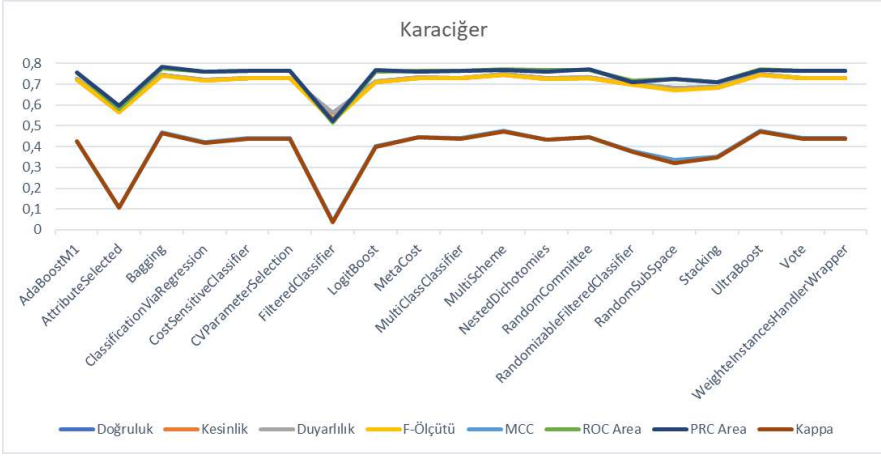


Şekil 1.4. Erken Saç Beyazlaması veri seti için en iyi temel ve meta algoritma

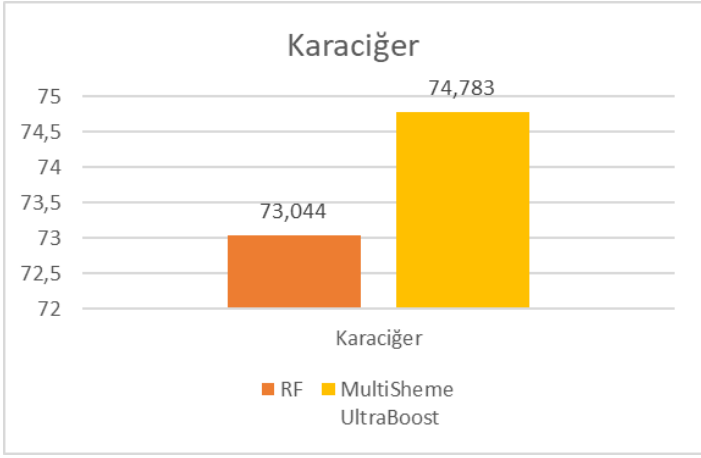
Şekil 1.5-1.7 ile karaciğer veri seti için temel ve meta algoritmalar için sınıflandırma sonuçları verilmiştir. Sonuçlara göre, temel sınıflandırıcı olan Random Forest algoritması ile %73,044 ve meta sınıflandırıcılardan MultiSheme ve UltraBoost yöntemleri ile %74,783 doğruluk oranı elde edilmiştir.



Şekil 1.5. Karaciğer veri seti için temel algoritma sonuçları

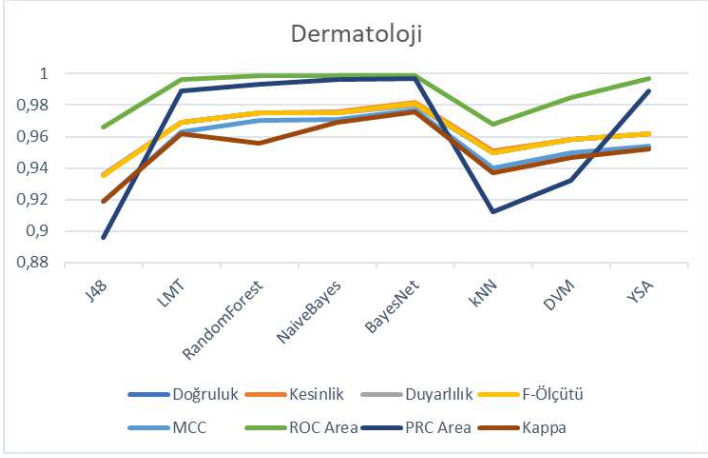


Şekil 1.6. Karaciğer veri seti için meta algoritma sonuçları

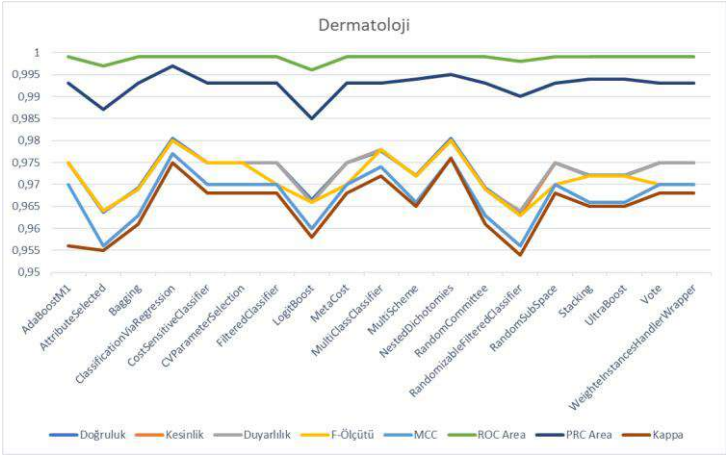


Şekil 1.7. Karaciğer veri seti için en iyi temel ve meta algoritma

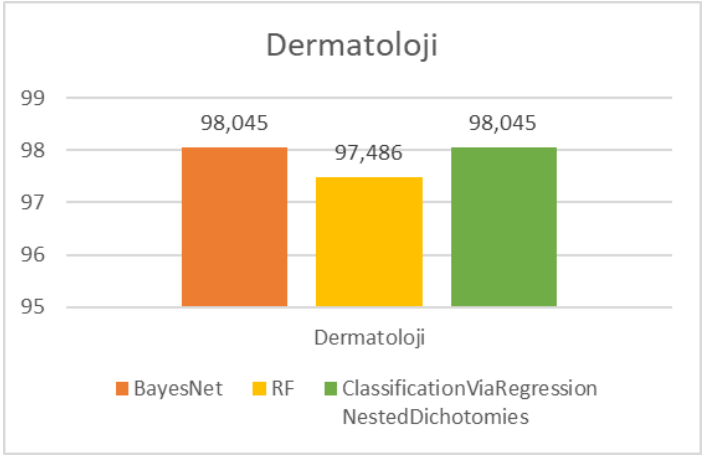
Şekil 1.8 - 1.10 ile dermatoloji veri seti için temel ve meta algoritmalar için sınıflandırma sonuçları verilmiştir. Sonuçlara göre, temel sınıflandırıcı olan Bayes Net ile %98,045, Random Forest algoritması ile %97,486 ve meta sınıflandırıcılardan ClassificationViaRegression ve Nested Dishotomies yöntemleri ile %98,045 doğruluk oranı elde edilmiştir.



Şekil 1.8. Dermatoloji veri seti için temel algoritma sonuçları

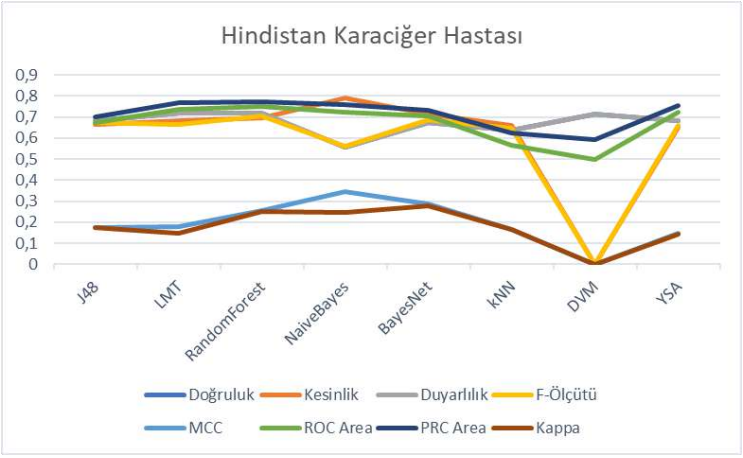


Şekil 1.9. Dermatoloji veri seti için meta algoritma sonuçları

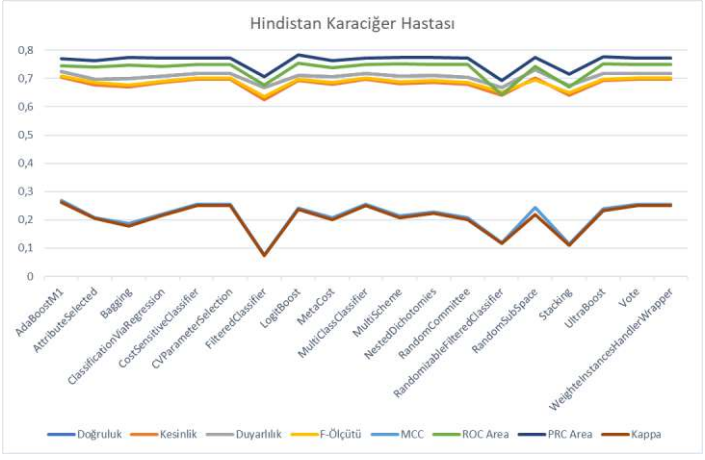


Şekil 1.10. Dermatoloji veri seti için en iyi temel ve meta algoritma

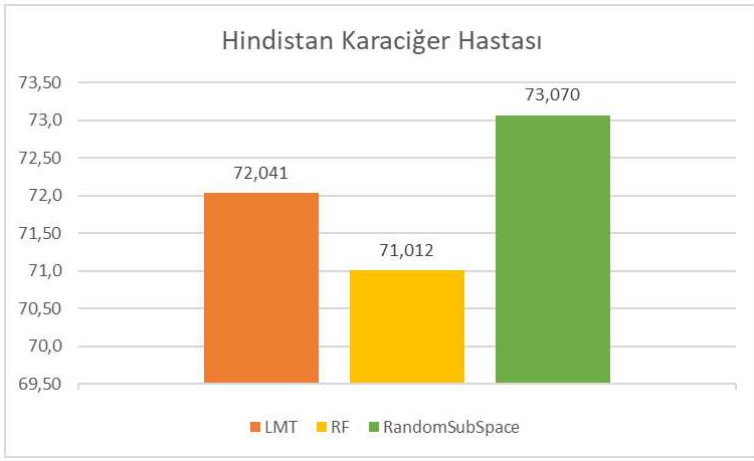
Şekil 1.11-1.13 ile Hindistan karaciğer hastası veri seti için temel ve meta algoritmalar için sınıflandırma sonuçları verilmiştir. Sonuçlara göre, temel sınıflandırıcı olan LMT ile %72,041 ve meta sınıflandırıcılardan RandomSubSpace yöntemi ile %73,070 doğruluk oranı elde edilmiştir.



Şekil 1.11. Hindistan Karaciğer Hastası veri seti için temel algoritma sonuçları

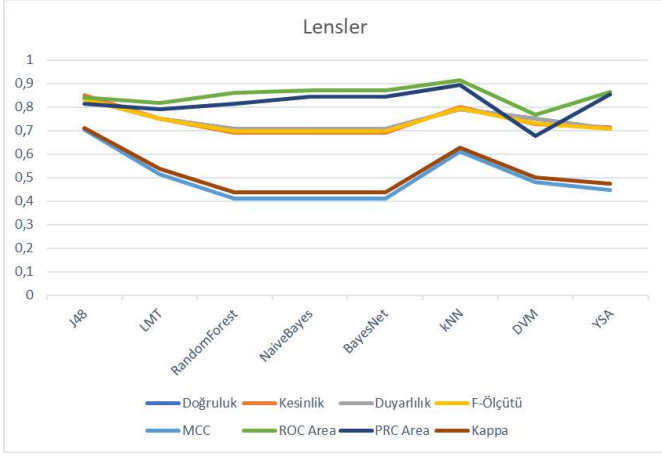


Şekil 1.12. Hindistan Karaciğer Hastası veri seti için meta algoritma sonuçları

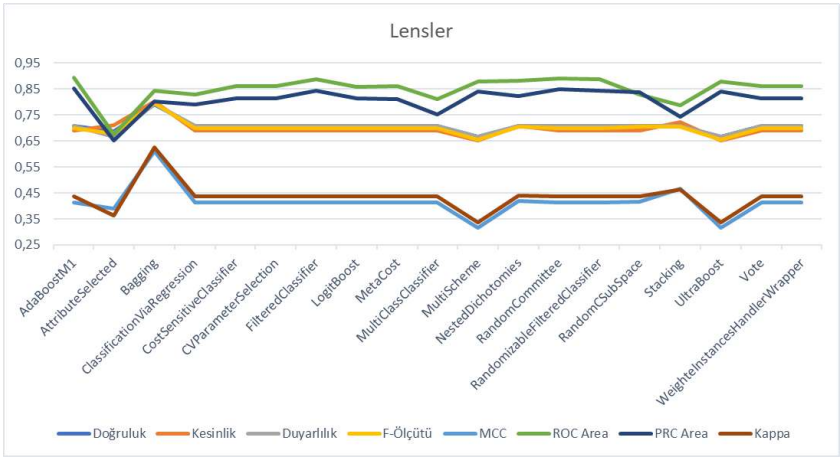


Şekil 1.13. Hindistan Karaciğer Hastası veri seti için en iyi temel ve meta algoritma

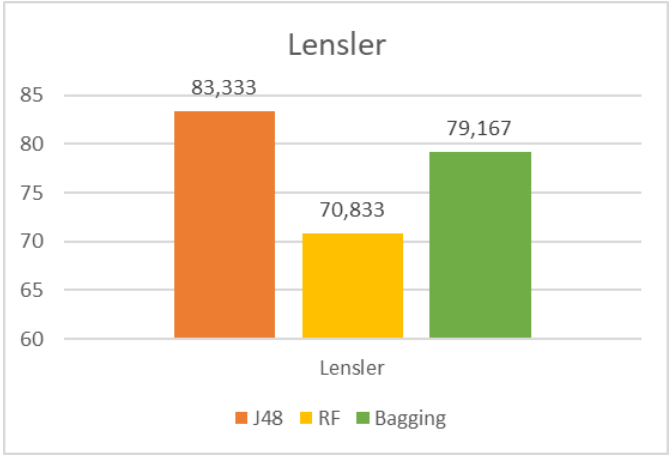
Şekil 1.14-1.16 ile lensler veri seti için temel ve meta algoritmalar için sınıflandırma sonuçları verilmiştir. Sonuçlara göre, temel sınıflandırıcı olan J48 ile %83,333, Random Forest ile %70,833 ve meta sınıflandırıcılardan Bagging yöntemi ile %79,167 doğruluk oranı elde edilmiştir.



Şekil 1.14. Lensler veri seti için temel algoritma sonuçları



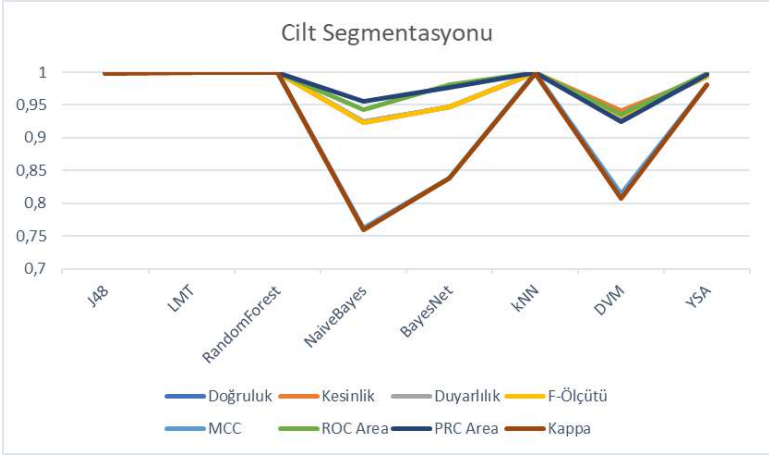
Şekil 1.15. Lensler veri seti için meta algoritma sonuçları



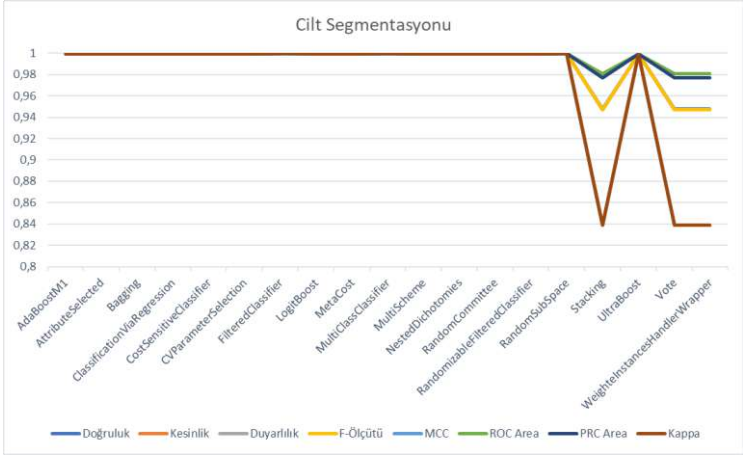
Şekil 1.16. Lensler veri seti için en iyi temel ve meta algoritma

Şekil 1.17-1.19 ile cilt segmentasyonu veri seti için temel ve meta algoritmalar için sınıflandırma sonuçları verilmiştir. Sonuçlara göre, temel sınıflandırıcı olan Random Forest ile %99,963 ve meta

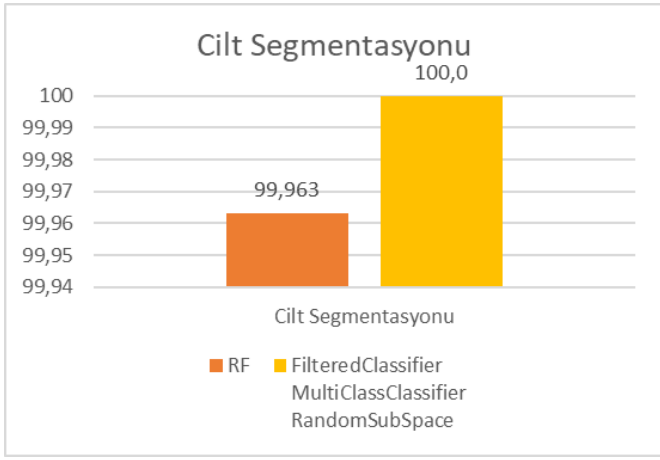
sınıflandırıcılardan FilteredClassifier, MultiClassClassifier ve RandomSunSpace yöntemleri ile %100 doğruluk oranı elde edilmiştir.



Şekil 1.17. Cilt segmentasyonu veri seti için temel algoritma sonuçları

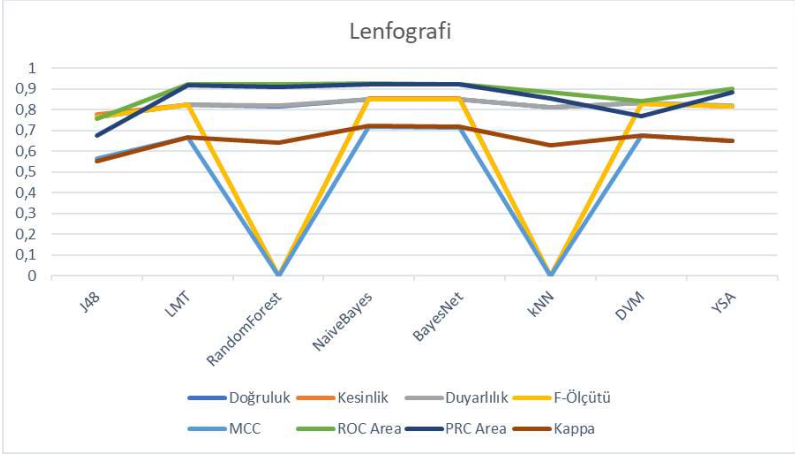


Şekil 1.18. Cilt segmentasyonu veri seti için meta algoritma sonuçları

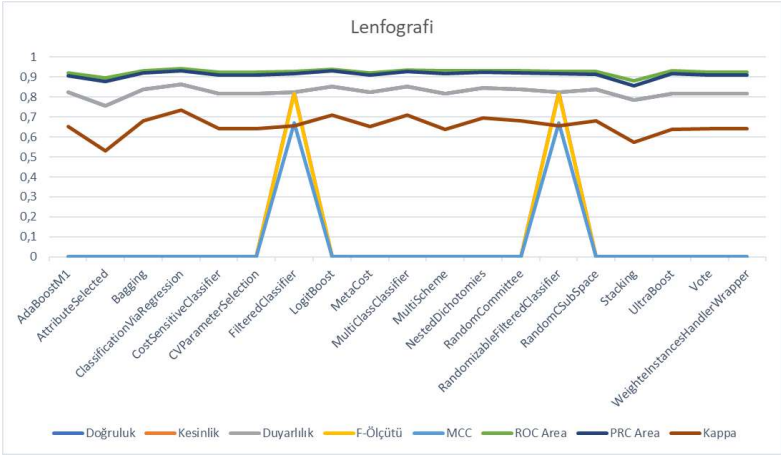


Şekil 1.19. Cilt segmentasyonu veri seti için en iyi temel ve meta algoritma

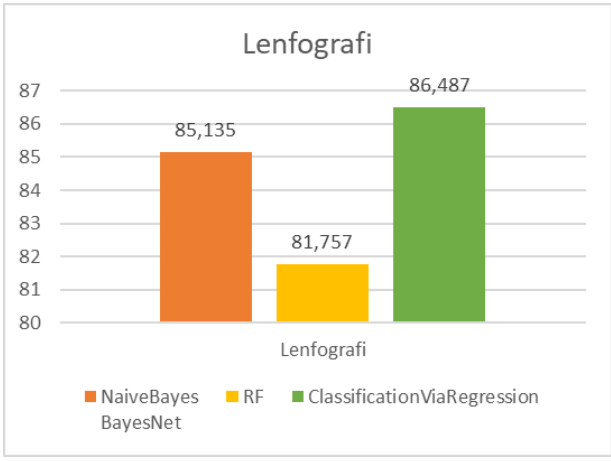
Şekil 1.20-1.22 ile lenfografi veri seti için temel ve meta algoritmalar için sınıflandırma sonuçları verilmiştir. Sonuçlara göre, temel sınıflandırıcı olan Naive Bayes ve Bayes Net ile %85,135, Random Forest ile %81,757 ve meta sınıflandırıcılardan ClassificationViaRegression yöntemi ile %86,487 doğruluk oranı elde edilmiştir.



Şekil 1.20. Lenfografi veri seti için temel algoritma sonuçları



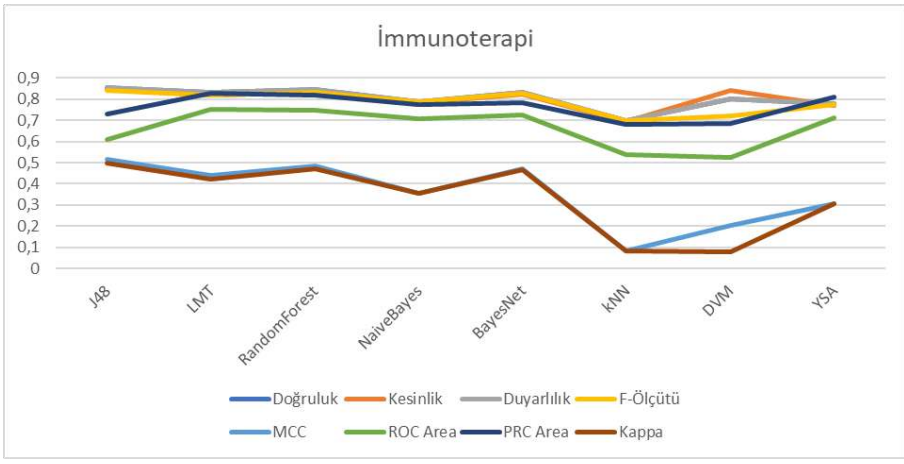
Şekil 1.21. Lenfografi veri seti için temel algoritma sonuçları



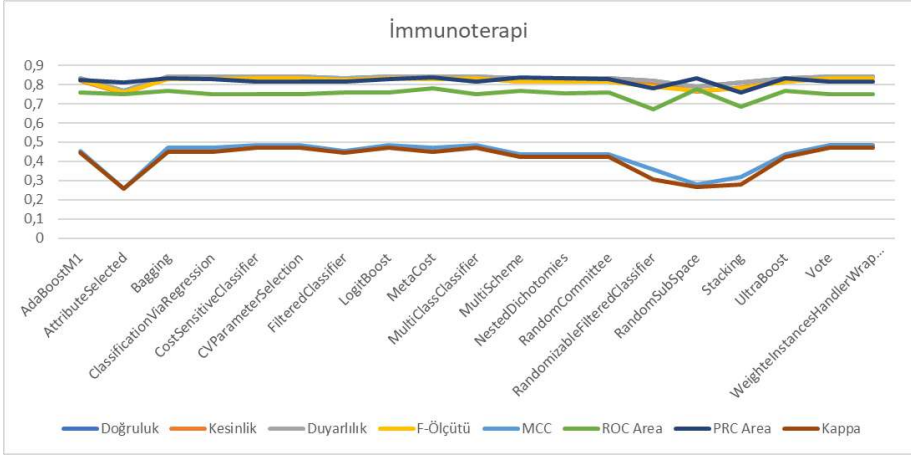
Şekil 1.22. Lenfografi veri seti için en iyi temel ve meta algoritma

Şekil 1.23-1.25 ile immünoterapi veri seti için temel ve meta algoritmalar için sınıflandırma sonuçları verilmiştir. Sonuçlara göre, temel sınıflandırıcı olan J48 ile %85,555, Random Forest ile %84,444

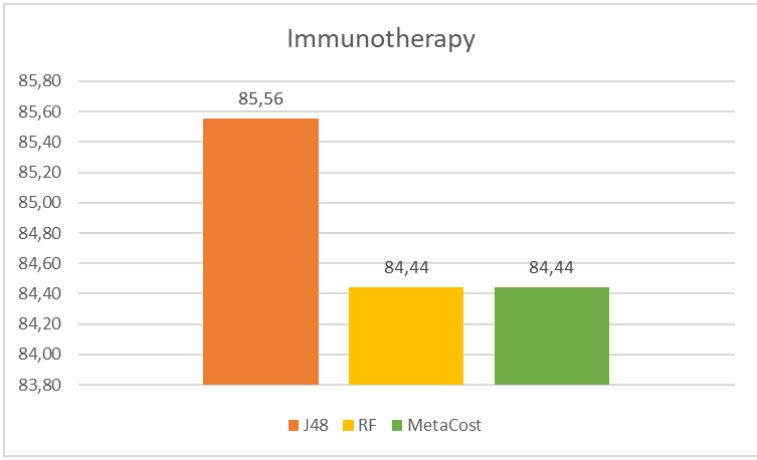
ve meta sınıflandırıcılardan MetaCost yöntemi ile %84,444 doğruluk oranı elde edilmiştir.



Şekil 1.23. İmmünoterapi veri seti için temel algoritma sonuçları



Şekil 1.24. İmmünoterapi veri seti için temel algoritma sonuçları



Şekil 1.25. İmmünoterapi veri seti için en iyi temel ve meta algoritma

Çalışmanın veri önişleme aşaması tamamlandıktan sonra her bir veri seti için temel ve meta yöntem seçimi yapılmıştır. Bu aşamada amaç her bir veri seti için tek tek doğruluk değeri elde etmek değildir.

Tablo 1.5. Sınıflandırma sonuçlarının Karşılaştırılması

	Temel		RF	Meta	
Dermatoloji	BayesNet	98,045	97,486	ClassificationViaRegression NestedDichotomies	98,045
Hindistan karaciğer	LMT	72,041	71,012	RandomSubSpace	73,070
Karaciğer	RF	73,044	73,044	MultiScheme UltraBoost	74,783
Lensler	J48	83,333	70,833	Bagging	79,167
Lenfografi	NaiveBayes BayesNet	85,135	81,757	ClassificationViaRegression	86,487
İmmünoterapi	J48	85,555	84,444	MetaCost	84,444
Erken saç beyazlaması	RF	94,56	94,56	Bagging CostSensitiveClassifier CVParameterSelection LogitBoost MetaCost MultiScheme RandomCommittee UltraBoost Vote WeightedInstancesHandlerWrapper	94,56
Cilt segmentasyonu	RF	99,963	99,963	FilteredClassifier MultiClassClassifier RandomSubSpace	100,00

Asıl amaç, genel bir çıkarım yapılıp yapılamayacağının sorgulanmasıdır. Bu doğrultuda veri setlerine uygulanan temel ve meta yöntemlerden en iyi doğruluk değerine ulaşan algoritma kaydedilmiştir. RF algoritmasının çalışma prensibi, meta yöntemlere benzemektedir. Bu nedenle RF algoritmasına ilişkin sonuçlar, ayrı bir kategori olarak incelenmiştir. Elde edilen sonuçlar Tablo 1.5'te yer almaktadır.

1.4. TARTIŞMA

Bu çalışma, sağlık alanı verilerinin sınıflandırılmasında temel ve meta yöntem seçiminin yanı sıra yapılan veri ön işleme tekniklerinin veri setine uyum sağlayacak şekilde seçimini vurgulamayı amaçlamıştır. Sınıflandırma modelleri incelendiğinde, temel veya meta yöntem seçimine ilişkin net bir bilgi olmadığı kanısına varılmıştır. Ancak, bu yoruma ulaşmadan önce veri önişleme süreçleri üzerine titiz bir çalışma yürütülmüştür. Bu sonuçların, verilerin sınıflandırılmasında kullanılan yöntemlerin sonuçları üzerinde etki yarattığı bilinmektedir.

Bu çalışmada sağlık alanından seçilmiş veri setleri kullanılmıştır. Veri setlerinde var olan kayıp veri ve normalizasyondan kaynaklı baskılı veri problemleri için önişleme tekniklerinin bir uygulaması yapılmıştır. Önişleme tekniklerinin veri setine uygunluğunu ölçmek için farklı yöntemler denenmiştir. Kayıp veri probleminin giderilmesi için ortalama, medyan, kNN, Rpart ve Mice yöntemleri kullanılmıştır. Kayıp veri seti için uygun yöntem RMSE değeri hesaplanarak tespit edilmiştir. Erken saç beyazlaması için kayıp verilerin silinmesi, dermatoloji verileri için ise kNN yöntemi en uygun yöntemlerdir. Veri normalizasyon işlemi genellikle farklı şekillerde ölçümlenmiş veriler arasındaki baskıyı ortadan kaldırmak için kullanılmaktadır. Çalışmada kullanılan veriler için normalizasyon

tekniğinin sınıflandırma analizi kapsamında uygun olup olmayacağının araştırılması için min-max, sigmoid, unit, decimal, softnorm ve ham veriler ile RF sınıflandırma sonuçları elde edilmiştir. Erken saç beyazlaması verileri için sigmoid, Hindistan Karaciğer Hastası verileri için decimal, cilt segmentasyonu için softnorm ve diğer veriler için orijinal gözlemler en yüksek doğruluğu vermiştir. Lensler ve lenfografi veri setleri sadece kategorik değişkenlerden oluştuğundan normalizasyon işlemi dışında tutulmuştur.

Çalışmanın sonuçları, etkili bir sınıflandırma analizi yapmak için temel ve meta algoritma seçimine ilişkin bir yargı bulunmadığını göstermektedir. Uygun model seçimi yapmak araştırmacının büyük emek harcadığı bir aşamadır. Temel ve meta öğrenme modellerinin kullanımına ilişkin her ne kadar genel bir çıkarım yapmak mümkün olmasa da model seçimi için zaman ve kaynak tasarrufu sağlanması açısından sonuçlar yol gösterici olabilir.

Bu çalışmada, sınırlı sayıda ele alınan veri kümesi kullanılarak elde edilen sonuçlar üzerine odaklanılmıştır. Daha fazla sayıda veri kümesinin kullanılması, modellerin performansının daha kapsamlı bir şekilde değerlendirilmesine olanak sağlayacaktır. Bu nedenle gelecekte yapılacak çalışmaların daha fazla veri seti ile tekrarlanması temel ya da meta algoritma seçimine ilişkin daha ayrıntılı bir çıkarımda bulunulmasını sağlayacaktır.

KAYNAKLAR

Akram, V. K. ve Taser, P. Y. (2020) Telsiz Duyarga Ağlarda Bizans Saldırılarının Topluluk Öğrenme-tabanlı Tespiti, DEÜ FMD, 22(66), 905-918.

Arora, T. ve Dhir, R. (2017) Correlation-Based Feature Selection and Classification via Regression of Segmented Chromosomes Using Geometric Features, Medical & Biological Engineering & Computing, 55(5), 733-745.

Balogun, A. O., Balogun, A. M., Sadiku, P. O. ve Adeyemo, V. E. (2017) Heterogeneous Ensemble Models For Generic Classification, Scientific Annals of Computer Science, 15(1), 92-98.

Başarslan, M. S. (2017) Telekomünikasyon Sektöründe Müşteri Kayıp Analizi, Düzce Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Düzce.

Belli, A. A., Etgu, F., Ozbas Gok, S., Kara, B. ve Dogan, G. (2016) Risk Factors for Premature Hair Graying in Young Turkish Adults. *Pediatric Dermatology*, *33*(4), 438-442.

Breiman, L. (1996) Bagging Predictors, Machine Learning, 24, 123-140.

Buuren, S. ve Groothuis-Oudshoorn, K., (2011) Mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software, 45.

Coşkun, C. ve Baykal, A. (2011) Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması, Akademik Bilişim, 11, 51-58.

Cutler, A., Cutler, D. R. ve Stevens J. R. (2011) Random Forest, Machine Learning, 45(1), 157-176.

Domingos, P. (1999) Metacost: A General Method for Making Classifiers Cost-Sensitive, Fifth International Conference on Knowledge Discovery and Data Mining, 155-164.

Dong, L., Frank, E. ve Kramer, S. (2005) Ensembles of Balanced Nested Dichotomies for Multi-class Problems, PKDD, 84-95.

Durmuş, B., Güneri, Ö.İ., İncekırık, A. (2023) Sosyal, İnsan ve İdari Bilimlerde İleri ve Çağdaş Çalışmalar, 1986-2021 Yılları Arasında Sağlık Alanında Yapılan Veri Madenciliği Çalışmaları, Edt: Halim Akbulut ve Rıfat Işık, Duvar Yayınları, 752-768.

Elmas, Ç. (2003) Yapay Zekâ Uygulamaları, Seçkin Yayıncılık, Ankara.

Erken, Ş. ve Şenyay, L. (2023) Makine Öğrenmesi ile Eksik Veri Tamamlama Yöntemlerinin Sınıflandırma Performansına Etkileri, Kayseri Üniversitesi Sosyal Bilimler Dergisi, 5(1), 51-71.

Frank, E. ve Kramer, S. (2004) Ensembles of Nested Dichotomies for Multi-Class Problems, Twenty-First International Conference On Machine Learning.

Friedman, J., Hastie, T. ve Tibshirani, R. (2000) Additive Logistic Regression: A Statistical View of Boosting, Ann Stat, 28(2), 337-407.

Gautam R., Vanga S., Ariese F. ve Umapathy S. (2015) Review of Multidimensional Data Processing Approaches for Raman and Infrared Spectroscopy, EPJ Techniques and Instrumentation, 2, 8.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. ve Witten, I. (2009) The Weka Data Mining Software: An Update, ACM SIGKDD Explorations Newsletter, 11(1), 10-18.

Ho, T. K. (1998) The Random Subspace Method for Constructing Decision, IEEE Transactions on Pattern Analysis and Machine Intelligence, Lucent Tech, No:l, AT&T Bell Labs., Murray Hill, 20(8), 832-844.

Jaiswal, S. (2024) What is Normalization in Machine Learning? A Comprehensive Guide to Data Rescaling, DataCamp.

Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S. ve Pintelas, P. (2007) A Wrapper for Reweighting Training Instances for Handling Imbalanced Data Sets, Boukis, C., Pnevmatikakis, A., Polymenakos, L. (eds) Artificial Intelligence and Innovations, Theory to Applications, IFIP The International Federation for Information Processing, 247. Springer, Boston, MA.

Kargar, K., Safari, M. J. S. ve Khosravi, K. (2021) Weighted Instances Handler Wrapper and Rotation Forest-Based Hybrid Algorithms for Sediment Transport Modeling, Journal of Hydrology, 598, 126452.

Kawade, D. R. ve Oza, K. S. (2015) SMS Spam Classification Using WEKA, International Journal of Electronics Communication and Computer Technology, 5(ICICC), 43-47.

Kégl, B. (2009) Introduction to Adaboost, 11-14.

Kim, J., Choi, K., Kim, G. ve Suh, Y. (2012) Classification Cost: An Empirical Comparison Among Traditional Classifier, Cost-Sensitive Classifier, and Metacost, Expert Systems with Applications, 39(4), 4013-4019.

Kittler, J., Hatef, M., Duin, R. P. W. ve Matas, J. (1998) On Combining Classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), 226-239.

Koçali, M. (2019) Türkiye'de İklimin Enerji Piyasasına Olan Etkileri Ve Bu Etkilerin Tahmin Edilebilirliği, Türk Hava Kurumu Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi

Kuncheva, L. I. (2004) Classifier Ensembles for Changing Environments, International Workshop on Multiple Classifier Systems, Heidelberg: Springer Berlin Heidelberg, 1-5.

Landwehr, N., Hall, M. ve Frank, E. (2005) Logistic Model Trees, Machine Learning, 59, 161-205.

Larose, D. T. (2014) Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons.

Li, H., Xiong, L., Ohno-Machado, L. ve Jiang, X. (2014) Privacy Preserving Rbf Kernel Support Vector Machine, BioMed Research International, 2014, 827371.

Little, R. J. A. ve Rubin, D. B. (2014) Incomplete Data, John Wiley & Sons, Inc.

Makhtar, M., Neagu, D. C. ve Ridley, M. J. (2011). Comparing Multi-Class Classifiers: On The Similarity of Confusion Matrices for Predictive Toxicology Applications, In International Conference on Intelligent Data Engineering and Automated Learning, Berlin, Heidelberg: Springer Berlin Heidelberg, 252-261.

Malik, M. R., Liu, Y. ve Shaikh, S. (2019) Analysis of Software Deformity Prone Datasets with Use of AttributeSelectedClassifier, International Journal of Advanced Computer Science and Applications, 10(7).

Michie, D., Spiegelhalter, D.J. ve Taylor C.C. (1994) Machine Learning, Neural and Statistical Classification, Ellis Horwood Limited.

Mustakim, N. A., Abdul Karim, Z. H., Sauid, M. K., Mokhtar, N., Hassan, Z. ve Mohamad Roseli, N. H. (2024) Gender-Based Analysis of Online Shopping Patterns on Shopee in Malaysia: A J48 Decision Tree Approach, Information Management and Business Review, 16(3(I)S), 844-854.

Özkan, Y. (2008) Veri Madenciliği Yöntemleri, Papatya Yayıncılık Eğitim.

Reddy, B. H. M., Reddy, S. V. ve Sarojamma, B. (2021) Data Mining Techniques for Estimation of Wind Speed Using WEKA, International Journal of Computer Sciences and Engineering, 9(9), 49-53.

Ruan, Y. X., Lin, H. T. ve Tsai, M. F. (2014) Improving Ranking Performance with Cost-Sensitive Ordinal Classification via Regression, Information Retrieval, 17, 1-20.

Salman, H. A., Kalakech, A. ve Steiti, A. (2024) Random Forest Algorithm Overview, Babylonian Journal of Machine Learning, 69-79.

Salzberg, S.L. (1994) C4.5: Programs for Machine Learning by J. Ross Quinlan, Morgan Kaufmann Publishers, Machine Learning, 16, 235-240.

Santosh, S. V. (2015) İlk Okuyucunun Ötesine Geçmek: Meme Ultrason Teşhisinde Maliyet ve Performansı Optimize Etmek İçin Bir Makine Öğrenimi Metodolojisi, Ultrason Med Biyo., 41(12), 3148-3162.

Sikora, R. ve Al-laymoun, O. H. (2014) A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms, Journal of International Technology and Information Management, 23(1).

Skurichina, M. ve Duin, R. P. W. (2002) Bagging, Boosting and The Random Subspace Method for Linear Classifiers Pattern, Analysis & Applications, 5, 121-135.

Sun, P., Reid, M. D. ve Zhou, J. (2014) An Improved Multiclass Logitboost Using Adaptive-One-Vs-One, Machine Learning, 97, 295-326.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N. ve Asadpour, M. (2020) Boosting Methods for Multi-Class Imbalanced Data Classification: An Experimental Review, Journal Big Data, 7, 70.

Trevor, H., Tibshirani, R. ve Friedman, J. (2011) The Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd ed.), Springer, International Publishing.

Tu, M. C., Shin D. ve Shin, D. (2009) A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms, Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 183-187.

URL-1: https://archive.ics.uci.edu/ Erişim Tarihi: 12.10.2023.

URL-2: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-normalization/ Erişim Tarihi: 30.05.2025

Van Harmelen, F., Lifschitz, V. ve Porter, B. (2008) Handbook of Knowledge Representation, Elsevier.

Witten, I. H. ve Frank, E. (2005) Data Mining: Practical Machine Learning Tools and Techniques (2nd ed.), Morgan Kaufmann, San Francisco.

Yıldırım, P., Uludağ, M. ve Görür, A. (2008) Hastane Bilgi Sistemlerinde Veri Madenciliği, Akademik Bilişim, 30 Ocak-01 Şubat, Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, 429-434.

BÖLÜM 2:

MUTLULUK GÖSTERGELERİNİN SEYREK TEMEL BİLEŞEN ANALİZİ

Nida ORUÇ[3]
Muzaffer GÖZTAŞ[4]
Doğan YILDIZ[5]

Özet

Bu çalışma, Dünya Mutluluk Raporu (2023) verileri kullanılarak, ülkelerin mutluluk göstergeleri doğrultusunda sosyo-ekonomik benzerliklerini değerlendirmeyi amaçlamaktadır. Bu kapsamda, veri setinde yer alan altı temel göstergeye Seyrek Temel Bileşenler Analizi (STBA) uygulanmış ve ülkelerin refah düzeylerini yansıtan iki boyutlu bileşen düzlemi elde edilmiştir. STBA sonucunda elde edilen bileşenler temelinde, K-Ortalamalar algoritması kullanılarak ülkeler benzer gösterge profillerine göre gruplandırılmıştır. Analiz bulguları, ülkelerin benzer sosyo-ekonomik özellikler çerçevesinde kümelendiğini ve özellikle Kişi Başına GSYİH (Log), Sosyal Destek, Sağlıklı Yaşam Beklentisi, Yaşam Seçimi Özgürlüğü, Cömertlik ve Yolsuzluk Algısı değişkenleri açısından anlamlı biçimde ayrıştığını göstermektedir. Çalışmanın sonunda, bu yöntemlerin ülkeler arası karşılaştırmalarda karar vericiler için daha etkili analiz araçları sunduğu vurgulanmıştır.

[3] Arş. Gör., İstanbul Topkapı Üniversitesi, İktisadi, İdari ve Sosyal Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü, İstanbul, Türkiye, nidaoruc@topkapi.edu.tr, ORCID ID: 0009-0005-0230-2234
[4] Arş. Gör., Yıldız Teknik Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, İstanbul, Türkiye, muzaffer.goztas@yildiz.edu.tr, ORCID ID: 0009-0001-7933-901X
[5] Dr. Öğr. Üyesi, Yıldız Teknik Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, İstanbul, Türkiye, dyildiz@yildiz.edu.tr, ORCID ID: 0000-0001-7430-2368

SPARSE PRINCIPAL COMPONENT ANALYSIS OF HAPPINESS INDICATORS

Abstract

This study aims to evaluate the socio-economic similarities of countries based on happiness indicators, using data from the 2023 World Happiness Report. In this context, Sparse Principal Component Analysis (SPCA) was applied to six core indicators included in the dataset, resulting in a two-dimensional component space that reflects countries' levels of well-being. Based on the component scores obtained through SPCA, countries were grouped according to their similar indicator profiles using the K-Means clustering algorithm. The analysis results reveal that countries clustered around similar socio-economic characteristics and diverged significantly, particularly in terms of Log GDP per Capita, Social Support, Healthy Life Expectancy, Freedom to Make Life Choices, Generosity, and Perceptions of Corruption. The study concludes that these methods offer decision-makers more effective analytical tools for conducting cross-country comparisons.

2.1. GİRİŞ

Ülkelerin sosyal, ekonomik ve bireysel refah düzeylerinin karşılaştırılabilir bir şekilde değerlendirilmesi, bazı akademik alanlarda önem taşımaktadır. Özellikle soyut kavramlar olan mutluluk, yaşam kalitesi ve bireylerin öznel iyi olma halleri, son zamanlarda nicel ölçümlerle daha fazla yer bulmaya başlamıştır. Bu bağlamda, sosyoekonomik gelişmişlik de bireylerin eşit imkânlara erişimi, gelir adaletinin sağlanması, yaşam standartlarının yükseltilmesi ve ekonomik-sosyal yapıların sürdürülebilirliği açısından dikkate değer bir ölçüt olarak öne çıkmaktadır. Aynı zamanda bölgeler arası dengenin sağlanmasına katkı sunması nedeniyle de kapsamlı biçimde analiz edilmesi gereken bir konudur (Aydemir Dev, 2023). Bu kapsamda, Dünya Mutluluk Raporu gibi veri setleri, veri setleri, uluslararası karşılaştırmaların daha sağlıklı yapılmasına katkı sunmuştur. Aynı zamanda bu veri setleri sayesinde, çok boyutlu refahın sayısal olarak ifade edilebilmesi mümkün kılınmış ve karar vericilere güçlü analiz olanakları sunulmuştur. Ancak bu veri setlerinin pek çok güçlü yanı olmasına rağmen, dezavantajları da mevcuttur. İçerdikleri çok sayıda değişkenin anlamlandırılması, geleneksel yöntemlerle her zaman kolay olmamaktadır. Bu bağlamda, boyut indirgeme ve yorumlanabilirlik sağlayan yöntemlerin kullanımı ön plana çıkmaktadır.

Tablo 2.1'de genel akışı görüldüğü üzere çalışmada, ülkelerin sosyal ve yapısal refah düzeylerini daha sade, yorumlanabilir ve görsel olarak ifade edilebilir biçimde analiz edebilmesi hedefiyle Seyrek Temel Bileşenler Analizi (STBA) yöntemi uygulanmıştır.

Tablo 2.1. Çalışmanın Süreç Akışı ve Matematiksel Modellemesi

Aşama	**İşlem**	**Matematiksel Model**
1	Veri Matrisi Tanımı	$X \in \mathbb{R}^{nxp}, x_{ij} \in$ [Mutluluk Göstergeleri]
2	Veri Standardizasyonu	$z_{ij} = (x_{ij} - \mu_j) / \sigma_j$
3	Kovaryans Matrisi	$\Sigma = (1/(n-1)) \cdot Z^T Z$
4	Boyut İndirgeme: TBA & STBA	$\max w^T \Sigma w - \lambda \|\|w\|\|_1 \text{ s.t. } \|\|w\|\|_2 = 1$
5	Kümeleme: K-MEANS	$\min_{(C_k)} \sum_{k=1}^{K} \sum_{z_i \in C_K} \|z_i - \mu_k\|^2$
6	Çıktı: Karar Modeli	Ülkelerin Gösterge Grupları

Bu çalışmanın temel amacı, geleneksel TBA yönteminin sınırlarının ötesine geçmek ve anlamlı değişkenlere odaklanan, açıklayıcılık gücü yüksek olan bir model ile ülkelerin çok boyutlu mutluluk profillerini analiz etmektir.

2.2. MATERYAL VE METHOD

2.2.1. Temel Bileşenler Analizi

Temel Bileşenler Analizi (TBA), çok değişkenli istatistiksel yöntemler arasında en eski ve yaygın tekniklerden biridir (Jolliffe, 2002).

Bu yöntem, ilk kez Karl Pearson'ın 1901 yılında yayımladığı "On lines and planes of closest fit to systems of points in space" çalışmasında tanımlanmış olup daha sonra Harold Hotelling, 1933'te yayımladığı "Analysis of a complex of statistical variables into principal components" makalesiyle yöntemsel temelleri literatüre kazandırılmıştır (Ersungur vd., 2007).

Bu analiz yönteminde amaç, çok sayıda değişkenin bileşkeleri olarak tanımlanabilecek daha az sayıda yeni değişken, diğer bir deyişle temel bileşenler üretmeyi hedeflemektedir. Bu temel bileşenler birbirinden bağımsızdır; böylelikle orijinal değişkenler arasındaki korelasyon yapısı da ortadan kaldırılmış olur (Ersungur vd., 2007).

TBA, literatürde kimi zaman temel faktör analizi olarak da isimlendirilen yöntemsel yaklaşımlarla ilişkilendirilmektedir. Bu analiz ile veri setinde yer alan ortogonal (birbirinden bağımsız) varyansların mümkün olan en büyük kısmını ortaya çıkarmaktır. Bu yöntemin çeşitli avantajları bulunmaktadır. Bu avantajlardan birisi, hata varyansını yok sayıp yalnızca değişkenler arasındaki ortak varyansa odaklanmasıdır. Bu açıdan faktör analizine benzer bir yaklaşım sergilemektedir. Ortak faktörler, değişkenleri etkileyen ve aralarındaki ilişkileri açıklayan temel yapılardır. Ortak varyans ise bu faktörler aracılığıyla açıklanabilen varyans miktarını ifade etmektedir (Turanlı vd., 2012).

TBA, başlangıçta yer alan p sayıda değişken, yine p adet birbirinden bağımsız (ortogonal) temel bileşen ile temsil edilir. Bu bileşenler, veri setindeki toplam varyansı aşamalı biçimde açıklar. İlk temel bileşen, toplam varyansa en yüksek katkıyı sağlayan yapıyı temsil ederken; sonraki bileşenlerin katkısı azalarak devam etmektedir. Bu yaklaşım, bileşenlerin sıralı olarak toplam varyans içindeki paylarına göre değerlendirilmesini sağlamaktadır. Genel anlamda, başlangıçtaki tüm değişken sayısı kadar bileşen elde edilebilmesine rağmen, bu bileşenlerin yalnızca bir kısmı anlamlı varyans açıklayıcı güce sahiptir. Bu nedenle analizde çoğunlukla, toplam varyansın büyük kısmını

açıklayan ilk birkaç bileşen dikkate alınırken, geri kalanların açıklayıcılığı daha düşük ve sınırlı kabul edilir (Albayrak vd., 2005).

TBA'nin, sıklıkla kullanılmasına rağmen, gerçek dünya uygulamalarındaki kullanımını sınırlayabilecek üç temel dezavantajı vardır. Bu dezavantajlardan bir tanesi, TBA, gerekli temel bileşenlerin sayısını belirlemek için doğrudan bir yöntem sunmaz. Bu sorun, model sırası seçimi kriterleri kullanılarak ele alınmıştır. İkinci olarak, TBA'de her bir değişkenin temel bileşenlere katkısını belirlemesinin zor olmasıdır. Bu durum, TBA'nin yükleme vektörlerinin yoğun yapısı nedeniyle meydana gelmektedir. Bu dezavantaj, açıklanabilir yapay zekâ ve özellik seçimi gibi alanlarda büyük ölçüde bir sorundur. Bu sorunun ortadan kalkabilmesi için, bazı değişkenlerin katkısının sıfır olması zorlanarak yükleme vektörlerinde yalnızca birkaç sıfır olmayan girişin bulunmasına izin verilmiş veya seyreklik kısıtlaması eklenerek ortogonallik kısıtlaması gevşetilmiştir. Bir diğeri ise, klasik TBA'nin analizinin çözümü aykırı değerlere karşı hassastır (Rekavandi vd., 2024).

TBA, çok değişkenli yapıları özetlemek amacıyla kullanılan doğrusal bir dönüşüm tekniğidir. STBA ise klasik bileşen analizine kıyasla daha sade ve yorumlanabilir sonuçlar elde etmeyi amaçlamaktadır. Bu yöntem, bazı değişkenlerin bileşen üzerindeki etkisini sıfıra indirerek yalnızca anlamlı katkı sunan değişkenlerin öne çıkmasını sağlar. Böylece, bileşenler daha yalın bir yapıya kavuşur ve yorumlama kolaylaşır.

2.2.2. Seyrek Temel Bileşen Analizi

Klasik Temel Bileşen Analizinin her bir bileşeni, tüm orijinal değişkenlerin lineer kombinasyonlarından oluştuğundan, sonuçların yorumlanabilirliği genellikle zordur. Bu sorunu gidermek için, 2006 yılında Journal of Computational and Graphical Statistics dergisinde Hui Zou, Trevor Hastie ve Robert Tibshirani tarafından "Seyrek Temel Bileşenler Analizi" (STBA) yöntemi geliştirilmiştir. Bu yöntem, bileşen yüklerine bir 'lasso' (L_1) cezası ekleyerek, bazı yüklerin sıfıra eşitlenmesini ve böylece daha az sayıda değişkenle temsil edilen, yorumlanabilir bileşenler elde edilmesini sağlar. (Zou vd., 2006).

STBA, temel bileşenler analizinin, seyrek yapı sağlayan elastik net regresyon yaklaşımıyla çözülebilecek bir optimizasyon problemine dönüştürülmesi esasına dayanır. Bu yöntem, bileşenleri oluştururken bazı katsayıların sıfırlanmasını sağlayarak hem değişken seçimi yapar hem de bileşenlerin daha yorumlanabilir hale gelmesini amaçlar (Todorov ve Filzmoser, 2013).

Ridge cezasının dahil edilmesinin, regresyon katsayılarını cezalandırmadığını gösterilmiştir. Bu sayede, $n \gg p$ olduğunda, ridge cezası λ için ceza parametresini sıfıra ayarlanabilir ve $n \ll p$ olduğunda, prensipte herhangi bir pozitif λ kullanabilir. STBA'nin amacı, açıklanan varyans oranını korurken yükleri seyrekleştirmektir. Bu bağlamda, açıklanan varyans ile ceza katsayısı arasındaki ilişkiyi görselleştiren düzeltilmiş toplam varyanstan hesaplanan açıklanan toplam varyansın yüzdesinin işaret grafiğini inceleyerek λ_1'i (L_1-ceza ayarlama parametresi) seçmek için veriye dayalı bir yaklaşım önermişlerdir (Zhao vd., 2020)

Son zamanlarda, STBA'nin teorik özelliklerini incelemeye olan ilgi önemli ölçüde artış göstermektedir (Berk ve Bertsimas 2019). STBA, verilerdeki yerelleştirilmiş mekansal yapıları belirleyerek ve farklı zaman ölçekleri arasındaki belirsizliği gidererek düşük sıralı yapıların daha iyi yorumlanmasını sağlayarak modern veri analizi için güçlü bir teknik olarak ortaya çıkmıştır (Erichson vd., 2020).

STBA, özellikle yüksek boyutlu verilerde hem boyut indirgemeyi hem de değişken seçimini bir arada gerçekleştirmesi sayesinde geniş bir uygulama alanı bulmuştur. STBA temel amacı, daha az önemli yüklemelerin bir kısmını sıfıra zorlayarak seyrek özvektörler elde etmektir. Çıkarılan bileşenlerde bu seyrekliği sağlamak için mevcut yöntemlerin çoğu, kovaryans matrisinin ana bileşenlerini bulurken TBA formülasyonuna bir kısıtlama veya ceza terimi ekler (Drikvandi ve Lawal, 2023).

TBA, yüksek boyutlu verileri, ilişkisiz ana bileşenler tarafından kapsanan daha düşük boyutlu bir alana yansıtarak verideki varyansın büyük bir kısmını korur. Ancak, temel bileşenlerin yorumlanmasının zor olduğu yaygın bir görüştür, çünkü her temel bileşen, orijinal değişkenlerin çoğunun veya tamamının doğrusal bir kombinasyonudur. Bu dezavantajı gidermek için geliştirilen STBA, yalnızca orijinal değişkenlerin küçük bir alt kümesini içeren "seyrek" ana bileşenleri tahmin eder (Chen ve Rohe, 2023). Klasik TBA ile karşılaştırıldığında, STBA'nin en belirgin farkı, bileşenlerin sadece az sayıda değişken içeriyor olmasıdır. Bu durum, yorumlanabilirliğin yanı sıra modelin gürültüden arındırılmasını da sağlar. Özellikle sosyal göstergelerin

analizinde, bileşenlerin hangi değişkenlerle tanımlandığının açıkça görülebilmesi, karar vericiler açısından oldukça değerlidir.

2.2.3. Kümeleme Analizi

Küme, benzer özellikler taşıyan gözlem birimlerinin oluşturduğu gruplar şeklinde tanımlanabilir. Kümeleme analizi ise gözlemler arasındaki benzerlik ya da farklılık ilişkilerini esas alarak bu birimleri belirli alt gruplara ayırmayı amaçlayan istatistiksel bir tekniktir. Bu yöntem gözetimsiz öğrenme yaklaşımı içinde yer alır; çünkü analiz sürecinde gözlemler önceden tanımlanmış herhangi bir sınıfa ait değildir. Kümelerin sayısı, merkezleri veya sınırları gibi bilgiler, modelin kendi yapısı ya da belirlenen bazı kriterler doğrultusunda ortaya çıkar. Literatürde kümeleme analizine ilişkin çok sayıda yöntem bulunmaktadır. Bu yöntemler temel yaklaşımlarına göre; bölümleme temelli, hiyerarşik, yoğunluk temelli, ızgara yapılı, model tabanlı ya da bulanık kümeleme gibi kategorilere ayrılmaktadır. Farklı disiplinlerde yoğun biçimde kullanılan bu yöntemler, kendi içlerinde çeşitli algoritmalar barındırır. Ancak tüm bu algoritmaların ortak hedefi; oluşturulan grupların içsel benzerliğinin yüksek, gruplar arası farklılıkların ise belirgin düzeyde olmasıdır (Özaltındış, 2025). Bu algoritmalardan birisi olan K-Ortalamalar (K-Means), gözlem değerleri arasındaki uzaklığı baz alarak, kümelerin merkezine göre yeni kümeler oluşturulması gerektiği düşüncesiyle çalışır (Akkartal, 2025).

Kümeleme analizinde yaygın olarak kullanılan K-Ortalamalar yöntemi, ilk kez 1957 yılında Stuart Lloyd tarafından önerilmiş, ancak literatürde geniş çapta yer bulması 1982 yılındaki yayımla mümkün

olmuştur. Bu yönteme "K-Means" adı ise 1967'de James MacQueen tarafından verilmiştir. Ayrıca, Edward W. Forgy'nin 1965'te önerdiği benzer algoritma nedeniyle bu yaklaşım bazı kaynaklarda Lloyd-Forgy algoritması olarak da anılmaktadır. K-Ortalamalar algoritması; sade yapısı, hızlı hesaplama süresi ve oldukça verimli sonuçlar sunması nedeniyle en sık tercih edilen kümeleme tekniklerinden biri olmuştur. Temel prensip olarak, gözlemler benzerliklerine göre belirlenen k sayıda kümeye atanır. Bu benzerliğin ölçülmesinde çeşitli uzaklık ölçütleri kullanılabilse de, en yaygın tercih edilen metrik genellikle Öklidyen uzaklıktır (Özaltındiş, 2025).

2.2.4. Veri ve Yöntem

Bu çalışmada analiz edilen veri seti, Birleşmiş Milletler Sürdürülebilir Kalkınma Çözümleri Ağı (UNSDSN) tarafından her yıl yayımlanan Dünya Mutluluk Raporu'nun 2023 yılına ait veri setidir. Bu veri setine resmî web sitesi üzerinden erişilmiştir (https://worldhappiness.report/ed/2023/#appendices-and-data). İlk rapor, 2012 yılında Birleşmiş Milletler Sürdürülebilir Kalkınma Çözümleri Ağı'nın düzenlediği toplantıya destek amacıyla hazırlanmış; bu sayede ekonomik refah kadar bireylerin mutluluk düzeylerinin de kalkınma gündemine dâhil edilmesi hedeflenmiştir (Kaya, 2022). Rapor, ülkelerin vatandaşlarına sunduğu yaşam kalitesini; ekonomik refah, sosyal destek, sağlıklı yaşam süresi, özgürlük, cömertlik ve yolsuzluk algısı gibi çok boyutlu göstergeler üzerinden değerlendirmektedir.

Çalışmada yer alan orijinal veri setinde 137 ülkeye ait bilgiler bulunmaktadır. Ancak bazı ülkelerde eksik veri bulunması nedeniyle

analiz kapsamına yalnızca tam gözlem içeren ülkeler dahil edilmiştir. Bu kapsamda, bazı temel değişkenlerinde eksiklik bulunan "State of Palestine" analiz dışında bırakılmış ve böylece analizde kullanılan ülke sayısı 137'den 136'ya düşmüştür. Veri ön işleme sürecinde eksik gözlemler çıkarılmış ve yalnızca altı değişkene odaklanılmıştır. Yalnızca mutluluk skorunu etkilediği düşünülen altı temel değişken ile ülke adı dikkate alınmıştır. Bu değişkenler Log GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices, Generosity, Perceptions of corruption. Çalışmada kullanılan değişkenler, bireylerin öznel mutluluğunu ve yaşam memnuniyetini etkileyen sosyal, ekonomik ve yapısal faktörleri temsil etmektedir. "Log GDP per capita" değişkeni, kişi başına düşen gayrisafi yurtiçi hasıla, bireylerin ekonomik refah düzeyini yansıtmaktadır. "Social support" değişkeni, bireylerin ihtiyaç duyduklarında yakın çevrelerinden destek alabilme düzeylerine ilişkin algılarını ölçmekte ve toplumsal dayanışmanın bir göstergesi olarak değerlendirilmektedir. "Healthy life expectancy" ise bireylerin yaşamlarını sağlıklı bir şekilde sürdürebilecekleri ortalama yaşam süresini ifade ederek hem yaşam süresi hem de yaşam kalitesine ilişkin bilgi sunmaktadır. "Freedom to make life choices" değişkeni, bireylerin kendi yaşamlarına dair karar alma özgürlüğü algısını temsil etmekte; demokratik haklar, ifade özgürlüğü ve bireysel özerklik gibi unsurları dolaylı biçimde yansıtmaktadır. "Generosity" değişkeni, toplumdaki gönüllülük faaliyetlerine katılım, bağış yapma eğilimi ve yardımseverlik gibi davranışlarla ilişkilidir. Son olarak, "Perceptions of corruption" değişkeni, bireylerin kamu kurumları ve özel sektöre dair yolsuzluk

algılarını ölçmekte olup, toplumda kurumsal güven düzeyinin dolaylı bir göstergesi olarak değerlendirilmektedir. Analiz sürecinde tüm değişkenler standardize edilerek z-puanlarına dönüştürülmüş ve böylece ölçüm birimlerinden bağımsız hale getirilmiştir.

Veri setinde yer alan orijinal değişken isimleri İngilizce olarak sunulmuş olup, çalışmanın bütünlüğünü ve anlatımını kolaylaştırmak amacıyla bu değişkenler Türkçeye çevrilerek analiz gerçekleştirilmiştir. Çalışmada kullanılan mutluluk göstergeleri, orijinal veri setinde yer alan İngilizce değişken adlarının Türkçeleştirilmiş karşılıkları ile analiz edilmiştir. Bu kapsamda, "Country Name" değişkeni "Ülke" olarak, "Log GDP Per Capita" değişkeni "Kişi Başına GSYİH (Log)" şeklinde ifade edilmiştir. Benzer şekilde, "Social Support" değişkeni "Sosyal Destek", "Healthy Life Expectancy" ise "Sağlıklı Yaşam Beklentisi" olarak adlandırılmıştır. "Freedom to Make Life Choices" değişkeni "Yaşam Seçimi Özgürlüğü", "Generosity" değişkeni "Cömertlik" ve "Perceptions of Corruption" değişkeni ise "Yolsuzluk Algısı" olarak çevrilmiştir. Tüm analizler bu Türkçe karşılıklar üzerinden yürütülmüş ve elde edilen bulgular ilgili değişken isimlerine göre yorumlanmıştır.

Veri analizi işlemleri, açık kaynaklı *Python* programlama dili kullanılarak *Google Colab* ortamında gerçekleştirilmiştir. Analizde, *sklearn* kütüphanesindeki *SparsePCA* sınıfı kullanılarak seyrek bileşenler elde edilmiş, sonrasında bu bileşenler *pandas*, *matplotlib* ve *seaborn* kütüphaneleri yardımıyla görselleştirilmiş ve yorumlanmıştır.

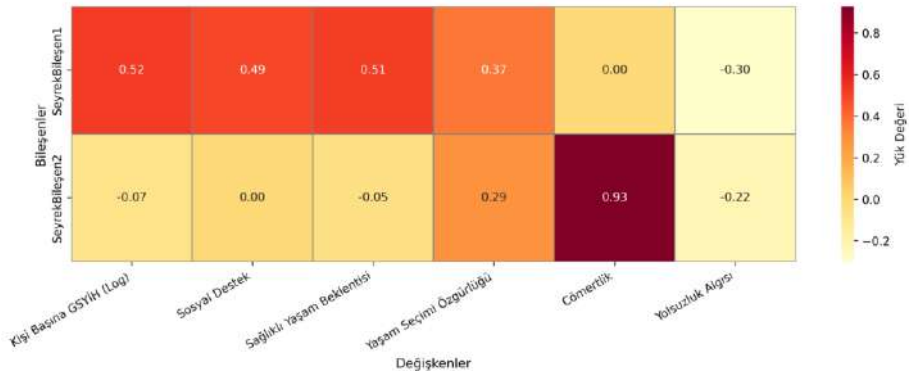
Bu aşamada sadece iki bileşenin çıkarılması tercih edilmiştir. Verinin yalnızca ilk iki temel bileşene indirgenmesi, gözlem noktalarının iki boyutlu bir düzlemde görselleştirilmesine imkân

tanıyarak kümelerin ve örüntülerin daha tutarlı ve anlaşılır biçimde tanımlanmasına ve yorumlanmasına olanak sağlar (Gewers vd., 2018). Bu tercih hem görselleştirmenin mümkün olması hem de ilk iki bileşenin açıklayıcılığının yüksek olması yönündeki literatür önerileriyle uyumludur. Bu çalışmada kapsamında kümeleme analizi için hiyerarşik olmayan yöntemler sınıfından bölümleme yönteminin bir yaklaşımı olan K-Ortalamalar yöntemi kullanılmıştır.

2.3. SONUÇLAR

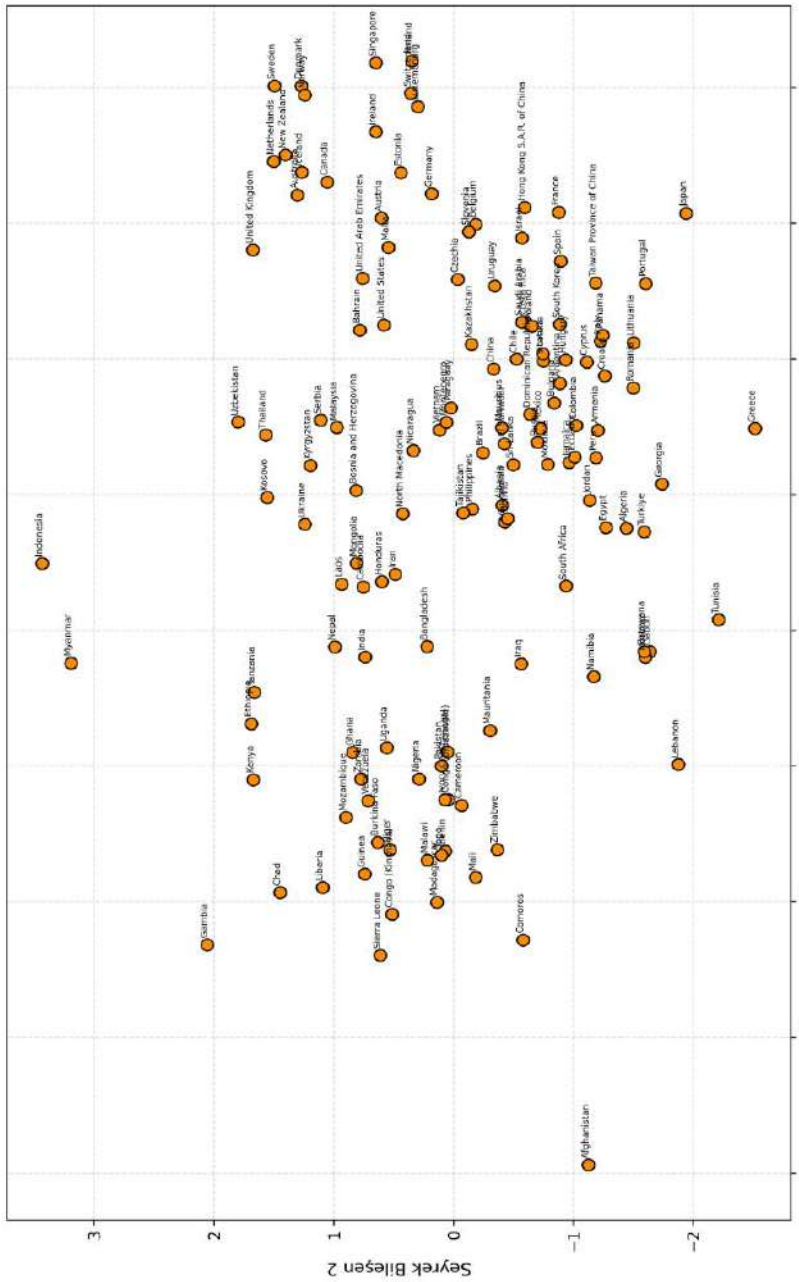
Bu başlık altında, STBA yöntemi kullanılarak elde edilen sonuçlara yer verilmiştir. Analiz kapsamında, mutluluk göstergelerinden oluşan çok boyutlu veri yapısı, iki temel bileşenle temsil edilen daha sade ve yorumlanabilir bir yapıya indirgenmiştir.

Elde edilen iki bileşenin yapısını açıklamak amacıyla, öncelikle her bir değişkenin bu bileşenlerdeki yük değerlerinin sunulduğu bir ısı haritası oluşturulmuştur. Bu görsel, bileşenlerin içeriksel yapısını ve yorumlanabilirliğini anlamak açısından temel bir referans niteliği taşımaktadır.



Şekil 2.1. Isı Haritası

Şekil 2.1'de yer alan ısı haritası, STBA yöntemi sonucunda elde edilen iki bileşenin, çalışmada kullanılan değişkenlerle olan ilişkisini göstermektedir. Bu görselde yer alan katsayılar (yükler), her bir değişkenin ilgili bileşeni tanımlamadaki katkı düzeyini ifade etmektedir. Elde edilen bileşenler, çeşitli değişkenlerin katkıları doğrultusunda şekillenmiş; her bir bileşenin hangi göstergelerden daha fazla etkilendiği, bileşen yükleri üzerinden analiz edilmiştir. STBA sonucunda elde edilen yük değerleri doğrultusunda, birinci bileşenin özellikle Kişi Başına GSYİH (Log), Sosyal Destek ve Sağlıklı Yaşam Beklentisi değişkenlerinden yüksek yükler aldığı görülmektedir. Bu durum, birinci bileşenin söz konusu göstergelerle yakından ilişkili olduğunu ve temel olarak bu üç değişkenin etkisiyle şekillendiğini ortaya koymaktadır. Bu durum, söz konusu bileşen olan SeyrekBileşen1'in daha çok ekonomik refah ve temel yaşam koşullarını yansıttığını göstermektedir. İkinci bileşen olan SeyrekBileşen2 ise yalnızca Cömertlik değişkeninden yüksek bir yük alarak, bu bileşenin büyük ölçüde sadece bu göstergeden etkilendiğini göstermektedir. Bu bileşen ise ülkelerdeki toplumsal güven, etik yapı ve bireysel özgürlüklerin etkisini yansıtan daha soyut bir boyutu temsil etmektedir. Bu değerlendirmeler, Şekil 2.1'de sunulan bileşen yükü ısı haritasına dayanmaktadır. Bu yapı sayesinde, çok boyutlu mutluluk verisi daha sade iki bileşen üzerinden temsil edilmiş ve her bileşenin hangi değişkenlerle tanımlandığı açık şekilde ortaya konmuştur. Bu durum, ilerleyen analizlerde ülkelerin neden belirli gruplarda toplandığını anlamak a. açısından da temel bir referans oluşturmaktadır.

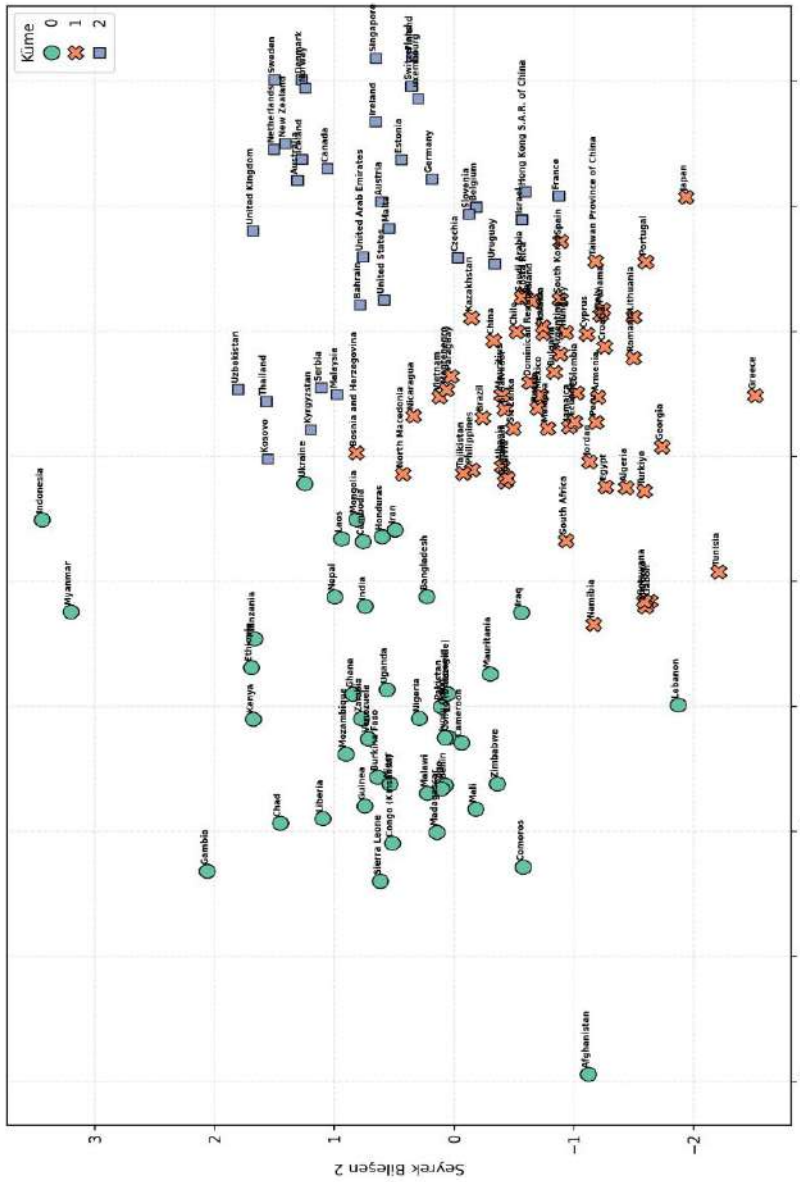


Şekil 2.2. İki Bileşenli Düzlemde Konumlanan Ülkeler

Şekil 2.2'de, analiz kapsamında yer alan ülkeler, iki bileşenli düzlemde konumlandırılmıştır. Her bir nokta bir ülkeyi temsil etmekte ve bu konumlar, ülkenin mutluluk göstergeleri doğrultusunda elde edilen bileşen skorlarına dayanmaktadır. Dağılım yapısı incelendiğinde, yapısal refah düzeyi yüksek olan ülkelerin benzer bileşen skorlarına sahip olduğu ve bu nedenle birbirlerine yakın konumlandığı görülmektedir. Öte yandan, bireysel değerler ve toplumsal güven unsurlarının belirgin olduğu ülkelerin daha farklı bir eksende kümelendiği dikkat çekmektedir. STBA'nın iki bileşeni, ülkelerin Beklenen Yaşam Süresi, Kentsel Nüfus Oranı, İnternet Kullanıcıları ve İşsizlik Oranı gibi yapısal ve sosyal göstergeler temelinde açık bir şekilde ayrıştırılmasını ve kümelenmesini mümkün kılmaktadır.

Şekil 2.2'de, birinci ve ikinci bileşen değerleri negatif olan ve sol alt çeyrekte konumlanan Afganistan, Lübnan, Tunus, Botswana ve Namibya gibi ülkelerin, hem ekonomik refah hem de toplumsal değer göstergeleri bakımından daha düşük bileşen skorlarına sahip olduğu görülmektedir. Bu görsel, veri setine dâhil edilen tüm ülkelerin birinci ve ikinci seyrek bileşenler ekseninde dağılımını sunarak, benzerlik ve ayrışma desenlerinin bütüncül bir görünümünü sağlar. Böylece düşük gelişmişlik göstergelerine sahip ülkeler ile yüksek gelişmişlik eğilimi gösteren Birleşik Krallık, İsveç, Yeni Zelanda gibi ülkeler arasında anlamlı farklılıkların ortaya çıktığı anlaşılmaktadır. Bu iki boyutlu görsel, çalışmanın amacı doğrultusunda ülkeler arası sosyo-ekonomik

yapı benzerlikleri ve farklıklarının hızlı biçimde tanımlanmasına ve yorumlanmasına olanak sağlamaktadır.



Şekil 2.3. STBA Sonuçlarına Dayalı K-Means Kümeleme ile Ülkelerin İki Boyutlu Görselleştirmesi

Şekil 2.3'te, çalışmaya dâhil edilen tüm ülkeler, STBA sonucu elde edilen iki bileşenli düzlemde konumlandırılmış ve K-means algoritması yardımıyla üç kümeye ayrılmıştır. Kümeleme sonucunda ülkeler hem renk hem de sembol farklılıkları ile ayrıştırılmıştır. Bu dağılıma göre, Küme 0 toplamda 47 ülkeyi, Küme 1 38 ülkeyi, ve Küme 2 ise 51 ülkeyi kapsamaktadır. Küme 0'da bulunan ülkeler kapsamında Afganistan, Hindistan, Kenya, Irak öne çıkarken; Endenozya, Laos, Burma da bu gruptadır. Küme 1'de bulunan ülkelerden birkaçı ise, Bosna Hersek, Yunanista ve Kıbrıs bulunmaktadır. Çin ve Türkiye'nin de turuncu renkle gösterilen Küme 1 içinde yer aldığı görülmektedir. Bu kümede yer alan ülkelerin seyrek bileşen skorlarına göre belirli göstergelerde benzer yapısal özellikler taşıdığı söylenebilir. Küme 2'de ise Kanada, Danimarka, İngiltere ve İrlanda gibi yüksek mutluluk skoruna sahip ülkelerle birlikte Almanya, Fransa, ve Kosova'yı içermektedir. ABD, İngiltere ve Batı Avrupa ülkelerinin önemli bir kısmı da mavi renkle gösterilen Küme 2 içerisinde toplanmıştır. Bu küme, ekonomik ve sosyal göstergeler açısından göreli bir homojenlik sunmakta olup, gelişmiş ülkelerin büyük bölümünü içermektedir. Bu örneklemeler, kümeleme analizinin ülkeler arasındaki mutluluk profillerini coğrafi ve kurumsal çeşitlilik açısından etkili biçimde ortaya koyduğunu göstermektedir.

Şekil 2.3 incelendiğinde, 0 numaralı kümede, genellikle düşük GSYİH, düşük sağlık ve sosyal destek göstergelerine sahip Afrika ve Asya ülkeleri yer almaktadır. Bu ülkeler bileşen düzleminin sol kısmında yoğunlaşmış, birikim gösteren bir grup oluşturmuştur. 1

numaralı küme, orta gelir grubundaki ülkelerden oluşmakta; ekonomik ve sosyal göstergeler açısından daha karma bir yapıya sahip olan ülkeleri içermektedir. Bu grup, bileşen düzleminin merkezine yakın, geniş bir alana yayılmıştır. 2 numaralı küme ise ağırlıklı olarak Avrupa, Kuzey Amerika ve Okyanusya ülkelerini içermektedir. Bu ülkeler daha yüksek düzeyde ekonomik refah, sosyal destek ve sağlık beklentisine sahiptir. Bu nedenle bileşen düzleminin sağ tarafında, birlikte kümelenmişlerdir. Görseldeki net ayrışma, analizin sunduğu yapısal gruplamaların anlamlı ve yorumlanabilir olduğunu göstermektedir.

Tablo 2.2. Ortalama Gösterge Değerleri

Küme	Kişi Başına GSYİH (Log)	Sosyal Destek	Sağlıklı Yaşam Beklentisi	Yaşam Seçimi Özgürlüğü	Cömertlik	Yolsuzluk Algısı
0	8.10	0.66	58.77	0.72	0.10	0.79
1	9.80	0.84	66.66	0.79	-0.08	0.79
2	10.63	0.91	70.10	0.88	0.11	0.53

Kümeleme sonucu elde edilen üç grubun ortalama gösterge değerleri Tablo 2.2'de sunulmuştur. Tabloya göre, Küme 2 yüksek GSYİH, yüksek sosyal destek ve yüksek yaşam beklentisi ile refah düzeyi yüksek ülkeleri temsil etmektedir. Küme 0 ise düşük GSYİH, düşük sosyal destek ve düşük yaşam beklentisi ile daha kırılgan yapıdaki ülkeleri kapsamaktadır.

Bu kümeleme yapısı, yalnızca sayısal farklılıkları değil, aynı zamanda ülkelerin mutluluğa ilişkin temel göstergelerdeki yapısal benzerliklerini de ortaya koymaktadır. Çalışmada kullanılan 6 değişken

üzerinden uygulanan STBA ile ülkeler iki boyutlu bileşen düzlemine indirgenen ve ardından K-Ortalamalar algoritması ile benzer gösterge profiline sahip ülkelerin aynı grupta toplandığı gözlemlenmiştir. Elde edilen bu coğrafi dağılım, STBA ile elde edilen bileşenlerin ülkeleri sadece nicel düzeyde değil, aynı zamanda yapısal ve ilişkisel açıdan da sınıflandırabildiğini ortaya koymaktadır. Görselleştirme, kümeleme sonuçlarının mekansal örüntülerle örtüşmesini sağlayarak analizin yorumlanabilirliğini güçlendirmiştir.

2.4. TARTIŞMA

Bu çalışmada, dünya mutluluk endeksine ilişkin belirleyici göstergeler kullanılarak STBA uygulanmış ve elde edilen bileşen skorları doğrultusunda ülkeler kümelenmiştir. Geleneksel temel bileşen analizinde, her bir bileşenin tüm değişkenlerin doğrusal birleşiminden oluşması, yorumlamayı zorlaştırmakta ve açıklayıcılığı sınırlı hâle getirmektedir. Bu nedenle, yalnızca bazı değişkenleri içeren ve bileşenleri daha anlamlı hâle getiren seyrek temel bileşen analizi tercih edilmiştir. Bu yöntem, verinin SeyrekBileşen1 ve SeyrekBileşen2 olarak 2 bileşene indirgenmesini sağlamıştır. Bu sayede birinci bileşen beklenen yaşam süresi, şehirleşme düzeyi ve internet erişimi gibi sosyo-ekonomik gelişmişlik göstergelerini temsil ederken ikinci bileşen istihdam oranı, kadın işgücüne katılım ve işsizlik düzeyi üzerinden istihdam yapısını yansıtmıştır. Çalışmanın temel katkısı, iki bileşenin sosyo-ekonomik gelişmişlik ve istihdam yapısına ilişkin göstergeleri ayrıştırarak yapısal olarak temsil edebilmesidir. Bu kapsamda

değişkenler ile bileşenler arasındaki ilişkilerin daha açık ve yorumlanabilir şekilde ortaya konmasına olanak tanımıştır.

Analiz sonucunda, ülkelerin benzer sosyo-ekonomik özellikler çerçevesinde kümelendiği ve özellikle Kişi Başına GSYİH (Log), Sosyal Destek, Sağlıklı Yaşam Beklentisi, Yaşam Seçimi Özgürlüğü, Cömertlik ve Yolsuzluk Algısı göstergeleri bağlamında anlamlı şekilde ayrıştığı görülmüştür. İlk iki seyrek bileşen düzleminde yapılan görselleştirme, ülkelerin bu göstergeler ekseninde nasıl konumlandığını açık biçimde ortaya koymuştur. Sosyo-Ekonomik Gelişmişlik Göstergeleri Bileşeni(SeyrekBileşen1), yüksek yük değerleriyle GSYİH (0.52), Sosyal Destek (0.49) ve Sağlıklı Yaşam Beklentisi (0.51) gibi değişkenlerle temsil edilmekte olup, İstihdam Yapısı Bileşeni (SeyrekBileşen2) ise en güçlü yükü Cömertlik (0.93) değişkeninden almaktadır.

Küme yapısı incelendiğinde, sosyal destek düzeyi yüksek ve yaşam memnuniyeti ortalaması daha fazla olan ülkelerin benzer kümelerde toplandığı; buna karşılık daha düşük göstergelere sahip ülkelerin farklı kümelerde yer aldığı gözlemlenmiştir. Bu sonuçlar, Tablo 2.2'de sunulan kümelere göre ortalama değişken değerleriyle istatistiksel olarak desteklenmiş, karşılaştırmalar görsel olarak da anlamlı biçimde sunulmuştur.

İleride yapılacak çalışmalarda yöntemin daha da güçlendirilmesi adına bazı açılımlar önerilmektedir. Öncelikle, çalışmada yalnızca tek yıl verisi kullanılmıştır; oysa zaman serileri üzerinden yapılacak çok yıllı karşılaştırmalar ile ülkelerin zaman

içindeki değişim dinamikleri analiz edilebilir. Kümeleme aşamasında yalnızca K-means yöntemi tercih edilmiştir; alternatif kümeleme algoritmaları kullanılarak yöntemsel karşılaştırmalar yapılması, farklı gruplama yapılarının ortaya çıkmasını sağlayabilir. Son olarak, kültürel, coğrafi veya politik değişkenlerin modele entegrasyonu ile analizin kapsamı genişletilebilir ve daha bütüncül bir bakış açısı elde edilebilir.

KAYNAKÇA

Akkartal, G. R. (2025). Türkiye, Orta Doğu ve Kuzey Afrika (MENA) Ülkelerinin Lojistik Performans Endekslerinin Kümeleme Analizi Yöntemi ile İncelenmesine Yönelik Bir Çalışma. İstanbul Gelişim Üniversitesi Sosyal Bilimler Dergisi, 12(1), 98-119. https://doi.org/10.17336/igusbd.1406509

Albayrak, A. S. (2005). Türkiye'de illerin sosyoekonomik gelişmişlik düzeylerinin çok değişkenli istatistik yöntemlerle incelenmesi. Uluslararası Yönetim İktisat ve İşletme Dergisi, 1(1), 153–177.

Berk, L., & Bertsimas, D. (2019). Certifiably optimal sparse principal component analysis. Mathematical Programming Computation, 11(3), 381–420. https://doi.org/10.1007/s12532-018-0153-6

Chen, F., & Rohe, K. (2023). A new basis for sparse principal component analysis. Journal of Computational and Graphical Statistics, 33(2), 421–434. https://doi.org/10.1080/10618600.2023.2256502

Aydemir Dev, M. (2023). Türkiye'deki İllerin Sosyoekonomik Gelişmişlik Düzeylerinin Analizi. Bulletin of Economic Theory and Analysis, 8(2), 333-355.

Drikvandi, R., & Lawal, O. (2023). Sparse principal component analysis for natural language processing. Annals of Data Science, 10, 25–41. https://doi.org/10.1007/s40745-020-00277-x

Erichson, N. B., Zheng, P., Manohar, K., Brunton, S. L., Kutz, J. N., & Aravkin, A. Y. (2020). Sparse principal component analysis via variable projection. SIAM Journal on Applied Mathematics, 80(2), 977–1002.

Ersungur, Ş. M., Kızıltan, A., & Polat, Ö. (2007). Türkiye'de bölgelerin sosyo-ekonomik gelişmişlik sıralaması: Temel bileşenler analizi. Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, 21(2), 55–66.

Gewers, F. L., Ferreira, G., Arruda, H., Silva, F., Comin, C., Amancio, D., & Costa, L. (2018). Principal component analysis: a natural approach to data exploration (2018). arXiv preprint arXiv:1804.02502.

Helliwell, J. F., Layard, R., Sachs, J. D., Aknin, L. B., De Neve, J.-E., & Wang, S. (Eds.). (2023). World Happiness Report 2023 (11th ed.). Sustainable Development Solutions Network.

Jolliffe, I. T. (2002). Principal component analysis. Springer-Verlag.

Kaya, H. (2022). OECD ülkelerinin ekonomik Özgürlük Endeksi ve Dünya Mutluluk Raporu açısından kümeleme analizi (Master's thesis, Dokuz Eylul Universitesi (Turkey)).

Özaltındış, T. (2025). Mekânsal ağırlıklandırılmış K-ortalamalar yaklaşımı ile enerji tüketim verilerinin kümelenmesi [Doktora tezi, Mimar Sinan Güzel Sanatlar Üniversitesi].

Rekavandi, A. M., Seghouane, A. K., & Evans, R. J. (2024). Learning robust and sparse principal components with the α-divergence. IEEE Transactions on Image Processing.

Todorov, V., & Filzmoser, P. (2013). Comparing classical and robust sparse PCA. In Synergies of soft computing and statistics for intelligent data analysis (pp. 283-291). Springer Berlin Heidelberg.

Turanlı, M., Cengiz, D., & Bozkır, Ö. (2012). Faktör analizi ile üniversiteye giriş sınavlarındaki başarı durumuna göre illerin sıralanması. Istanbul University Econometrics and Statistics e-Journal, 17, 45–68.

Zhao, Y., Lindquist, M. A., & Caffo, B. S. (2020). Sparse principal component based high-dimensional mediation analysis. Computational statistics & data analysis, 142, 106835.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15(2), 265–286. http://www.jstor.org/stable/27594179

BÖLÜM 3:

AVRUPA BİRLİĞİ ÜLKELERİNİN MADENCİLİK VE TAŞ OCAKÇILIĞI İTHALAT VE İHRACATININ SEYREK İSTATİSTİKSEL YÖNTEMLER İLE SINIFLANDIRILMASI

Ece Nur GENÇER [6]

Özet

Bu çalışma, Avrupa Birliği'ne üye 27 ülkenin 2019-2023 yılları arasındaki madencilik ve taş ocakçılığı ürünlerine ilişkin ithalat ve ihracat miktarlarını ve değerlerini inceleyerek, ülkeler arasındaki dış ticaret yapısal benzerliklerini seyrek istatistiksel yöntemlerle sınıflandırmayı amaçlamaktadır. Yüksek boyutlu ve seyrek yapılara sahip veri setlerine uygun olarak, analizde Seyrek K-Ortalamalar, Seyrek Temel Bileşenler Analizi (Sparse PCA) ve Ufak Örneklem K-Ortalamalar (Mini-Batch K-Means) yöntemleri kullanılmıştır. Seyrek PCA sonrası yapılan ufak-örneklem k-ortalamalar kümeleme analizinde daha yüksek doğruluk oranı elde edilmesine rağmen, değişken seçimini doğrudan yapması ve yapısal yorumlanabilirliği sağlaması nedeniyle Seyrek K-Ortalamalar yöntemi tercih edilmiştir. Elde edilen sonuçlar, ülkelerin ithalat ve ihracat dengeleri, ekonomik entegrasyon düzeyleri ve sektörel bağımlılıkları açısından anlamlı kümelere ayrıldığını göstermektedir. Özellikle Belçika, Fransa, Almanya ve Hollanda gibi ülkeler yüksek dış ticaret hacimleriyle ayrışırken; çevresel ve ticari olarak daha sınırlı etkiye sahip ülkeler farklı kümelerde gruplanmıştır. Bu bulgular, Avrupa Birliği içinde madencilik ve taş ocakçılığı sektörlerinin ülkeler arası ekonomik yapılarla nasıl ilişkili olduğunu ortaya koymak açısından önemlidir.

[6] Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı., İstanbul, Türkiye, ecenurgencer@gmail.com, ORCID ID: 0009-0003-1283-1303

CLASSIFICATION OF THE MINING AND QUARRYING IMPORTS AND EXPORTS OF EUROPEAN UNION COUNTRIES USING SPARSE STATISTICAL METHODS

Abstract

This study aims to classify the structural similarities in foreign trade among 27 European Union (EU) member states between 2019 and 2023 by analyzing the import and export quantities and values of mining and quarrying products using sparse statistical methods. Considering the high-dimensional and sparse nature of the dataset, methods such as Sparse K-Means, Sparse Principal Component Analysis (Sparse PCA), and Mini-Batch K-Means clustering were applied. Although higher clustering accuracy was achieved through Sparse PCA combined with Mini-Batch K-Means, the Sparse K-Means method was ultimately preferred due to its ability to directly select variables and enhance interpretability. The findings reveal that countries are grouped into meaningful clusters based on trade balances, levels of economic integration, and sectoral dependencies. Nations such as Belgium, France, Germany, and the Netherlands distinguished themselves with their high trade volumes, while countries with more limited economic influence were grouped into different clusters. These results are significant in understanding the relationship between mining and quarrying trade patterns and the broader economic structures of EU countries.

3.1. GİRİŞ

Avrupa Birliği (AB), dünyada ekonomik açıdan en gelişmiş bölgelerden biri olarak kabul görmektedir. Bu gelişmişlik, uzun yıllar boyunca madencilik sektörü tarafından desteklenmiştir. Madencilik sektörü hem dünya genelinde hem de Avrupa Birliği (AB) ülkelerinde ekonomik kalkınmanın temel direklerinden biridir.

AB ülkeleri, madencilik ve taş ocakçılığı sektörünün büyüklüğü açısından büyük çeşitlilik göstermektedir; bu durum, emisyonların hem hacmi hem de yapısının ülkeden ülkeye değişmesine neden olmaktadır. Ancak, dinamik sosyo-ekonomik gelişim, enerji ve mineral kaynaklara olan talebi artırmaktadır; bu kaynaklar neredeyse tüm ekonomi sektörlerinde kullanılmaktadır.

Avrupa Topluluğu Ekonomik Faaliyetlerin İstatistiki Sınıflaması-Nace Rev 2 (Europen Union, 2025) baz alındığında madencilik sektörü, "madencilik ve taş ocakçılığı" olarak birlikte adlandırılmaktadır. Bu sınıflama; katı (örneğin kömür ve cevherler), sıvı (örneğin petrol) ve gaz (örneğin doğal gaz) hâlinde kayaç kütlesinden mineral çıkarımıyla ilgili tüm faaliyetleri kapsamaktadır. Ayrıca, maden yataklarının aranması ve madencilikle ilgili diğer destek faaliyetleri de bu sektöre dâhildir.

Her ne kadar AB ülkelerinde madencilik sektörünün önemi bir miktar azalmış olsa da bu sektör hâlâ ekonominin önemli bir parçasıdır. AB ülkelerinde çıkarılan mineral ve enerji hammaddeleri, enerji ihtiyacının önemli bir kısmını karşılamaktadır (örneğin Polonya, Almanya) ve neredeyse tüm ekonomik sektörlerde kullanılan hammaddeleri sağlamaktadır (inşaat, kimya, ilaç, uzay, otomotiv,

elektronik ve diğer sanayiler). Günümüzde bahsi geçen bu hammaddeler, ülkelerin birçoğunda stratejik ürünler olarak kabul görmektedir (Brodny ve ark., 2020).

Günümüzde insanlığın gelişmesi ve nüfusun artmasıyla birlikte, barınma gibi temel gereksinimlerini karşılayabilmeleri için inşaat faaliyetlerinin fazlalaşması, ekonomi için inşaat ham maddelerine karşı talebi de artırmıştır. Taş ocakçılığı sektörü, inşaat sanayisinin temel ham maddelerini üretmektedir.

Taş ocakçılığı, genellikle kayaçların yeryüzeyinde ya da hemen altındaki alanlardan çıkarılması sürecidir. Madencilik ile taş ocakçılığı arasındaki fark, taş ocakçılığının metalik olmayan kayaçlar ve agregaları çıkarması, madenciliğin ise mineral yataklarını çıkarmak için kazı yapmasıdır. Kumtaşı, kireçtaşı, perlit, mermer, demirtaşı, arduvaz, granit, kaya tuzu ve fosfat kayaçları gibi çeşitli taşlar, bu sektör tarafından bir dizi işlemle çıkarılmaktadır (George, 2021).

3.2. MATERYAL VE METOD

3.2.1. Seyrek Kümeleme Metodu

Seyrek kümeleme, klasik kümeleme algoritmalarının yüksek boyutlu veri setlerinde karşılaştığı temel problemlere çözüm olarak geliştirilmiş bir yöntemler bütünüdür. Geleneksel kümeleme yöntemleri (örneğin K-ortalamalar veya hiyerarşik kümeleme), veri setindeki tüm değişkenleri eşit önemde kabul ederek kümeleme yapar. Ancak gerçek hayattaki birçok veri setinde, özellikle genetik veriler, metin madenciliği veya görüntü işleme gibi yüksek boyutlu veri yapılarında, tüm değişkenlerin

küme yapısını belirlemede aynı ölçüde anlamlı olmadığı bilinmektedir (Witten ve ark., 2010).

Bu aşamada seyrek kümeleme yaklaşımı, veri kümesi içerisinde bulunan ve kümeleme açısından anlamlı bilgi barındıran gözlemleri otomatik olarak seçerek kümelemeyi bu seçilmiş gözlemler üzerinden yapmayı hedefler (Witten ve ark., 2010; Gaynor ve ark., 2017). Bu sayede değişken sayısının azaltılmasının yanında (boyut indirgeme sağlanır) kümeler arası heterojenlik daha bariz ve yorumlanabilir olur.

Seyrek kümeleme yöntemlerinde temel olarak veri setindeki değişkenlere ağırlıklar atanır. Bu ağırlıklar, kümeler arası varyansı artıran ve kümeler içi varyansı azaltan değişkenleri ön plana çıkaracak şekilde optimize edilir (Witten ve ark., 2010). Yöntemde kullanılan cezalandırıcı (penalty) terimler sayesinde önemsiz değişkenlerin ağırlıkları sıfırlanır (L_1 cezası) veya azaltılır (L_2 cezası). Bu cezalandırıcı terimler sayesinde örnek seçim işlemi otomatik olarak modelin bir çıktısı olmuş olur.

Bu metodun genel formülasyonunda optimizasyon şu şekilde yapılır:

$$min_{w,C} \sum_{j=1}^{p} w_j . \Delta_j(C) \tag{3.1}$$

Burada:

- C kümeleme çözümünü,
- w_j $j.$ değişkenin ağırlığını,
- $\Delta_j(C)$ ise kümeler içi $j.$ değişken bazında hesaplanan toplam uzaklığı ifade eder.

Cezalandırma koşulları:

$$\sum_{j=1}^{p} {w_j}^2 \leq 1, \qquad \sum_{j=1}^{p} |w_j| \leq s, \quad w_j \geq 0 \tag{3.2}$$

Buradaki s seyreklik parametresi olup, modelde seçilecek değişken sayısını kontrol eder (Witten ve ark., 2010).

Seyrek kümeleme yöntemleri özellikle aşağıdaki durumlar için uygundur:

- Değişkenlerin sayısının örneklerin sayısından daha fazla olduğu (p > n) veri kümelerinde,
- Gürültülü değişkenlerin bulunduğu durumlarda,
- Kümeleme sonucu değişken seçiminin de istendiği durumlarda,
- Yorumlanabilir model çıktısı beklendiğinde.

Seyrek kümelemenin yaptığı en kayda değer katkı, yalnızca kümeleri doğru belirlemek değil; bunun yanında kümeler arasındaki farklılığın nedeni olan değişkenleri daha net gösterebilmesidir (Arias-Castro ve ark., 2017; Floriello ve ark., 2017).

3.2.1.1. *Seyrek Konveks Kümeleme*

Bu yöntemde hem gözlem hem de değişken seviyesinde kümeleme ve değişken seçimi aynı anda gerçekleştirilmektedir (Wang ve ark.., 2017).

3.2.1.2. *Seyrek Alt Uzay Kümeleme*

Bu türde, her veri noktası diğer veri noktalarının seyrek doğrusal birleşimi olarak modellenir. Böylece her gözlem, bulunduğu alt uzaydaki diğer gözlemlerle ilişkilendirilmiş olur ve kümeleme işlemi spektral yöntemlerle gerçekleştirilir (Elhamifar ve ark., 2011).

3.2.1.3. *Büyük Ölçekli Seyrek Kümeleme (LSSC)*

LSSC, önce daha taslak bir gruplama işlemi yapar (örneğin K-Ortalamalar), sonrasında ise daha detaylı bir seyrek kodlamayla kümeleri hassaslaştırır (Zhang ve ark., 2016).

3.2.1.4. *Seyrek K-Ortalamalar Kümeleme*

Seyrek K-Ortalamalar, klasik K-Ortalamalar algoritmasının daha fazla boyuta sahip veri kümelerinde karşılaştığı değişken seçimi problemine çözüm sunan bir kümeleme metodudur (Witten ve ark., 2010). Bu metot, k-ortalamalar yönteminin nesnel fonksiyonuna L_1 (lasso) ve L_2 (ridge) cezalarının eklenmesiyle değişken seçme işlemini otomatik hale getirir (Vouros ve ark., 2020).

Seyrek K-Ortalamalar algoritmasında esas amaç, örnekler ile küme merkezleri arasında bulunan mesafelerin kareler toplamını azaltırken, önemli değişkenleri seçerek gürültü değişkenlerini elemektir (Witten ve ark., 2010). K-ortalamalar algoritmasındaki standart optimizasyon problemi aşağıdaki gibi ifade edilir:

$$J_{kmeans} = \sum_{k=1}^{K} \sum_{x_i \epsilon c_k} \sum_{j=1}^{p} (x_{ij} - m_{kj})^2 \qquad \textbf{(3.3)}$$

Burada x_{ij} veri matrisindeki *i.* gözlemin *j.* değişkenidir, m_{kj} ise *k.* kümenin *j*. değişkenindeki merkezidir (Vouros ve ark., 2020).

Seyrek K-Ortalamalar algoritması bu yapıya değişkenlerin ağırlıklarını da ekleyerek tekrar ifade edilmiştir:

$$J_{skmeans} = \sum_{j=1}^{p} w_j \left[\sum_{i=1}^{n} (x_{ij} - \mu_{ij})^2 - \sum_{k=1}^{K} \sum_{x_i \epsilon C_k} (x_{ij} - m_{kj})^2 \right] \qquad (3.4)$$

Burada w_j ağırlıklandırılmış değişken vektörüdür ve değişken seçimi bu ağırlıklar aracılığıyla gerçekleştirilir. İlgili optimizasyon probleminde değişken seçimi ve regularizasyonu şu kısıtlarla sağlanır (Witten ve ark., 2010):

$$\sum_{j=1}^{p} {w_j}^2 \leq 1, \qquad \sum_{j=1}^{p} |w_j| \leq s, \quad w_j \geq 0 \qquad (3.4)$$

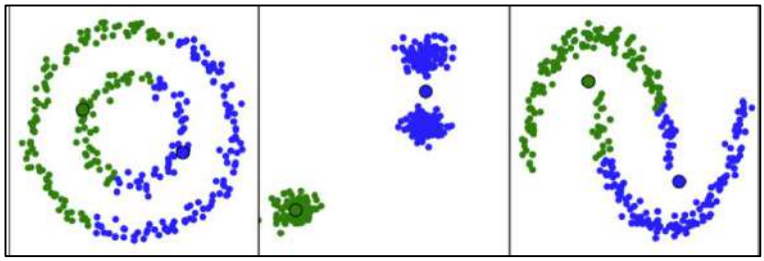
Burada s parametresi modelin seyreklik derecesini yani seçilecek değişken sayısını belirler (Chang, 2018).

Seyrek K-Ortalama yönteminin önemli özelliklerinden bir diğeri ise yüksek boyuta sahip veri kümelerinde değişken seçimi yaparken gruplar arası varyansı maksimum yapmaya çalışmasıdır. Grup seyreklik metodu, özellikle karma yapıya sahip veri setleriyle (sayısal, kategorik bir arada) çalışırken kategorik değişkenler için kukla değişken kullanılır ve bu dönüşümün ardından ayrı ayrı gruplar şeklinde değişken seçimi gerçekleştirilir (Chavent, 2020).

Ayrıca seyrek k-ortalamanın geliştirilmiş bir versiyonu olan L_1/L_0 cezalandırılmış seyrek k-ortalama metodunda, hem L_1 hem de L_0 cezaları beraber kullanılarak daha güçlü bir değişken seçimi sağlanmıştır. Bu yöntemde amaç fonksiyonu şu şekilde tanımlanmıştır:

$$max_{C,w} \sum_{j=1}^{p} w_j \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{ii'j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \epsilon C_k} d_{ii'j} \right) \qquad (3.5)$$

Buradaki $d_{ii'j}$, veri setindeki iki gözlem arasındaki j. değişkende bulunan uzaklığı temsil eder (Chang, 2018).



Şekil 3.1. Farklı Veri Kümelerinde Seyrek K-Ortalamalar Grafikleri (https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/modules/clustering.html)

3.2.1.5. *Ufak Örneklem K-Ortalamalar Metodu (Mini-Batch)*

Büyük boyutlu veri setleriyle çalışırken klasik K-Ortalamalar algoritmasının bellek ve çalışma süresi anlamında yetersiz kalması, farklı optimizasyon tekniklerinin geliştirilmesine sebep olmuştur.

Ufak Örneklem K-Ortalamalar, her tekrarda tüm veri kümesini değil de, rasgele bir şekilde seçilen daha ufak bir örneklem (mini-batch) ile çalışır. Bu şekilde hem zaman hem de hafıza tasarrufu sağlanmış olur (Bejar, 2013). Temel yapı şu adımlardan oluşur:

1. Başlangıçta, k adet merkez noktası genellikle K-Ortalamalar++ Kümeleme yöntemiyle rastgele seçilir (Zappia ve ark., 2021).

2. Her tekrarda, veri kümesinden a boyutunda bir örnek seçilir:

$$M^{(t)} = \{x_1{}^{(t)}, x_2{}^{(t)}, \dots, x_a{}^{(t)}\} \tag{3.6}$$

3. Mini-batch içerisindeki örneklerin herbiri, en yakın merkez noktaya atanır.

4. Bu atamalar sonucu merkezler güncellenir. Güncelleme formülü şu şekildedir:

$$\mu_c{}^{(t)} = (1 - \eta_c)\mu_c{}^{(t-1)} + \eta_c x_i \quad (3.7)$$

Burada

$$\eta_c = \frac{1}{v_c}$$

v_c merkeze atanmış gözlem sayısını temsil eder (Bejar, 2013).

Bu güncelleme yöntemi, bir tür ağırlıklı ortalama işlevidir ve yeni gelen verinin merkeze etkisini iterasyonlar ilerledikçe azaltır. Bu sayede algoritmanın zaman içinde yakınsaması (convergence) garanti altına alınır.

Ufak Örneklem K-Ortalamalar, klasik yönteme kıyasla her tekrar esnasında daha az bilgiyle çalıştığı için yakınsama kriteri farklı şekilde ifade edilir. Genellikle, belirli sayıda iterasyon boyunca kümelerde bir değişiklik gözlenmemesi, algoritmanın durağanlaştığına işaret eder. Bu özellik, yüksek hacimli veri setlerinde algoritmanın uygulanabilirliğini artırır (Bejar, 2013; Wahyuningrum ve ark., 2021).

Ufak Örneklem K-Ortalamalar'ın en önemli avantajı hesaplama süresi açısından sağladığı tasarruftur. Ne var ki bu tasarruf, özellikle daha ufak ufak örneklem boyutları için kümelerin ayrışım kalitesinde az miktarda da olsa azalmaya neden olabilir. Zappia ve ark. (2021) çalışmasında, ufak örneklem boyutunun artırılmasıyla klasik K-Ortalamalar'a benzer kaliteye ulaşmanın mümkün olduğu gösterilmiştir. Örneğin, ufak örneklem boyutunun 500 ve üzeri olduğu durumlarda elde edilen küme merkezleri ve kümeler arası ayrışma düzeyleri, tüm veriyle çalışan klasik K-Ortalamalar ile neredeyse eşdeğerdir (Zappia ve ark., 2021). Bu nedenle özellikle yüksek boyutlu ve büyük hacimli veri setlerinde Ufak Örneklem K-Ortalamalar, pratikte oldukça etkili bir alternatif olarak öne çıkar (Wahyuningrum ve ark., 2021).

3.2.2. Seyrek Temel Bileşenler Analizi (Sparse PCA) Yöntemi

Geleneksel Temel Bileşenler Analizi (PCA), yüksek boyutlu veri setlerinde değişken sayısını azaltarak veri yapısındaki temel örüntüleri ortaya çıkarmada yaygın olarak kullanılan bir boyut indirgeme yöntemidir (Zou ve ark., 2006). Klasik PCA yönteminde temel bileşenler, tüm değişkenlerin ağırlıklı doğrusal birleşimi olarak tanımlandığı için, özellikle değişken sayısının fazla olduğu veri setlerinde bu bileşenlerin ekonomik ve yapısal olarak yorumlanması güçleşebilmektedir (d'Aspremont, 2008).

Seyrek Temel Bileşenler Analizi (Seyrek PCA- Sparse PCA) metodu, klasik PCA'nın sahip olduğu yorumlama yapma yetisinin kısıtlı olması problemini çözebilmek maksadıyla geliştirilen alternatif bir metottur. Seyrek PCA yöntemi, temel bileşenlerin yükleme (loading) vektörlerinde katsayılardan bazılarının sıfıra indirgenmesini sağlayarak, yalnızca anlamlı ve güçlü etkisi olan değişkenlerin modelde kalmasını hedefler (Johnstone ve ark., 2009). Bu sayede hem boyut indirgeme hem de değişken seçimi aynı anda gerçekleştirilmiş olur.

Seyrek PCA sayesinde elde edilmiş olan temel bileşenler, sadece belli sayıdaki değişkenin doğrusal birleşiminden oluşur. Bu durum, veri yapısının yorumlanabilirliğini ciddi ölçüde artırırken, değişken seçimi problemiyle de doğrudan ilişkili hale gelir (Cai, 2013).

Klasik PCA, veri setine ait kovaryans matrisini $\sum$ ile ifade edersek (Zou ve ark, 2006), varyansı maksimize eden bir yükleme vektörü v bulmayı amaçlar:

$$max_v v^T \sum v \quad \text{koşulu ile} \quad \|v\|_2 = 1 \tag{3.8}$$

Buradaki $||v||_2$ ifadesi, vektörün L_2 normunu, yani uzunluğunu ifade etmektedir (Zou ve ark., 2006). Bu optimizasyon problemi sonucunda elde edilen v vektöründeki tüm katsayılar genellikle sıfır olmayan değerler almaktadır.

Seyrek PCA metodu, klasik PCA hesaplamasına bir ceza terimi eklenmesiyle yükleme katsayılarının bazılarının sıfırla indirgenmesine katkıda bulunur. Bunun için genellikle L_1 normuna dayalı bir ceza fonksiyonu (lasso constraint) kullanılarak aşağıdaki şekilde değiştirir:

$$max_v v^T \sum v - \rho\|v\|_1 \quad \text{koşulu ile} \quad \|v\|_2 = 1 \tag{3.9}$$

Burada $\|v\|_1$ ifadesi, vektörün elemanlarının mutlak değerlerinin toplamını, $\|v\|_2$ ifadesi, vektörün uzunluğunu ifade ederken; ρ parametresi ise modeldeki seyrekliğin (sıfırlanacak katsayıların) derecesini kontrol eder (d'Aspremont, 2008). Bu parametre ne kadar büyük seçilirse, modelde sıfırlanacak katsayı sayısı da o kadar fazla olur. Bu formülasyon sayesinde, vektördeki birçok katsayının sıfıra yaklaşması sağlanarak temel bileşenlerin yalnızca belirli bir değişken alt kümesinden oluşması mümkün hale gelir (Cai, 2013).

Son yıllarda geliştirilen bazı yöntemler ile Seyrek PCA problemi, regresyon temelli Seyrek PCA (LASSO veya Elastic Net penalizasyonu ile) (Zou ve ark., 2006), yarı-pozitif tanımlı programlama (Semidefinite Programming - SDP) tabanlı formüle edilmiş yöntemler (d'Aspremont, 2008), kavramsal olarak en iyi Seyrek PCA modelleri (Deshpande ve ark., 2014), uç gözlem içeren veri kümeleri için geliştirilmiş Robust Sparse PCA (ROSPCA) metodu

(Hubert, 2015) gibi yöntemler geliştirilerek, özellikle gözlem sayısının değişken sayısına oranla çok daha küçük olduğu yüksek boyutlu veri setlerinde bile sparse yüklemeler etkin bir şekilde hesaplanabilmektedir.

3.2.3. Literatür Taraması

Elhamifar ve ark. (2011), birden fazla doğrusal olmayan manifold üzerinde konumlanmış verilerin kümeleme ve boyut indirgemesini aynı anda gerçekleştiren Seyrek Manifold Kümeleme ve Yerleştirme (Sparse Manifold Clustering and Embedding – SMCE) algoritmasını önermişlerdir. Elde edilen deneysel sonuçlar, önerilen yöntemin çok yakın konumlanmış manifoldlar, homojen olmayan veri dağılımı ve eksik veri bölgeleri gibi zorluklara karşı dayanıklı olduğunu ve manifoldların içsel boyutlarını başarılı bir şekilde kestirebildiğini göstermektedir.

Chen ve ark. (2012), seyrek grafiklerde kümelenme problemini ele alan çalışmada, çekirdek konveks programlama tabanlı yeni bir algoritma geliştirilmiştir. Algoritma, farklı hata türlerini ayrı ağırlıklarla değerlendirerek seyrek grafiklerde yüksek doğrulukta kümelenme sağlamıştır.

Jha ve ark.. (2015), perakende sektöründe seyrek zaman serileriyle çalışmak için ürünleri semantik özelliklere göre benzerlik tahmin ederek kümelendiren yeni bir yöntem geliştirilmiştir. Böylece daha az veriye sahip olan ürünler anlamlı gözlem grupları olarak model yapılmış ve tahmin başarısı yükseltilmiştir.

Tzagkarakis ve ark. (2015), finansal zaman serilerinde öğrenilen sözlükler üzerinden seyrek modelleme ile hem veri sıkıştırma hem de

analiz süreçleri geliştirilmiştir. FTS-SRC (Financial Time Series-Sparse Representation Coding) yöntemi, volatilite kümeleme ve tahminleme işlemlerinde yüksek doğruluk ve verimlilik sağlamıştır.

Benidis ve ark. (2017), finansal verilerde seyrek portföy optimizasyonu için yeni bir model geliştirilmiş; LAIT (Linear Approximation for Index Tracking) ve SLAIT (Specialized LAIT) algoritmaları ile hem takip hatası minimize edilmiş hem de işlem maliyeti azaltılmıştır. Yöntem, büyük portföylerde uygulanabilirliği ile öne çıkmıştır.

Liu ve ark. (2017) çalışmalarında, Seyrek Gömülü K-Ortalamalar (SE) algoritmasını geliştirerek seyrek rasgele projeksiyon metodu ile boyut azaltma işlemini hızlandırmışlardır. SE algoritması, büyük ve seyrek veri setlerinde hem hızlı hem de yüksek doğrulukta kümelenme sağlamıştır.

Qiang ve ark. (2020) yapmış oldukları çalışmada, kanser alt türlerini belirlemek amacıyla özellik etkileşim ağı kullanan yeni bir benzerlik metriği (Feature Alignment Similarity-FAS) geliştirilmiştir. NetAP algoritması ile yüksek boyutlu ve seyrek mutasyon verilerinde biyolojik olarak anlamlı kümeler elde edilmiştir. Bu metod, klasik yöntemlere kıyasla daha yüksek doğruluk sağlamıştır.

Balsor ve ark. (2021) yaptıkları çalışmada, insan beyninin moleküler gelişimi için seyrekliğe dayalı kümeleme yöntemlerinden Sağlam Seyrek K-Ortalamalar Kümeleme (Robust Sparse K-Means clusterering – RSKC) algoritmasını kullanarak, insan beyninin gelişiminde yaş gruplarına göre başarılı kümelenmeler elde edilmiştir. Özellikle küçük örneklemli ve yüksek boyutlu verilerde klasik

yöntemlere göre daha işlevsel sonuçlar sağlanmıştır. RSKC yöntemi, aykırı değerlere karşı dayanıklı yapısıyla öne çıkmıştır.

Löffler ve ark. (2022), seyrek Gaussian karışım modeli (sparse Gaussian mixture model) üzerinden temel bileşenler analizi (PCA) tabanlı yeni bir kümeleme algoritması geliştirmiştir. Önerilen yöntem minimax optimal hata oranına ulaşırken, düşük örneklem boyutunda polinomsal zamanlı algoritmaların başarısız olacağını teorik olarak göstermiştir. Çalışmada kuramsal analiz ön plandadır.

Huang ve ark. (2023), yüksek boyutlu ve seyrek kategorik verilerin kümelenmesinde geleneksel yöntemlerin yetersiz kaldığını belirtmiş; k-FreqItems algoritmasını geliştirmiştir. Algoritmada, kümelerin merkezleri sık kullanılan boyutlar (FreqItem) ile tanımlanmıştır. Ayrıca, büyük veri kümelerinde tohumlama (seeding) problemini çözmek için SILK yöntemi geliştirilmiştir. Algoritma, gerçek veri setlerinde düşük hata ve yüksek yorumlanabilirlik sağlamıştır.

Centofanti ve ark. (2024), fonksiyonel verilerin seyrek ve pürüzsüz yapısını dikkate alarak geliştirilen SaS-Funclust yöntemi ile hem kümeler hem de bilgi verici domain bölgeleri başarılı şekilde tespit edilmiştir. Model, yüksek yorumlanabilirlik ve doğruluk sağlamaktadır.

3.2.4. Kullanılan Veri ve Yöntemler

Bu çalışma kapsamında 2019 – 2023 yılları arasında Avrupa Birliği'ne üye 27 ülkenin, 34 farklı inşaat ve endüstriyel kullanım için taş ocakçılığı ve madencilik ürünlerinin ithalat/ihracat miktarlarını ve değerlerini içeren bir veri seti kullanılmıştır. İlgili veri seti Avrupa İstatistik Ofisi (Eurostat) sitesinden elde edilmiştir.

Zaman kesiti almanın kümeler üzerinde bir etkisi görülmediği için çalışmada 2019 – 2023 yıllarının tamamı ayrıştırılmadan dikkate alınmıştır. Uygulanacak analizler için Python programı tercih edilmiştir.

İlk olarak; veri setindeki boş gelen değerler sıfıra çok yakın değerler ile otomatik olarak doldurulmuştur. Ardından veri setinde herbir değişkenin dağılımı farklılaştığı için veriler Z-Skor ile standardize edilmiştir. Kullanılan veri bir panel serisi verisi olması sebebiyle dağılımların çarpıklığını (simetrisini) düzeltmek ve daha doğru sonuçlar elde edebilmek adına, standardize edilmiş olan veri setinin logaritması alınarak analizlere sokulmuştur.

Veri hazırlama işlemlerinin tamamlanması akabinde veri setinde 136 değişken bulunması sebebiyle sadece önemli değişkenleri analizde kullanabilmek için, seyrek kümeleme analizinde sıkça kullanılan ve ilgisiz değişkenlerin katsayılarını sıfıra indirgeyen L_1 düzeltmesi, boyut indirgemeyi sağlayarak değişken seçimi yapmayı sağlayan seyrek temel bileşenler analizi ve L_1 düzeltmesine benzer şekilde çalışan, Python üzerinden manuel yazılmış bir fonksiyon ayrı ayrı denenmiştir.

L_1 düzeltmesi veri setindeki tüm değişkenlerin katsayılarını sıfıra indirmiştir. Ayrım ve boyut indirgeme gücü olarak kullanılan veri setinde başarılı sonuç vermemesi nedeniyle boyut indirgeme işlemi seyrek temel bileşenler analizi yöntemiyle yapılmıştır. Seyrek temel bileşenler analizi sonrası çıkan sonuçlar ufak örneklem k-ortalamalar kümeleme analizine sokularak test edilmiştir.

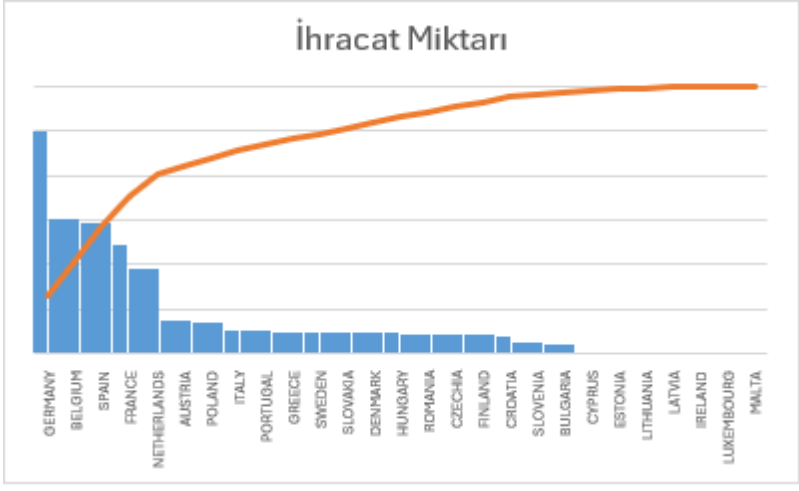
Sonrasında Python üzerinde fonksiyon oluşturularak L_1 düzeltmesine benzer şekilde çalışacak ve bağımsız değişkenlerin

katsayılarını düşürecek veya sıfıra çekecek bir fonksiyon tanımlanarak seyrek kümeleme modeli uygulanmıştır.

Son olarak; bu iki yöntemin sonuçları karşılaştırılıp, hangi modelin daha iyi olduğu ve ne amaçla kullanılabileceği açıklanmıştır.

3.3. SONUÇLAR

Herbir ülkenin 5 yıllık süre zarfı için ithalat miktarı, ithalat değeri, ihracat miktarı ve ihracat değeri kendi içlerinde ortalamaları alınarak herbir madencilik ve taş ocakçılığı malzemesi için Avrupa Birliği ülkelerinin ticaret hacmi görülmek istenmiştir.

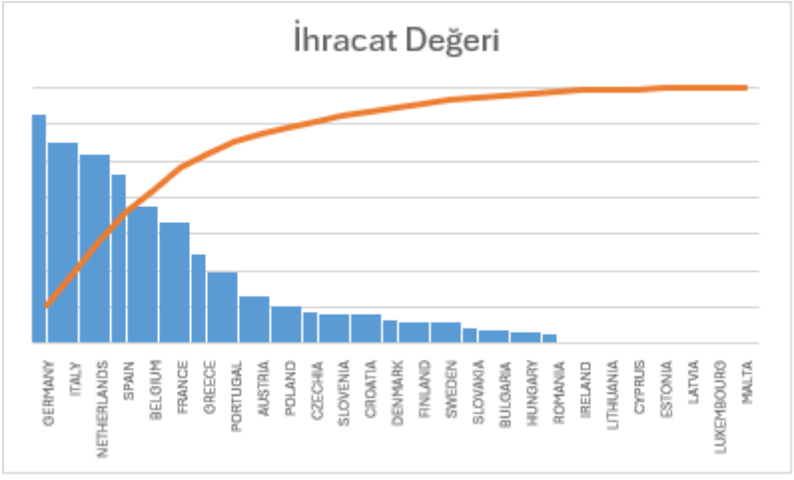


Şekil 3.2. Avrupa Birliği Ülkelerinin İhracat Miktarları

Şekil 3.2'de en yüksek ihracat miktarına sırasıyla Almanya Belçika, İspanya, Fransa ve Hollanda'nın sahip olduğu görülmektedir. Turuncu ile gösterilmiş olan çizgi ise bu ülkelerin taş ocakçılığı ve

madencilik ihracat miktarında ne oranda rol aldıklarını göstermekte ve çizginin bitiş noktası %100'ü ifade etmektedir.

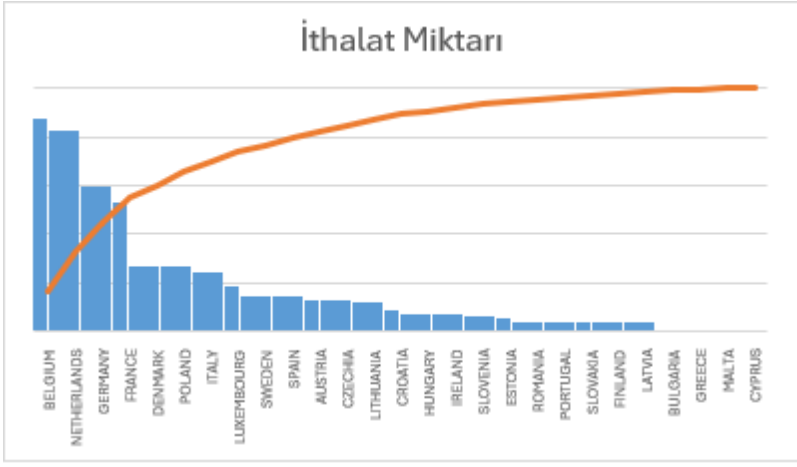
Şekil 3.3, taş ocakçılığı ve madencilik ihracat değerlerinin sırasıyla en yüksek Almanya, İtalya, Hollanda, İspanya ve Belçika'ya ait olduğunu göstermektedir. Belçika'nın daha fazla ihracat yapmasına karşın (Şekil 3.2), yaptığı taş ocakçılığı ve madencilik ürünlerinin satış değerinin daha düşük olduğu yorumu yapılabilir.



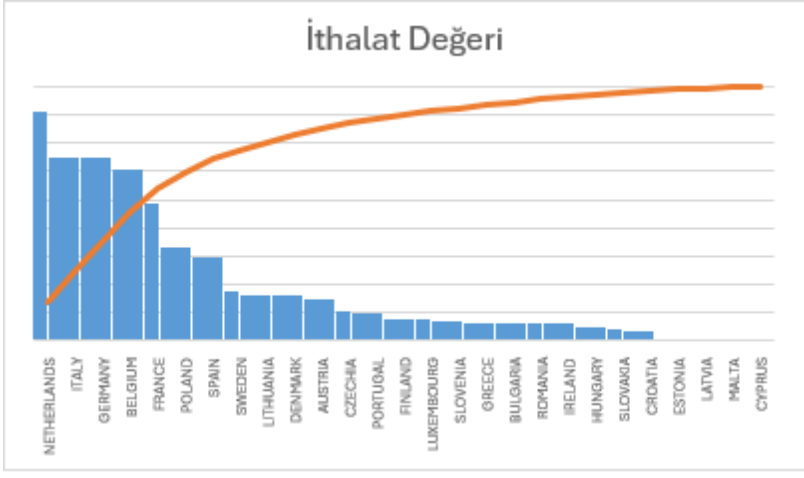
Şekil 3.3. Avrupa Birliği Ülkelerinin İhracat Değerleri

Şekil 3.4, taş ocakçılığı ve madencilik ithalat miktarlarını göstermektedir. Grafik incelendiğinde, sırasıyla en fazla ithalat miktarına Belçika, Hollanda, Almanya ve Fransa sahiptir. Belçika hem yüksek ihracat (Şekil 3.2) hem de yüksek ithalat yapıyor yorumu yapılabilir. Kıbrıs ise daha fazla ihracat yapmakta (Şekil 3.2) ve en az ithalat yapan ülke konumundadır.

Şekil 3.5, Avrupa Birliği'ne üye ülkelerin yapmış olduğu taş ocakçılığı ve madencilik ürünlerinin ithalat değerini göstermektedir. Hollanda, İtalya, Almanya ve Belçika sırasıyla ithalata en fazla harcama yapan ülkelerdir.



Şekil 3.4. Avrupa Birliği Ülkelerinin İthalat Miktarları



Şekil 3. 5. Avrupa Birliği Ülkelerinin İthalat Değerleri

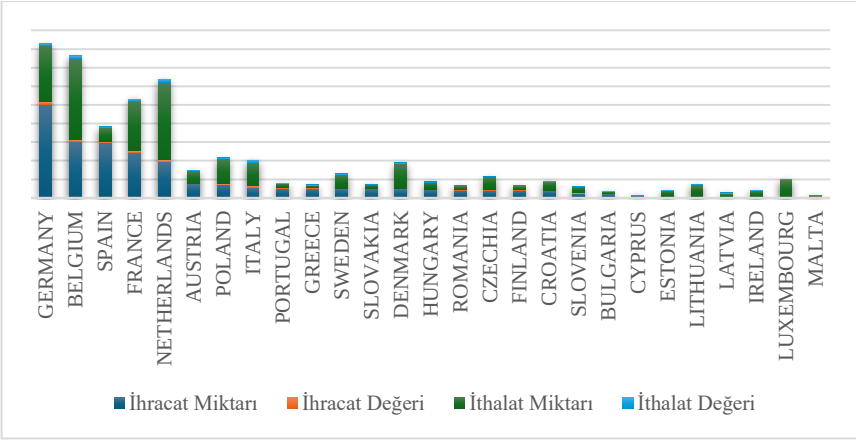
Bazı ülkeler ihracat ve ithalat hacmi arasında dengeli bir yapıya sahipken, bazı ülkelerde bu dengenin bozulduğu görülmektedir.

Tablo 3.1. Bazı Ülkelerde İthalat/İhracat Dengesi

Ülke	İhracat Miktarı	İthalat Miktarı	Yorum
Avusturya	3.72 milyar	3.43 milyar	Dengeli ticaret yapısına sahip.
Belçika	15.2 milyar	21.9 milyar	İthalat ihracattan daha yüksek.
Bulgaristan	1.10 milyar	632 milyon	İhracat ithalattan daha yüksek.
Hırvatistan	2 milyar	2.4 milyat	İthalat ihracattan daha yüksek.

Belçika ve Hırvatistan gibi ülkelerde ithalat ihracattan daha fazla hacme sahiptir. Bu durum bu ülkelerin dışarıdan hammadde veya ara malı temin eden, üretimi ithalata dayalı ekonomilere sahip olabileceğini gösterir. Bulgaristan gibi ülkelerde ise ihracat hacmi ithalat hacminden yüksek olup ihracat odaklı büyüme stratejisi izledikleri söylenebilir.

İhracat ve ithalat miktarları kadar bu ürünlerin değerleri de önem taşımaktadır. Belçika'nın ihracat ve ithalat değerlerinin diğer ülkelere nazaran daha yüksek olması; katma değere sahip ürünlerin ticaretinde daha fazla bulunduğu şeklinde yorumlanabilir. Kıbrıs gibi ülkelerde hem miktar hem de değer düşük olduğu için, büyük oranda düşük hacimli ve düşük değerli ürünlerin ticaretinin söz konusu olduğu söylenebilir.

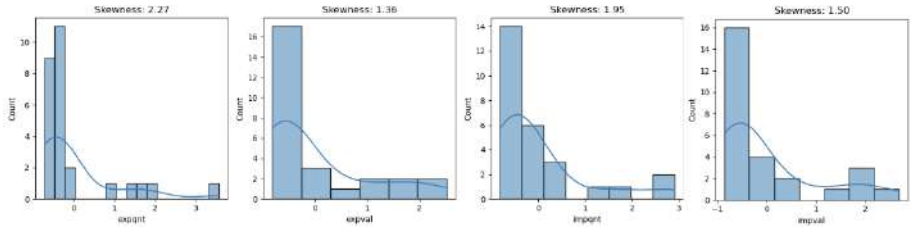


Şekil 3.6. AB Ülkelerinin İthalat/İhracatına İlişkin Bilgileri

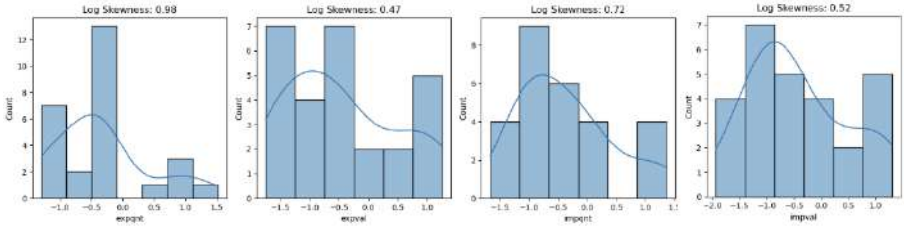
Kullanılan veride, bazı değişkenlerin çeşitli yıllarda boş geldiği görülmüş olup, seyrek kümelemenin böyle bir yapıda daha iyi çalıştığı bilinmektedir. Bu sebeple boş gelen değerler analiz esnasında sıfıra çok yakın bir değerle doldurulmuş olup bu yapının bozulmaması hedeflenmiştir.

Kullanılan veri seti panel bir veri seti olması sebebiyle gözlemler logaritmaları alınarak daha simetrik bir dağılım haline getirilmiştir. Aşağıda Şekil 3.7'de gözlemlerin logaritması alınmadan önceki lineer model halindeki çarpıklıklar, Şekil 3.8'de ise logaritması alındıktan sonraki çarpıklıkları yer almaktadır.

Sırasıyla seyrek PCA + ufak örneklem k-ortalamalar ve seyrek k-ortalamalar analizleri uygulandığında küme ayırma gücü olarak en optimal sonuçları bulmak için sırasıyla 3 küme ve 2 kümeye ayrılarak analizler yapılmıştır.

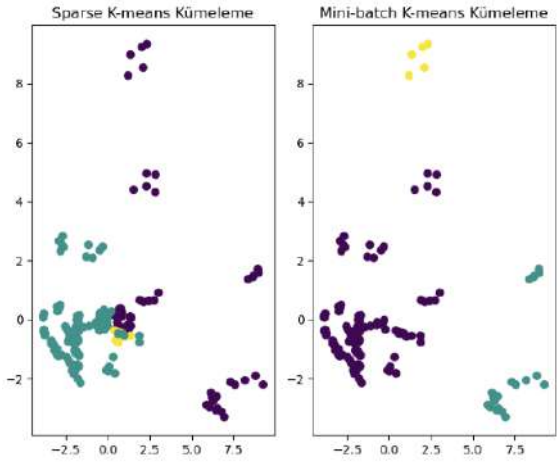


Şekil 3.7. Logaritmik Dönüşüm Öncesi Verilerin Ortalamalarının Çarpıklıkları



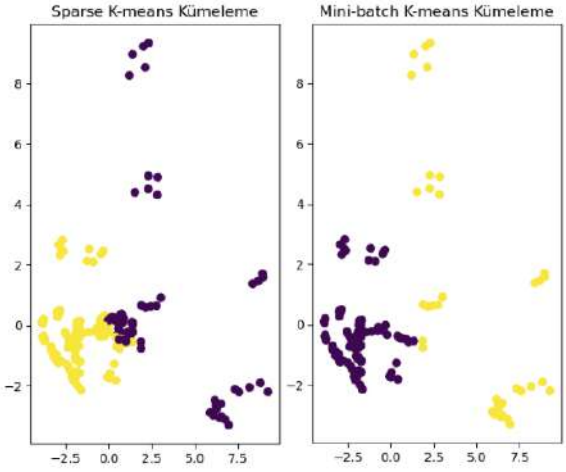
Şekil 3.8. Logaritmik Dönüşüm Sonrası Verilerin Ortalamalarının Çarpıklıkları

3 kümeye ayrıldığı takdirde algoritmaların küme ayırma gücü sırasıyla seyrek k-ortalamalar modeli için 0.46 ve ufak örneklem k-ortalamalar modeli için 0.56 çıkmıştır. Bu değerler bire yaklaştıkça kümelerin ayrılma gücü ve küme içi homojenlikleri artmaktadır. Bu, 3 küme ile çalışmaya devam edildiği takdirde en iyi ayırma gücüne ufak örneklem k-ortalamalar kümeleme sahip olacağı şeklinde yorumlanabilir. Seyrek k-ortalamalar algortiması ise daha düşük ayırma gücüne sahip olmuştur.



Şekil 3.9. 3 Kümeye Ayırma Sonucu Dağılımların Karşılaştırması

2 kümeye ayrıldığı takdirde algoritmaların küme ayırma gücü sırasıyla 0.53 ve 0.56 çıkmıştır. Küme sayısının artırılması ufak örneklem k-ortalama algoritmasının ayırma gücünü değiştirmemiş, seyrek k-ortalama algoritmasının ise daha homojen olacağını göstermiştir.



Şekil 3.10. 2 Kümeye Ayırma Sonucu Dağılımların Karşılaştırması

3 küme için seyrek k-ortalamalar (Tablo 3.2) ve ufak örneklem k-ortalamalar (Tablo 3.3) sonucunda oluşan kümelerde yer alan ülkeler aşağıdaki gibidir.

Tablo 3.2. Seyrek K-Ortalamalar ile 3 Kümeye Ayırma Sonucu Kümelerdeki Gözlemler

Küme	Ülkeler				
1	AUSTRIA	BULGARIA	CROATIA	CYPRUS	DENMARK
	SLOVENIA	HUNGARY	IRELAND	LATVIA	LITHUANIA
	ESTONIA	FINLAND	LUXEMBOURG	MALTA	SWEDEN
	PORTUGAL	ROMANIA	SLOVAKIA	GREECE	
2	BELGIUM	FRANCE	GERMANY	ITALY	
	POLAND	SPAIN	NETHERLANDS		
3	CZECHIA				

Tablo 3.3. Ufak Örneklem K- Ortalamalar ile 3 Kümeye Ayırma Sonucu Kümelerdeki Gözlemler

Küme	Ülkeler				
1	AUSTRIA	BULGARIA	CROATIA	CYPRUS	CZECHIA
	FINLAND	GREECE	HUNGARY	IRELAND	LATVIA
	MALTA	ROMANIA	SLOVAKIA	SLOVENIA	SWEDEN
	DENMARK	ESTONIA	LITHUANIA	LUXEMBOURG	
2	BELGIUM	FRANCE	GERMANY	NETHERLANDS	
3	ITALY	POLAND	PORTUGAL	SPAIN	

Her iki yöntemde de 2. küme ağırlıklı olarak Batı Avrupa ülkelerinden oluşmaktadır. Bu durum, bu ülkelerin benzer sosyo-ekonomik veya yapısal özellikleri taşıdığını göstermektedir.

Seyrek k-ortalamalarda (Tablo 3.2) Czechia, tek başına 3. kümeye düşerken; ufak örneklem k-ortalamalarda (Tablo 3.3) İtalya, Portekiz ve İspanya ile birlikte üçüncü kümeyi oluşturmuştur.

Ufak Örneklem K-Ortalamalar'da kümeler daha dengeli dağılmış görünürken, Seyrek K-Ortalamalar daha "ayırt edici" davranarak bazı ülkeleri (örneğin Czechia) tekil kümeye atamıştır. Bu durum, seyrek yöntemin değişken seçimi üzerinden daha ayrıştırıcı olabileceğini göstermektedir.

2 küme için sırasıyla seyrek k-ortalamalar ve ufak örneklem k-ortalamalar sonucunda oluşan kümelerde yer alan ülkeler aşağıdaki Tablo 3.4 ve Tablo 3.5'teki gibidir.

Tablo 3.4. Seyrek K-Ortalama ile 2 Kümeye Ayırma Sonucu Kümelerdeki Gözlemler

Küme	Ülkeler				
1	AUSTRIA	BULGARIA	CROATIA	CYPRUS	CZECHIA
	FINLAND	GREECE	HUNGARY	IRELAND	LATVIA
	MALTA	PORTUGAL	ROMANIA	SLOVAKIA	SLOVENIA
	DENMARK	ESTONIA	LITHUANIA	LUXEMBOURG	
2	BELGIUM	FRANCE	GERMANY	ITALY	SWEDEN
	NETHERLANDS	POLAND	SPAIN		

Tablo 5. Ufak Örneklem K-Ortalama ile 2 Kümeye Ayırma Sonucu Kümelerdeki Gözlemler

Küme	Ülkeler				
1	AUSTRIA	BULGARIA	CROATIA	CYPRUS	CZECHIA
	FINLAND	GREECE	HUNGARY	IRELAND	ITALY
	LUXEMBOURG	MALTA	POLAND	PORTUGAL	ROMANIA
	DENMARK	ESTONIA	LATVIA	LITHUANIA	SLOVAKIA
	SLOVENIA	SPAIN	SWEDEN		
2	BELGIUM	FRANCE	GERMANY	NETHERLANDS	

2 kümeye ayrılmış senaryoda her iki yöntemde de Belgium, France, Germany ve Netherlands 2. kümede yer almakta ve bu durum bu ülkelerin ekonomik/demografik olarak birlikte hareket ettiklerini göstermektedir.

Seyrek K-Ortalama (Tablo 3.4) İsveç'i de bu Batı Avrupa kümesine dahil ederken, Ufak Örneklem K-Ortalama (Tablo 3.5) Sweden'i 1. kümeye yerleştirerek daha homojen bir Kuzey-Doğu Avrupa grubu oluşturmuştur.

Ortak gözlemler arasında Finlandiya, Estonya, Danimarka ve Baltık ülkelerinin istikrarlı şekilde aynı kümelerde yer aldığı görülmektedir. Bu durum, coğrafi veya ekonomik yakınlıklarının kümelenme üzerinde etkili olduğunu göstermektedir.

3.4. TARTIŞMA

Bu analiz, Avrupa Birliği ülkeleri arasındaki taş ocakçılığı ve madencilik ürünlerinin ticaret hacmindeki farklılıkları ortaya koymaktadır. Bu farklılıklar; ülkelerin ekonomik büyüklüğü, üretim kapasiteleri, ithalata bağımlılıkları ve ticaret stratejileri ile doğrudan ilişkilidir. İthalat ve ihracat değerleri ile miktarları, bir ülkenin dış ticaret hacmini yansıtarak, kümeleme analizlerinin anlamlılığını artırmaktadır.Bu çalışmada ülkeler arasında dış ticaret yapıları açısından yapısal benzerlikleri tespit etmek amacıyla iki farklı kümeleme stratejisi uygulanmıştır:

- Seyrek K-Ortalamalar (Sparse K-Means) yöntemi hem değişken seçimi hem de kümeleme işlemini aynı anda gerçekleştirerek,

kümelerin oluşmasında etkili olan değişkenleri doğrudan belirlemektedir.

- Seyrek PCA + Ufak Örneklem K-Ortalamalar yöntemi öncelikle Seyrek PCA ile değişken boyutu azaltılmış, ardından elde edilen bileşenler kullanılarak ufak örneklem k-ortalamalar yöntemiyle kümeleme yapılmıştır.

1. Küme: Bu kümede, Avrupa Birliği'nin coğrafi ve ekonomik olarak çevresel konumunda bulunan, ticaret hacmi düşük ya da orta seviyede olan ülkeler yer almaktadır. Bu ülkeler, iç ticarette göreli olarak dengeli bir yapıya sahipken, dış ticarette sınırlı etki göstermektedir. AB içindeki ekonomik entegrasyona (ortak pazara ve mali iş birliklerine) katılım düzeyleri düşük, üretim kapasiteleri sınırlı ve dışa açıklıkları da zayıftır. Ancak bu ülkelerin, ekonomik yapıları istikrarlıdır ve ani dalgalanmalardan uzak bir seyir izler. Bu özellikleriyle bu küme, düşük ama yatay büyüme eğiliminde olan ve birbirine yapısal olarak benzeyen ülkelerden oluşmaktadır.

2. Küme: Bu kümede Belçika, Fransa, Almanya ve Hollanda gibi yüksek dış ticaret hacmine sahip, Avrupa'nın ekonomik açıdan merkezinde yer alan ülkeler bulunmaktadır. Bu ülkeler yüksek Gayrisafi Yurtiçi Hasıla (GSYİH) seviyelerine, gelişmiş sanayi altyapılarına ve birbirine benzer dış ticaret dengelerine sahiptir. Avrupa Birliği içinde güçlü ekonomik entegrasyon sergilemekte, hem birbirleriyle hem de küresel piyasalarla yoğun ticari ilişkiler kurmaktadırlar. Bu küme, dışa açık, büyük ölçekli ve yüksek düzeyde entegre ekonomilerin oluşturduğu bir merkez yapısını yansıtmaktadır.

3. Küme (yalnızca 3 kümeli analiz için): Bu kümede İtalya ve İspanya gibi Akdeniz ülkeleri yer almakta ve tarım, turizm ve hizmet sektörlerinin baskın olduğu özgün ekonomik yapılarıyla ayrışmaktadır. Dış ticaret yapıları mevsimsel dalgalanmalara daha açıktır ve enerji ithalatına yüksek bağımlılık göstermektedirler. Bu durum, dış ticaret açıklarının da göreli olarak daha yüksek olmasına neden olmaktadır.

Her iki stratejiyle elde edilen doğruluk oranları karşılaştırıldığında, ufak örneklem yöntemi 2 ve 3 kümeli analizlerde %56 doğruluk sağlarken, seyrek k-ortalama yöntemi sırasıyla %53 ve %46 doğruluk sunmuştur. Ancak seyrek k-ortalamalar yöntemi aşağıdaki nedenlerle tercih edilmiştir:

1. Seyrek K-ortalamalar yöntemi, kümeler üzerinde etkili olan değişkenleri doğrudan belirleyerek sonuçların ekonomik ve yapısal olarak yorumlanabilirliği artırır. Oysa Sparse PCA sonrası elde edilen bileşenler, değişken bazında doğrudan yorumlanamaz.

2. Seyrek k-ortalamalar algoritmasında değişken seçimi ile kümeleme işlemi aynı anda yürütüldüğü için süreç daha şeffaftır. Seyrek PCA + Ufak Örneklem K-Ortalamalar yaklaşımında bu iki aşamanın ayrı olması, analizin bütünselliğini zayıflatır.

3. Seyrek PCA ile elde edilen bileşenler, birden fazla değişkenin ağırlıklı birleşiminden oluşur. Bu bileşenlerin kümeler üzerinde ne şekilde etkili olduğu, ekonomik yapılarla doğrudan

ilişkilendirilemez. Dolayısıyla, sonuçlar sınıflandırma açısından başarılı olsa da, hangi yapısal değişkenlerin kümelenmeye neden olduğu net biçimde görülemez.

4. Seyrek k-ortalama yöntemi, özellikle küçük ama etkili değişken gruplarının kümeler üzerindeki belirleyiciliğini vurgular. Böylece, ülkeler arasındaki ekonomik ve ticari yapısal farklar daha net biçimde görünür hâle gelir ve sonuçların yorumlanabilirliği artar. Seyrek PCA'da bu etkiler ortalamalaşabilir veya bileşenlere dağılabilir.

Seyrek PCA ile boyut indirgeme sonrası yapılan ufak örneklem kümeleme, daha yüksek doğruluk oranları sunsa da, bu yöntemin sınırlı yorumsal gücü ve analizdeki aşamaların ayrık olması, yöntemin uygulama değerini düşürmektedir. Buna karşılık seyrek k-ortalama yöntemi, doğrudan değişken bazlı ayrıştırma, yorumsal şeffaflık, ve yapısal ayrışmaların net bir şekilde izlenebilmesi gibi nitel avantajlar sunmaktadır.

Bu nedenle, daha düşük doğruluk oranlarına rağmen analitik tutarlılığı, yapısal anlamlandırma kapasitesi ve değişken düzeyinde açıklayıcılığı yüksek olan seyrek k-ortalama yöntemi tercih edilmiştir.

KAYNAKÇA

Arias-Castro, E., & Pu, X. (2017). A simple approach to sparse clustering. arXiv preprint arXiv:1602.07277.

Balsor, J. L., Arbabi, K., Singh, D., Kwan, R., Zaslavsky, J., Jeyanesan, E., & Murphy, K. M. (2021). A practical guide to sparse k-means clustering for studying molecular development of the human brain. Frontiers in Neuroscience, 15, 668293. https://doi.org/10.3389/fnins.2021.668293

Bejar, J. (2013). K-means vs Mini Batch K-means: A Comparison. Universitat Politècnica de Catalunya. R13-8.

Benidis, K., Feng, Y., & Palomar, D. P. (2017). Sparse portfolios for high-dimensional financial index tracking. IEEE Transactions on signal processing, 66(1), 155-170.

Brodny, J., & Tutak, M. (2020). The use of artificial neural networks to analyze greenhouse gas and air pollutant emissions from the mining and quarrying sector in the European Union. Energies, 13(8), 1925.

Cai, T.T., Ma, Z., & Wu, Y. (2013). Sparse PCA: Optimal Rates and Adaptive Estimation. The Annals of Statistics.

Centofanti, F., Lepore, A., & Palumbo, B. (2024). Sparse and smooth functional data clustering. Statistical Papers, 65(2), 795-825.

Chang, X., Wang, Y., Li, R., & Xu, Z. (2018). Sparse k-means with l∞/l0 penalty for high-dimensional data clustering. Statistica Sinica, 28(3), 1265-1284. https://doi.org/10.5705/ss.202015.0261

Chavent, M., Lacaille, J., Mourer, A., & Olteanu, M. (2020, October). Sparse k-means for mixed data via group-sparse clustering. In ESANN 2020-28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (Vol. 978).

Chen, Y., Sanghavi, S., & Xu, H. (2012). Clustering sparse graphs. Advances in neural information processing systems, 25.

d'Aspremont, A., Bach, F., & El Ghaoui, L. (2008). Optimal Solutions for Sparse Principal Component Analysis. Journal of Machine Learning Research.

Deshpande, Y., & Montanari, A. (2014). Information-theoretically optimal sparse PCA. *IEEE Transactions on Information Theory*, 62(5), 2741-2759.

Elhamifar, E., & Vidal, R. (2011). Sparse manifold clustering and embedding. Advances in neural information processing systems, 24.

European Union. “Business economy by sector - NACE Rev. 2” . Erişim: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Business_economy_by_sector_-_NACE_Rev._2. Erişim Tarihi: 18 Nisan 2025.

Floriello, D., & Vitelli, V. (2017). Sparse clustering of functional data. Journal of Multivariate Analysis, 154, 1–18.

Gaynor, S., & Bair, E. (2017). Identification of relevant subtypes via preweighted sparse clustering. Computational Statistics & Data Analysis, 116, 139–154.

George, A. (2021). Economic and environmental issues of quarrying. J Emerg Technol Innov Res, 8(3), 41-45.

Huang, Q., Luo, P., & Tung, A. K. (2023). A new sparse data clustering method based on frequent items. Proceedings of the ACM on Management of Data, 1(1), 1-28.

Hubert, M., Reynkens, T., Schmitt, E., & Verdonck, T. (2015). Sparse PCA for high-dimensional data with outliers. Technometrics.

Jha, A., Ray, S., Seaman, B., & Dhillon, I. S. (2015, April). Clustering to forecast sparse time-series data. In 2015 IEEE 31st international conference on data engineering (pp. 1388-1399). IEEE.

Johnstone, I.M., & Lu, A.Y. (2009). Sparse Principal Components Analysis. ArXiv preprint arXiv:0901.4392.

Liu, W., Shen, X., & Tsang, I. (2017). Sparse Embedded $ k $-Means Clustering. *Advances in neural information processing systems*, *30*.

Löffler, M., Wein, A. S., & Bandeira, A. S. (2022). Computationally efficient sparse clustering. Information and Inference: A Journal of the IMA, 11(4), 1255-1286.

Qiang, J., Ding, W., Kuijjer, M., Quackenbush, J., & Chen, P. (2020). Clustering sparse data with feature correlation with application to discover subtypes in cancer. IEEE Access, 8, 67775-67789.

Tzagkarakis, G., Caicedo-Llano, J., & Dionysopoulos, T. (2015). Sparse modeling of volatile financial time series via low-dimensional patterns over learned dictionaries. Algorithmic Finance, 4(3-4), 139-158.

Vouros, A., & Vasilaki, E. (2020). A semi-supervised sparse k-means algorithm. arXiv preprint arXiv:2003.06973. https://arxiv.org/pdf/2003.06973

Wahyuningrum, T., Khomsah, S., Meliana, S., Yunanto, P. E., Suyanto, S., & Al Maki, W. F. (2021). Improving clustering method performance using K-Means, Mini Batch K-Means, BIRCH and Spectral. In 2021 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 226–231). IEEE. https://doi.org/10.1109/ISRITI54043.2021.9702823

Wang, B., Zhang, Y., Sun, W. W., & Fang, Y. (2017). Sparse convex clustering. Journal of Computational and Graphical Statistics, 26(4), 747–757.

Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. Journal of the American Statistical Association, 105(490), 713–726.

Yang, C., Robinson, D., & Vidal, R. (2015). Sparse subspace clustering with missing entries. International Conference on Machine Learning (ICML), 2463–2472.

Zappia, L., et al. (2021). Clustering trees: A visualization for evaluating clusterings at multiple resolutions. PLoS Computational Biology, 17(2), e1008625. https://doi.org/10.1371/journal.pcbi.1008625

Zhang, R., & Lu, Z. (2016). Large scale sparse clustering. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), 333–339.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics.

BÖLÜM 4:

ZAMAN SERİLERİNDE TEKRARLAYAN YAPILARIN KESİT VERİ MODELLEMESİNDE HAREKETLİ EPSİLON TABANLI YEREL ORTALAMA YAKLAŞIMI: İKLİMSEL VERİLERDE BİR UYGULAMA

Muzaffer GÖZTAŞ[7]
Nida ORUÇ[8]
Doğan YILDIZ[9]

Özet

Zaman serileri analizlerinde, bağımlı ve bağımsız değişkenleri kullanarak klasik doğrusal regresyon, ARIMAX, SARIMAX ve benzeri yöntemlerin uygulanabilmesi için bazı temel varsayımların sağlanması gerekmektedir. Bu çalışmada, zaman serileri varsayımlarından yola çıkılarak önerilen teorik temelli yaklaşımda zaman serilerini klasik doğrusal regresyon modellerinde uygulanabilirliğini kolaylaştırmak için epsilon uzaklıklı kümeler kullanılarak zaman serisi verilerinden kesit veri üretimi önerilmiştir. Ancak bu teorik önerinin veri seti üzerinden uyulabilmesi için zaman serileri analizlerinde test edilen bazı varsayımların sağlanması gerekmektedir. Örneğin, değişken kendi içinde trendsiz olmalı, aynı

[7] Arş. Gör., Yıldız Teknik Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, İstanbul, Türkiye, muzaffer.goztas@yildiz.edu.tr, 1ORCID ID: 0009-0001-7933-901X
[8] Arş. Gör., İstanbul Topkapı Üniversitesi, İktisadi, İdari ve Sosyal Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü, İstanbul, Türkiye, nidaoruc@topkapi.edu.tr, ORCID ID: 0009-0005-0230-2234
[9] Dr. Öğr. Üyesi,Yıldız Teknik Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, İstanbul, Türkiye, dyildiz@yildiz.edu.tr, ORCID ID: 0000-0001-7430-2368

dereceden durağanlığa sahip olmalı ve kesit kümeleri arasında yüksek korelasyonun sağlanıyor olması gerekmektedir.

Yöntemin etkinliği, iklim verileri kullanılarak sıcaklık tahmini üzerinden test edilmiştir. İlk olarak DBSCAN denetimsiz kümeleme algoritması ile zaman serisi verileri kümelenmiş ve anomaliler belirlenerek analiz süreçlerinden çıkartılmıştır. Ardından, elde edilen kesit verileri Elastik-Net regresyon modeliyle analiz edilmiştir. Elastik-Net modelinde, çoklu doğrusal bağlantı problemini yönetmek için farklı L (Alpha) ve l1_ratio değerleri üzerinden hiperparametre optimizasyonu gerçekleştirilmiştir. Böylece katsayılar küçültülerek değişken seçim de sürece dâhil edilmiştir.

Sonuçların, önerilen teorik yöntemin tahminleme gücünde ve bu veri seti üzerinde olan testlerinde %79 R2 ve %78 düzenlenmiş R2 değerleri ile iyi performans göstermiş olup hata terimlinin normalliği varsayımı da sağlanmıştır. Bu yaklaşım ile özellikle klasik zaman serileri analizlerinden karşılaşılan varsayımların zorluklarının giderilmesi süreçlerini biraz daha daraltarak regresyon modellerinde genelleştirebilirliğini ve uygulanabilirliğini arttırması amaçlanmıştır.

MOVING EPSILON-BASED LOCAL MEAN APPROACH FOR CROSS-SECTIONAL DATA MODELING OF RECURRING STRUCTURES IN TIME SERIES: AN APPLICATION ON CLIMATIC DATA

Abstract

In time series analyses, basic assumptions must be met to apply classical linear regression, ARIMAX, SARIMAX, and similar methods using dependent and independent variables. In this study, based on time series assumptions, a theoretically grounded approach is proposed to facilitate the application of time series data in classical linear regression models by generating cross-sectional data from time series data using epsilon-distance clusters. However, for this theoretical proposal to be implemented on datasets, certain assumptions tested in time series analysis must be satisfied. For example, variables must be trend-free within themselves, possess the same order of stationarity, and there must be high correlations between variables in the cross-sectional clusters

The effectiveness of the method was tested on climate data through temperature prediction. Initially, time series data were clustered using the DBSCAN unsupervised clustering algorithm, and anomalies were identified and excluded from the analysis process. Subsequently, the cross-sectional data obtained were analyzed using the Elastic-Net regression model. In the Elastic-Net model, hyperparameter optimization was performed through different *L* (*Alpha*) and *l1_ratio* parameter values to manage the multicollinearity problem. Thus, coefficient shrinkage and variable selection were included in the process.

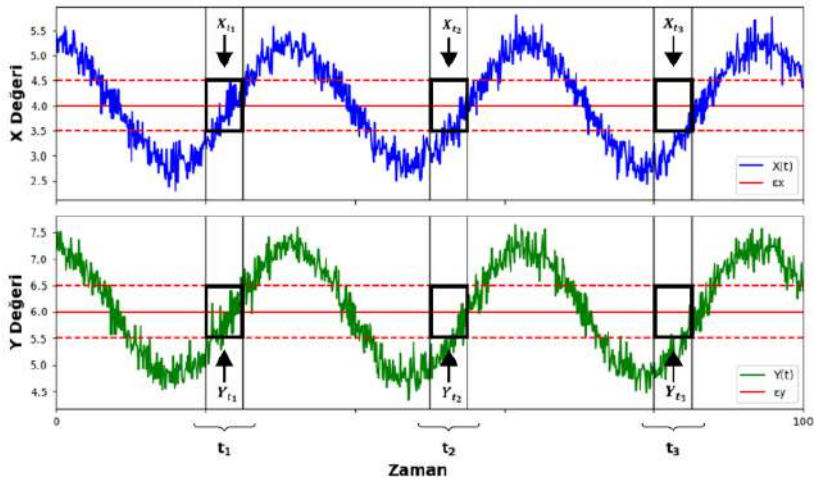
The results indicated that the proposed theoretical method performed well in predictive power and tests conducted on this dataset, achieving an R^2 value of 79% and an adjusted R^2 value of 78%, along with meeting the normality assumption of the error terms. This approach aims to reduce the difficulties encountered with assumptions in classical time series analysis, thereby enhancing the generalizability and applicability of regression models.

4.1. GİRİŞ

Zaman serileri analizlerinde bağımsız değişkenler ile bağımlı değişkeni doğrudan regresyon modelleri kullanarak modellenmesinde farklı varsayımların sağlanması gerekmektedir. Bunun için zaman serileri analizlerinde ARIMAX veya mevsimselliğin olduğu serilerde ise SARIMAX gibi yöntemler tercih edilmektedir (Box vd., 2015; Hyndman & Athanasopoulos, 2021). Bu işlemin uygulanabilir olması için bağımsız ve bağımlı değişkenlerde aynı dereceden durağanlık şartı aranmaktadır (Kirchgässner vd., 2013). Ancak bu koşullar sağlandığı noktada veriler klasik regresyona uyarlanabilirlik kazanabilir.

Doğrudan uyarlanabilir olmamasının temel nedenlerinden birisi klasik doğrusal regresyon modellerinde kesit verileri üzerinden analizlerin yapılabiliyor olmasıdır. Bu noktada uyarlanabilirlik için bağımlı ve bağımsız değişkenlerin bulunduğu zaman serisi değişkenleri arasında yüksek korelasyon olması, trend (eğilim) bulunmaması ve aynı dereceden durağanlık olması durumunda serilerde epsilon uzaklıklı kümelerin ortalamaları üzerinden kesit veri yaklaşımı oluşturulabilir (Lütkepohl, 2005).

Bu yaklaşımın uyarlanabilir olması için varsayımlardan birisi ise zaman aralıklarında serilerin diğer bir değişken üzerindeki izdüşümleri ile olan serinin ortalama kesiti ile oransal olarak yakın ve/ya aynı ana kütleden gelme varsayımlarını sağlıyor olması beklenir. Buna bir örnek olarak Şekil 4.1'deki X ve Y olmak üzere 2 değişken üzerinden teorik açıdan simüle edilmiştir.



Şekil 4.1. Ortalama Yaklaşımlı Zaman Serisi İz Düşüm Kesitleri

Gürültülü Uygulamaların Yoğunluk Tabanlı Uzamsal Kümelenmesi (DBSCAN) algoritmasına ait epsilon (ε) parametresinin X ve Y değişkenleri için değerleri $\varepsilon_X = 0{,}5$ ve $\varepsilon_Y = 0{,}5$ olarak belirlenmiş olsun. Böylece her değişken için ilgili kümeler oluşturulsun ve tüm veriler her 2 değişken için de 0.5 uzaklıktaki değerler ile kümelenmiş olsun. Elde edilen küme yapılarında X değişkenine ait bir küme için merkez noktayı X=4 için ele alalım, ε parametresinden dolayı ±0.5 küme aralığı olacak biçiminde kırmızı noktalı sınırları Şekil 4.1'deki gibi belirlemiş olalım.

Farklı zaman dilimlerinde aynı değer ve örüntünün yakalandığı tekrarlı gözlemlerin aynı küme aralıkları ele alınsın, bu izdüşümü Y değişkeni içinde gerçekleştirdiğimizde üst üste denk gelen yapı ve örüntüler arası ilişkilerin varlığı ortalamalar yaklaşımını kullanarak kesit veri örüntü modelini oluşturmamıza olanak sağlamaktadır. Ancak örüntülerin ortalamaları üzerinden oluşturulan kesitlerin bilimsel

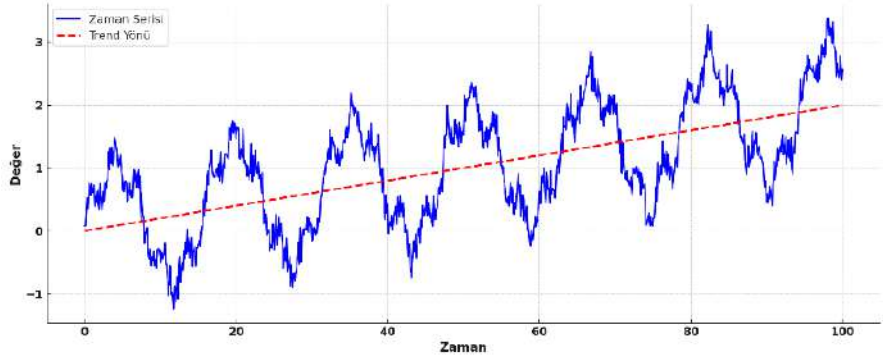
literatüre uygun bir biçimde sağlanabilmesi için temel varsayımları karşılaması gerekmektedir. Bu yaklaşımın modellenebilmesi için veriler üzerinde sağlanması gereken varsayımlar aşağıda yer alan Tablo 1'de listelenmiştir.

Tablo 1. Veri Ortalamasından Kesit Veri Üretim Varsayımları

Değişken	***X***	***Y***
Epsilon	ε_X	ε_Y
Zaman Aralığı	$t_1, t_2, \cdots, t_n$	
Zaman Boyutu	$n_{t_1}, n_{t_2}, \cdots, n_{t_n}$	
Korelasyon	$\underset{Corr(X_{t_i}, Y_{t_i})}{max} Corr(X_{t_i}, Y_{t_i}) \neq 0$	
Durağanlık	$I(d)[X(t), Y(t)]$	
Trend	$X(t) = \mu + \epsilon_t, (\beta_t = 0)$	$Y(t) = \mu + \epsilon_t, (\beta_t = 0)$
Zaman Serisi (Örüntü)	$X_{t_{1,1}}, \cdots, X_{t_{1,n_{t_1}}}, \cdots, X_{t_{n,1}}, \cdots, X_{t_{n,n_{t_n}}}$	$X_{t_{1,1}}, \cdots, X_{t_{1,n_{t_1}}}, \cdots, X_{t_{n,1}}, \cdots, X_{t_{n,n_{t_n}}}$
Küme	I	J
Kesit Veri (Ortalama)	$\bar{X}_I = \frac{\sum_{k=t_{1,n_{t_1}}}^{t_{n,n_{t_n}}} X_k}{n_{t_1} + n_{t_2} + \cdots + n_{t_n}}$	$\bar{Y}_J = \frac{\sum_{k=t_{1,n_{t_1}}}^{t_{n,n_{t_n}}} Y_k}{n_{t_1} + n_{t_2} + \cdots + n_{t_n}}$
Regresyon Kesiti	$(x, y) = (\bar{X}_I, \bar{Y}_J)$	

Her kesitin kendi içerisinde yüksek negatif ya pozitif korelasyonlu olması gerekmektedir. Örneğin, aynı iz düşüme sahip t_1 zaman aralığındaki zaman serisi verileri X_{t_1}ve Y_{t_1} serilerinin aralarındaki korelasyon yüksek olmalıdır. Aynı durum Şekil 4.1'de bulunan ve benzer biçimde Tablo 4.1'de detaylıca işlenen $t_1, t_2, \ldots, t_n$ zaman aralıklarında bulunan örüntülerin izdüşümlerine ait

korelasyonları olan $Corr(X_{t_2}, Y_{t_2})$, $Corr(X_{t_3}, Y_{t_3})$, ..., $Corr(X_{t_n}, Y_{t_n})$ için de geçerlidir.



Şekil 4.2. Artış Eğilimli Zaman Serisi Örneği

Serilerden elde edilen ortalama kesit yaklaşımı sadece eğilim (trend) içermeyen ve aynı dereceden durağanlık içeren verilerde anlamlılık kazanır. Bu noktadaki amaç, aynı iz düşüme sahip kesitlerin hareketlerindeki benzerliklerin, aynı eğriyi veren farklı zaman dilimlerinde yakalanmasıdır (Bhattacharjee vd., 2019). Bir veri seti zaman içerisinde artış ve azalış açısından aynı örüntüyü sağlıyor olabilir ancak bu örüntü bir trend yapısına sahip olabilir veya durağan olmayabilir. Şekil 4.2 buna bir örnek olarak verilebilir. Grafik belirli bir örüntüde ilerliyor ancak trend yönüne baktığımızda artan bir eğilimin varlığını kırmızı noktalı artan eğri ile de görmekteyiz (Dette & Heinrichs, 2020). Bu ve benzeri trend içeren durumlarda ortalama bazlı kesitsel yaklaşımlar yanlış sonuçların üretilmesine neden olabilir. Çünkü örüntünün tekrarlanmasında, her tekrarlı benzer örüntü noktasının aynı olduğunu söyleyemeyiz (Epskamp vd., 2016).

Şekil 4.1'de simüle gösterimde verilerin ortalama bazlı yaklaşımları için Tablo 4.1'de gösterilen *X* bağımsız değişkeni için $X(t)=\mu+\varepsilon_t$, *Y* bağımlı değişkeni için ise $Y(t)=\mu+\varepsilon_t$ denkliğini β_t trend parametresinin 0 olması şartı altında sağlıyor olması gerekmektedir. Bu koşul, verilerin durağan bir süreç izlediğini ve eğilim (trend) içermediğini varsayar (Dickey & Fuller, 1979). Ortalama bazlı yaklaşımların uygulanabilirliği, bu denkliklerin zaman serisi verilerinde sabit bir ortalama etrafında rastgele dalgalanmalar sergilemesine bağlıdır (Kwiatkowski vd., 1992). Trend parametresinin sıfır olması, serilerin uzun vadeli eğilimlerden arındırılmış olduğunu ve yalnızca rastgele hata terimi (ε_t) içerdiğini garanti eder (Phillips & Perron, 1988).

Çalışmada zaman serisindeki tekrarlı örüntülerin oluşturduğu ortak kümelerden elde ettiğimiz ortalama (kesit) verilerinin modellenebilmesi için Elastik-Net algoritması kullanılmıştır (Zou & Hastie, 2005). Elastik-Net, Lasso ve Ridge regresyonlarının özelliklerini bir araya getiren ve maliyet fonksiyonunu minimize etmeyi amaçlayan bir algoritmadır. Penalizasyon ve aşırı uyum açısından son derece güçlü bir algoritma olmasının yanında özellik seçimi yaparak gereksiz değişkenlerin modelden çıkartılmasına olanak tanır.

Bağımsız değişkenler arası yüksek korelasyon sonucu ortaya çıkan çoklu doğrusal bağlantıyı önlemek için de kullanılan bu algoritma özellikle yüksek boyutlu veri setlerinde, aşırı uyumu (Overfitting) önlemek için düzenlileştirme (regularization) uygular (Tibshirani, 2011). L_1 (Lasso) terimi gereksiz özellikleri tamamen sıfırlayıp

değişken seçimi yaparken, L_2 (Ridge) terimi ağırlıkları küçülterek modelin genelleme performansını artırır. Böylece, Elastik-Net regresyonu hem model sadeliğini korur hem de aşırı uyum riskini azaltır. Özetle hem L_1 hem de L_2 düzenlileştirme tekniği kullanarak model geliştirme sürecinde olası hataları önleme doğrultusunda çalışan bir tür denetimli öğrenme algoritmasıdır (Friedman vd, 2010).

$$RSS_{ElasticNet} = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2 + L\left(\eta \sum_{j=1}^{p}|\beta_j| + (1-\eta)\sum_{j=1}^{p}(\beta_j)^2\right) \qquad (4.1)$$

Formül 4.1'de belirtilen η ifadesi Lasso modeline ait olan ağırlık katsayısı; $(1-\eta)$ ise Ridge modelinin ağırlık katsayısını temsil etmektedir. Ağırlık η katsayısı 0 ile 1 arasında bir ağırlıklandırmaya sahiptir. Çalışmanın giriş bölümünde örnek bir veri ile modellediğimiz zaman serilerinin kesit veriye dönüştürülmesi ve kullanılan denetimli makine öğrenmesi algoritmasının teorik açıklamalarıyla birlikte bahsedilmiştir. Bu teorik yaklaşımın kullanılabilirliğini göstermek açısından iklim verileri üzerinden bir uygulama gerçekleştirilmiştir. Bu uygulamada tüm değişkenlerimizin zaman serisi verisi olduğu ve teorik uygulamamıza örnek olarak seçilen *Temperature* (*Sıcaklık*) bağımlı değişkeninin *Radiation* (*Radyasyon*), *Humidity* (*Nem*), *Pressure* (*Basınç*), *WindDirection* (*Rüzgâr Yönü*) ve *WindSpeed* (*Rüzgâr Hızı*) bağımsız değişkenleri tarafından açıklanabilirliği Elastik-Net algoritması ile test edilmiştir.

4.1.1. Teorik Yaklaşıma Literatürden Örnekler

Çalışmamızın temel odak noktası, zaman serilerinde düzenli aralıklar ile aynı dağılım ve örüntüyü izleyen değişkenlerden elde edilen verilerle regresyonun uygulanabilirliğini inceleyen teorik yaklaşımızdır. Bu yaklaşımda, zaman serilerindeki benzer örüntülerin kümelenmesini ele almakta olup, farklı disiplinlerdeki veri setleriyle literatürde uygulanmasından birkaç örnek ele alınmıştır.

Yamashita vd. (2020), epsilon-tau prosedürüyle zaman serilerinde segmentasyon yaparak yerel trendleri ortaya çıkarmış ve trend kümelerini tespit etmiştir. Bu çalışma, bizim çalışmamız olan epsilon tabanlı mesafeler ile kümelerin belirlenmesinin ardından uyguladığımız zaman serilerinden segmentasyon ile kesit verilerin elde edilmesi teorik yaklaşımına benzerlik gösteren ve altyapısını destekleyen önemli bir literatür katkısı sunmaktadır.

Jaén-Vargas vd. (2020), zaman serisi verilerinde kayan pencere yaklaşımıyla özellikler elde ederek derin öğrenme modellerinde elde edilen pencere boyutlarının model başarısına etkisini gözlemlemiştir. Bu çalışma, iklim verileri üzerinde denediğimiz teorik yaklaşımımızdaki zaman serilenin kümelenmesi ve karşılaştırmalı korelasyon yöntemlerimizle teorik olarak benzer bir yaklaşım sunarken, örüntü analizi ve veri segmentasyonu açısından çalışmamıza altyapı sağlayan literatürden önemli bir örnektir.

Moy vd. (2023), sağlık verilerinden oluşan zaman serilerini DBSCAN algoritmasıyla kümelemiş ve uygun hiperparametre seçiminin model başarısına etkisini elbow gibi grafiksel yöntemler

kullanarak analiz etmiştir. Bu çalışma, DBSCAN kullanarak zaman serileri kümelerini oluşturduğumuz çalışmamıza metodolojik açıdan benzerlik göstermekte olup farklı veri türlerinde örüntü analizinin uygulanabilirliğini destekleyen önemli bir literatür katkısı sunmaktadır.

Djenouri vd. (2022), enerji sistemlerindeki büyük ölçekli zaman serisi verilerinde tekrarlayan desenleri keşfetmek için RNN algoritması ile DBSCAN algoritmasını birleştiren ortak bir çalışma sunmuştur. Bu yaklaşımı, iklim verilerinde uyguladığımız zaman serilerinde kümelenme ve anomali tespiti yöntemlerimize teorik açıdan benzerlik göstermekte olup katkı sunmaktadır. RNN ve DBSCAN kombinasyonunun örüntü gruplarını otomatik olarak tespit etme özelliği, iklim verilerinde yakaladığımız düzenli örüntülerin ve aykırı değerlerin analizine metodolojik bir benzerlik sağlamaktadır.

Dumler vd. (2023), seri üretim hatlarındaki endüstriyel sensör verilerini kullanarak otomatik zaman serisi segmentasyonu ve kümelenmesi yöntemini kullanmıştır. Bu çalışma, çalışmamıza metodolojik olarak altyapı sağlamış olup özellikle örüntü çıkarımı ve süreç takibi açısından teorik katkı sağlamaktadır.

Literatürde, DBSCAN ve temel istatistiksel yöntemler kullanılarak zaman serilerinde örüntü ve tekrarlayan yapıların elde edilmesine yönelik olarak benzer metodolojik yaklaşımlar, bu bölümde beş farklı çalışma örneği üzerinden incelenmiştir. Ele alınan örnekler, metodolojik açıdan çalışmamızın teorik alt yapısını destekler niteliktedir. Literatürde, bu tarz yaklaşımların farklı veri setleri ve

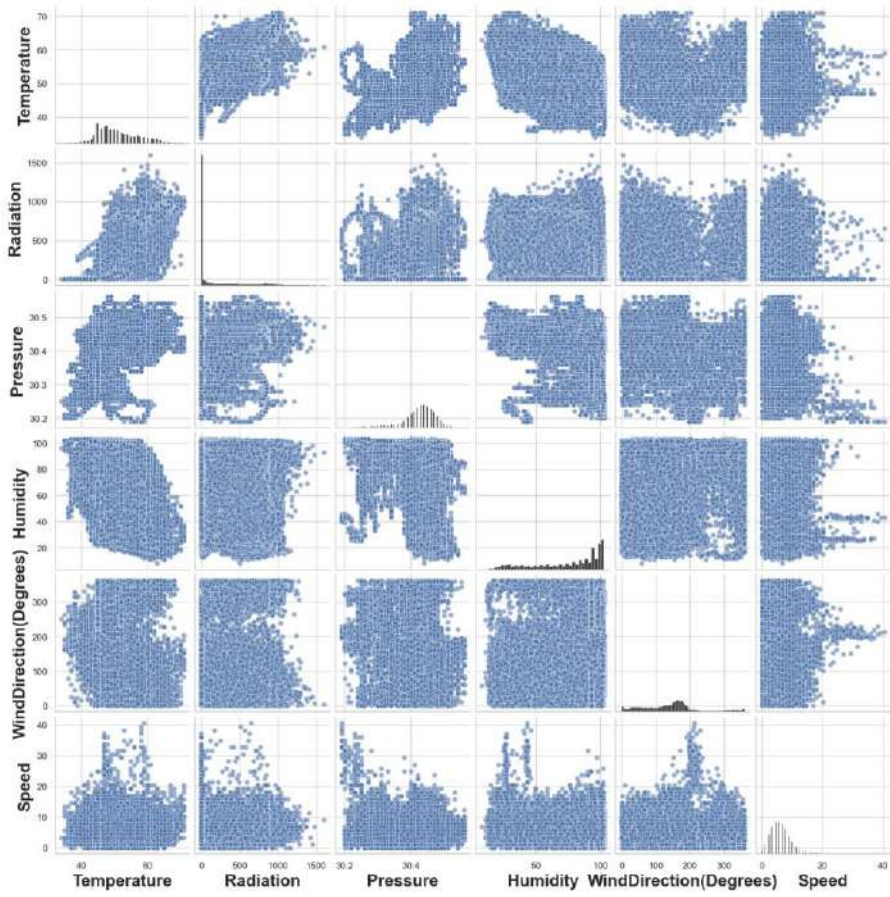
disiplinlerde uygulandığını birçok çalışma bulunmakla birlikte, seçilen örnekler, yöntemin çeşitli alanlardaki kullanımını yansıtmaktadır.

4.2. MATERYAL VE METOD

HI-SEAS uzay ve hava araştırma merkezi tarafından alınan veriler, 4 aylık bir süre boyunca toplanmış yaklaşık 32086 gözlemden oluşmaktadır. Veri seti, güneş radyasyonu, hava sıcaklığı, basınç, nem, rüzgâr yönü ve rüzgâr hızı gibi meteorolojik değişkenleri içermektedir. *Radiation* (W/m²), güneş radyasyonu miktarını metrekare başına watt olarak ifade ederken, *Temperature* (°F), hava sıcaklığını Fahrenheit cinsinden ölçmektedir. *Pressure* (Hg), barometrik basıncı inç cıva sütunu biriminde belirtir ve *Humidity* (%), havadaki nem oranını yüzde olarak ifade eder. Ayrıca, *WindDirection* (°), rüzgâr yönünü derecelerle belirtirken, *WindSpeed* (mph), rüzgâr hızını saatlik mil biriminde ölçmektedir. Bu değişkenler, modelleme ve analizlerde kullanılmak üzere veri setinde tanımlanmıştır.

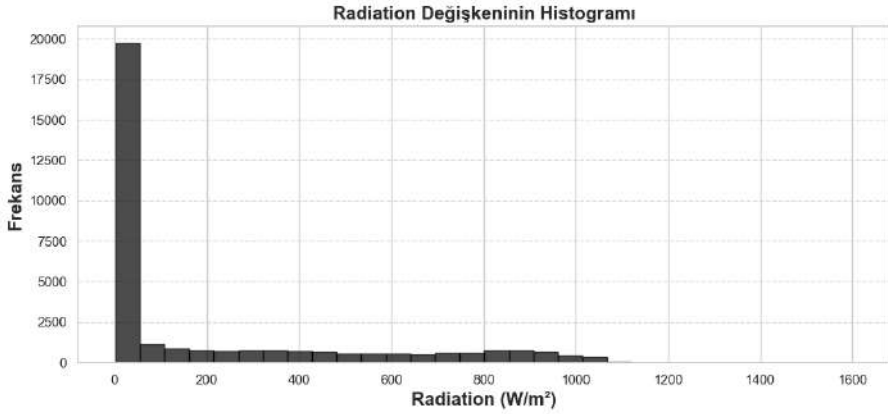
4.2.1. Anomali Tespiti Öncesi Değişkenlerin Dağılımı

Şekil 4.3'te yer alan diyagonal grafiklerde, veri setinde kullanılan altı temel değişkenin kendi iç dağılımları incelenmiştir. ***Temperature*** değişkeninin dağılımı normal dağılıma yakın bir yapı sergilemekte olup, diğer değişkenler belirgin biçimde normal dağılımdan sapmaktadır. Özellikle ***Radiation*** ve *Wind**Speed*** değişkenleri tipik olarak **sağa çarpık** bir dağılım göstermekte olup bu durum, verinin düşük değerlerde yoğunlaşıp yüksek değerlere doğru seyrekleştiğini göstermektedir.



Şekil 4.3. Değişkenlerin Saçılım ve Dağılım Grafiği

Humidity değişkeninde ise bu durumun aksine, yüksek değerlerde yığılma olup dağılım **sola çarpık** bir yapıdadır. ***Pressure*** değişkeninin dağılımı ise **bimodal** karakter taşımakta, yani iki ayrı yoğunluk bölgesiyle dikkat çekmekte ve normal dağılımdan ciddi ölçüde sapmaktadır.



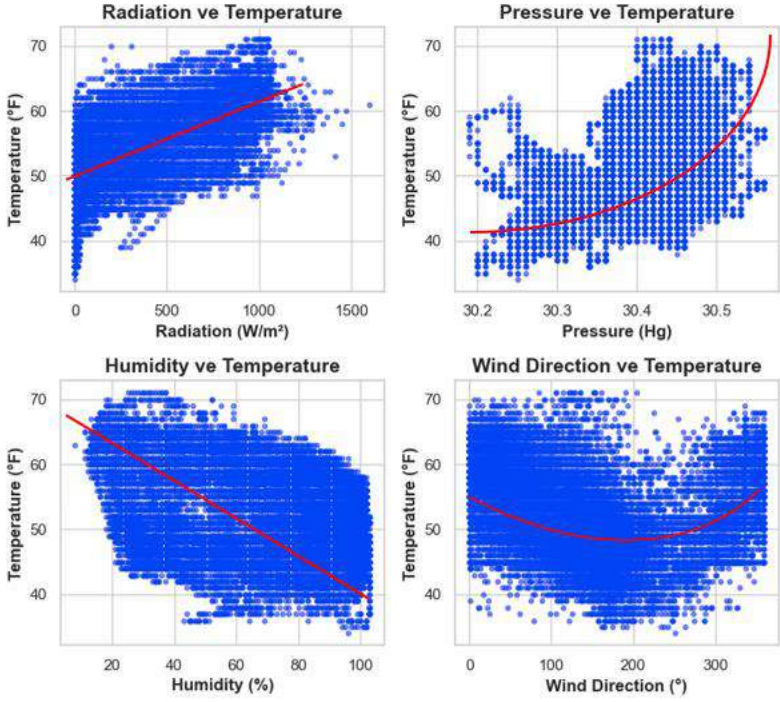
Şekil 4.4. Radiation Değişkeni Histogramı

Şekil 4.4'te yer alan histograma bakıldığında *Radiation* bağımsız değişkenine ait veri setinde sağa çarpıklık ve yanlılık göze çarpmaktadır. Bu yanlılık diğer değişkenlerin histogramlarına göre çok daha yüksek olduğundan model performansını etkileyebilir. Bu noktada çalışmamızın temel varsayım ve amacı doğrultusunda seriler ortalamaları tercih edilmiştir. Bu nedenle de önişleme ve anomali tespitine öncelikle bu değişken üzerinden başlanması gerekmektedir.

Veri önişleme hangi değişken üzerinden anomali gerçekleşeceği tamamen değişkenlerin dağılımı ve birbiriyle olan uyumuna göre değişmektedir.

Şekil 4.5'deki grafiklerde kırmızı bir çizgi grafiği ile ilişki yönleri gösterilen serpilme diyagramları incelendiğinde ise dikkat çeken diğer bir nokta *Temperature* bağımlı değişkeni ile *Radiation* bağımsız değişkeni arasındaki serpilmeye göre aralarında çok muhtemel pozitif yönlü doğrusal bir ilişki olma ihtimali göze çarpmaktadır. *WindSpeed* değişkeninin Şekil 4.3'te yer alan serpilme

diyagramındaki *Temperature* değişkeni ile olan dağılımında herhangi bir yönlü ilişki görülmemektedir. Bu tür görsel analizler verileri yorumlamak için tek başına yeterli olmayabilir.



Şekil 4.5. Olası Güçlü İlişkili Değişkenler

4.2.2. Normallik Varsayımı

Değişkenler üzerinde hangi analizin hangi testler ile yapılacağına karar vermeden önce bu karar için en önemli ve temel testlerden birisi olan dağılım yani normallik sınama testleridir. Bu testler veri setinin dağılımı hakkında bize çok fazla bilgi verirken aynı zamanda ileride uygulanacak diğer analizlerin türleri açısından da yol göstericidir.

Hipotezler:

H_0: Değişken normal dağılıma uyum sağlamaktadır.

H_1: Değişken normal dağılıma uyum sağlamamaktadır.

Tablo 4.2. Bağımlı ve Bağımsız Değişkenlerin Normallik Testleri

Değişken	KS	SW	DAK
Radiation	$p < 0.001$	$p < 0.001$	$p < 0.001$
Temperature	$p < 0.001$	$p < 0.0534$	$p < 0.001$
Pressure	$p < 0.001$	$p < 0.001$	$p < 0.001$
Humidity	$p < 0.001$	$p < 0.001$	$p < 0.001$
WindDirection	$p < 0.001$	$p < 0.001$	$p < 0.001$
WindSpeed	$p < 0.001$	$p < 0.001$	$p < 0.001$

KS: Kolmogorov-Smirnov, SW: Shapiro-Wilk, DAK: D'Agostino's K^2

Bağımlı ve bağımsız değişkenlerin normalliğini test etmek için Kolmogorov-Smirnov, Shapiro-Wilk ve D'Agostino's K^2 Testleri uygulanmış ve Tablo 4.2'deki *p* değerleri elde edilmiştir.

Bir hipotez testinde elde edilen *p* değeri $\alpha = 0.05$ değerinden küçük ise H_1 hipotezinin kabulüne karar verilirken, H_0 hipotezinin kabulü için yeterince bulgu olmadığı yorumu yapılabilir. Model değişkenleri için 3 farklı normallik teste ait *p* değerleri incelendiğinde her değişken için $p<0.001$ olduğundan anlamlılık düzeyi $\alpha = 0.05$ değerinden küçük olduğu görülmektedir. Dolayısıyla bu 6 değişkenden hiçbirinin normallik varsayımına uymadığına karar verilebilir. Ancak bu tarz çalışmalarda normallik varsayımı için genellikle histogram, çarpıklık ve basıklık değerlerine bakılması daha doğru bir yaklaşım olur. Bunun temel nedeni merkezi limit teoreminin de bize göstermiş olduğu gibi büyük veri setlerinde normallik varsayımı için istatistiksel testler

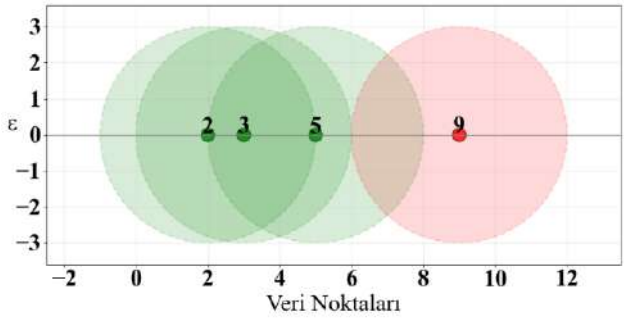
kullanmak normalliği yorumlamak açısından oldukça yanıltıcı olabilir, bunun yerine grafikler üzerinden gitmek daha doğru yaklaşım izlememizi sağlar (Ghasemi & Zahediasl, 2012).

4.2.3. Anomali Tespiti ve Kesitler Ortalaması Yaklaşımı

DBSCAN algoritması nokta ve yoğunluğu temel alarak her nokta için tüm birleşimlerinin denendiği bir kümeleme algoritmasıdır. Çalışma biçiminin temelinde, bir noktanın diğer noktalarla kümelenmesini sağlayan yarıçap değeri olan ε (epsilon) yer almaktadır. Bu değerin yanında ek olarak MinPts adında farklı bir hiperparametresi de mevcuttur. Bu parametre bir kümenin oluşabilmesi için gerekli olan minimum noktayı yani veri noktasını temsil etmektedir (Ester vd., 1996).

Örneğin, hiperparametrelerimiz ε = 3 ve MinPts = 2 alınmış olsun. Bu durumda elimizdeki veri setinde yer alan değerlerin birbiriyle kümelenebilmesi için birbirleri arasındaki farklar 3'ten fazla olmamalıdır ve bir kümenin oluşabilmesi için minimum 1 adet veri noktası gerekmektedir. Bu hiperparametreler için elimizde 4 farklı veri noktası olduğunu düşünelim. Bu noktalar sırasıyla 2, 3, 5, 9 sayılarından oluşsun. Bu sayılar belirtilen epsilon ve MinPts parametreleri altında kümelediğimizde oluşabilecek en uygun küme {2, 3, 5} ve {9} biçiminde olacaktır. 2, 3 ve 5 sayıları arasındaki mutlak farklar 3 epsilon değerinden büyük değildir, bu nedenle birbirleriyle kümelenebilir durumdadırlar. Ancak 9 sayısı bir küme oluşturmaz çünkü epsilon değerine göre başka hiçbir sayı ile birleşemez ve tek başına da bir küme oluşturamaz çünkü bir küme oluşumu için minimum

değer sayısı 2 olarak belirtilmiştir. Bu tür durumlara anomali ya da sapan değer denmektedir.



Şekil 4.6. DBSCAN Algoritma Simülasyon Çıktısı

DBSCAN algoritması için yani Şekil 4.6'daki gibi kümelenemeyen değerlere gürültü (Noise) denmektedir. MinPts hiperparametresi 1 değerinden farklı olmak koşuluyla herhangi bir kümeye dâhil olmayan veriler -1 ile işaretlenirler.

Tablo 4.3. Epsilon Uzaklıklı Radiation Kesit Kümeleri

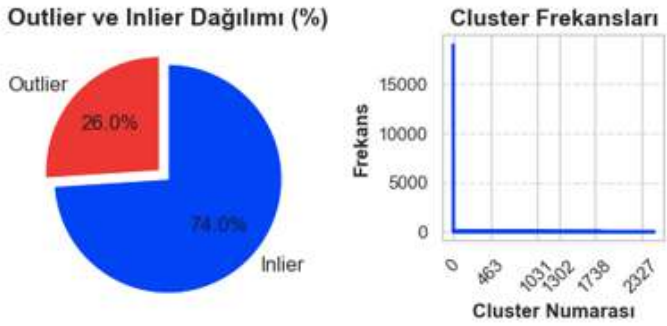
Küme	Frekans	Minimum Değer	Maksimum Değer	Ortalama	Standart Sapma	Aralık
0	18984	1,11	29,13	2,77	4,72	28,02
2	123	47,47	51,48	49,46	1,17	4,01
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1994	1	651,78	651,78	651,78	-	0,00
2477	1	91,64	91,64	91,64	-	0,00

DBSCAN ile kümeleme işleminin ardından elde edilen kümelerin frekanslarının dağılımı ise bize yanlılığı da göstermektedir. Tablo 4.3'te kümelerin frekansları görülmektedir.

Dengeli bir veri yapısının varlığı durumunda, kümeler içerisindeki değerlerin frekanslarında bu ölçüde belirgin bir yanlılık

meydana gelemez. Ancak 0 kümesine ait frekans sayısı 18.984 adet veri yer almaktadır. Bu da veri setinde bir yanlılığa işaret etmektedir. Yaklaşık olarak 32.086 *Radiation* verisinin %58'inde bir noktada yığılma gözlemlenmiştir. Böylece ortalamalar yaklaşımı oluşturulabilir.

Şekil 4.7'de dilim grafiği ile gösterilen grafikte ise elde edilen kümelerin içerisinde sadece 1 adet *Radiation* değeri rastlanması durumunda yani tek başına kümelenen ve anomali olarak hesaplanmış ifadelerin benzersiz kümeler bazlı dağılımı gösterilmiştir.



Şekil 4.7. DBSCAN -1 İşaretli Anomali ve Küme Boyut Dağılımı

Şekil 4.7'de belirtilen 2478 adet kümenin yaklaşık olarak %26'sında frekansı 1 olan anomaliler tespit edilmiştir. Bu değerlerin veri setinden çıkartılması, ileride model geliştirme sürecinde anlamsız sonuçlar üretilmesinin önüne geçilmesini sağlamıştır.

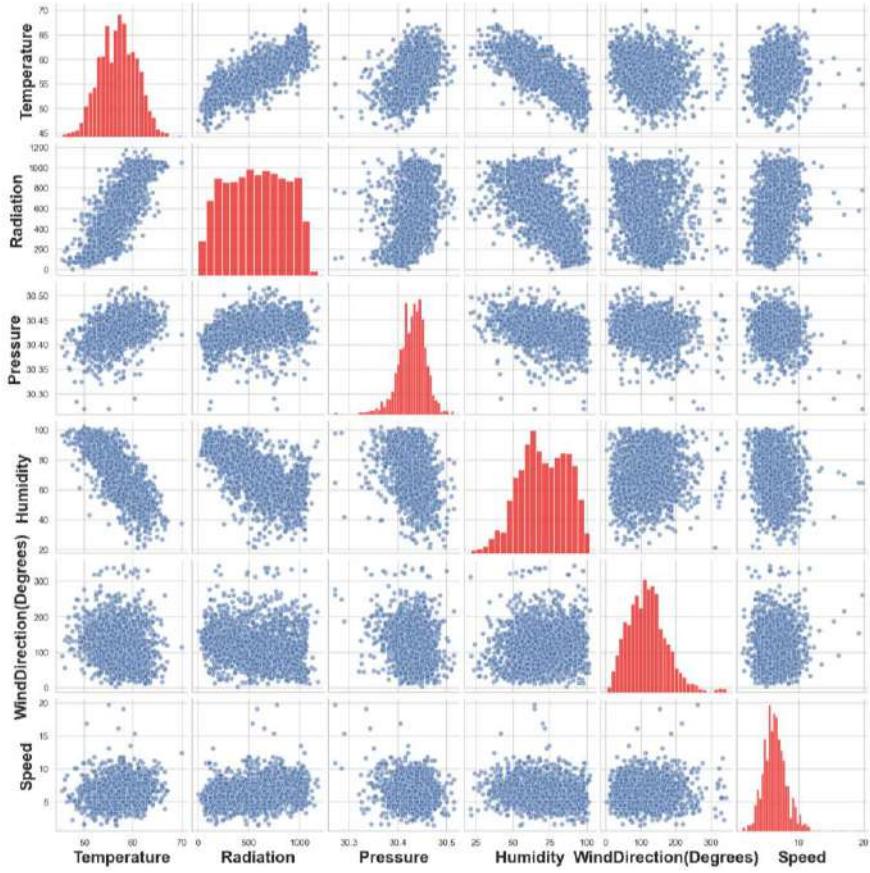
Tablo 4.4'te yer alan ilk dört küme, 2478 adet küme içerisinden anomali içeren kümelerin çıkarılmasının ardından seçilmiş; bu kümelere ait bağımlı ve bağımsız değişkenlerin karşılık gelen değerlerinin ortalamaları alınarak modelleme süreci yürütülmüştür.

Tablo 4.4. Radiation Değişkeni Seri Ortalamalarına Ait Kesit Verileri

Cluster	*Radiation* (x̄)	*Temperature* (x̄)	*Pressure* (x̄)	*Humidity* (x̄)	*Wind Direction* (x̄)	*Wind Speed* (x̄)
0	2.77	47.57	30.42	76.32	159.05	6.19
1	33.66	50.87	30.4	85.0	110.12	5.07
2	49.45	50.58	30.4	90.11	149.23	4.76
3	59.42	51.42	30.4	88.11	135.95	4.95

4.2.4. Anomali Sonrası Dağılımlar

Anomali tespiti sonrası verilerin dağılımı ve saçılım Şekil 4.8'da gösterilmiştir. Anomali öncesi dağılım ve saçılıma göre daha anlamlı örüntüler göze çarpmaktadır.



Şekil 4.8. Anomali Eliminasyonu Sonucu Değişken Dağılımı

Anomali tespiti sonrasında değişkenlerin histogramlarına baktığımızda normal dağılım eğrisine yakınsadıkları görülmektedir. Bu durumu test etmek için normallik testleri yapılmıştır.

Tablo 4.5. Anomali Sonrası Değişken Normallik Test Çıktıları

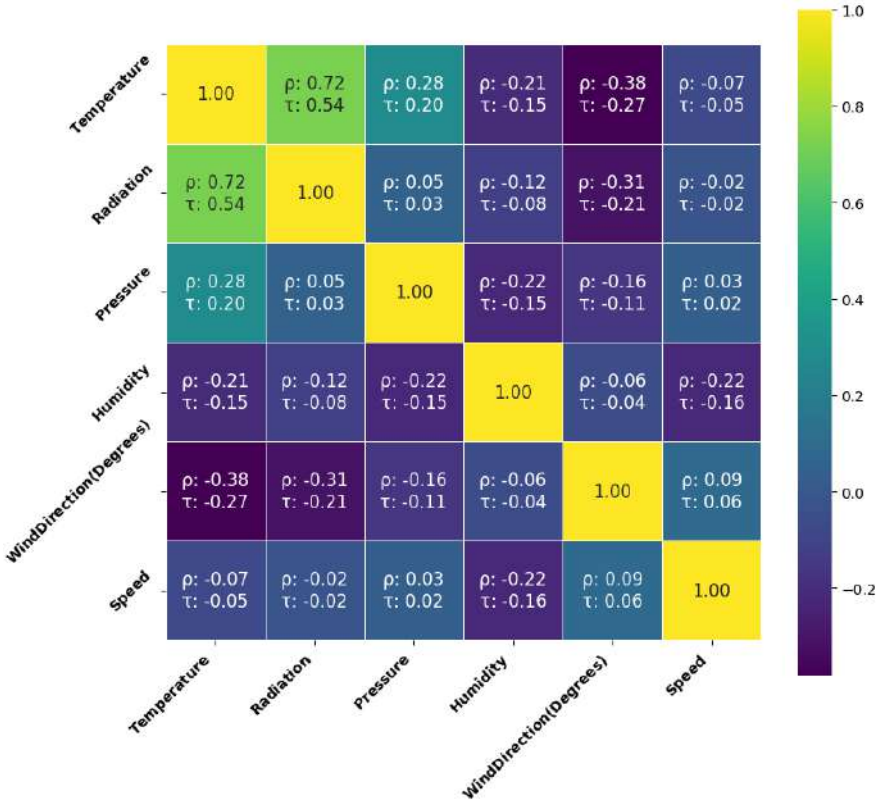
Değişken	KS	SW	DAK
Radiation	$p < 0.001$	$p < 0.001$	$p < 0.001$
Temperature	$p < 0.001$	$p < 0.0534$	$p < 0.001$
Pressure	$p < 0.001$	$p < 0.001$	$p < 0.001$
Humidity	$p < 0.001$	$p < 0.001$	$p < 0.001$
WindDirection	$p < 0.001$	$p < 0.001$	$p < 0.001$
WindSpeed	$p < 0.001$	$p < 0.001$	$p < 0.001$

KS: Kolmogorov-Smirnov, SW: Shapiro-Wilk, DAK: D'Agostino's K^2

Tablo 4.5'te yer alan *Temperature* bağımlı değişkeninin Kolmogorov-Smirnov için verdiği p=0.0534 değeri hariç diğer tüm değişken ve testler için p değerleri 0.05'ten küçük olduğu için bu verilerin normal dağılıma uygun olmadığı sonucuna varılmıştır. Bu durumda, analizlerde normal dağılım varsayımına dayanmayan parametrik olmayan testler üzerinden devam edilmelidir. Değişkenlerde normalizasyonlar yaparak normal dağılıma daha da yakınsaması sağlanabilir. Ancak normallik varsayımı regresyon model geliştirme süreçlerinde değişkenlerin normallik varsayımının aksine model sonucu test verisindeki hataların normal bir dağılım sergilemesi ve eş varyanslılık (Homoskedastisite) açısından önemlidir (Long & Ervin, 2000).

4.2.5. Değişkenler Arası Bağlantısallık

Normal dağılıma uymayan ve aralarındaki bağlantıyı ölçmek istediğimiz verilerde, geleneksel istatistiksel korelasyon tekniklerinden Spearman'ın rho (ρ) ve Kendall'ın Tau (τ) korelasyon katsayıları sıklıkla tercih edilir. Bu yöntemler, sıralama temelli çalıştıkları için normal dağılım varsayımına ihtiyaç duymaz ve verilerdeki monotonik ilişkileri ölçmek için oldukça uygundur (Puth vd., 2015).



Şekil 4.9. Değişkenler Arası Parametrik Olmayan Korelasyon Değerleri

Spearman'ın rho korelasyonu veriler arasındaki monotonik ilişkiyi ölçerken, Kendall'ın tau sıralamalar arasındaki tutarlılığı daha

derinlemesine analiz eder (Bonett & Wright, 2000). Büyük veri setlerinde Spearman'ın rho, küçük ve hassas veri setlerinde ise Kendall'ın tau daha uygun bir seçimdir. Ancak doğrusal ilişkide karşılaştırmalar açısından birlikte kullanımları oldukça mantıklı olduğu söylenebilir.

Şekil 4.9'a bakıldığında *Temperature* (*Sıcaklık*) ile diğer değişkenler arasındaki ilişkiler incelendiğinde, Spearman'ın rho korelasyon analizi sonuçlarına göre en güçlü pozitif ilişki *Radiation* (*Radyasyon*) ile gözlemlenmiştir. *Humidity* (*Nem*) ve *WindSpeed* (*Rüzgâr hızı*) ile sıcaklık arasında orta düzeyde negatif korelasyonlar tespit edilmiştir. Bu sonuçlar, nem ve rüzgâr hızı arttıkça sıcaklığın düşme eğiliminde olduğunu göstermiştir. *Pressure* (*Basınç*) ile sıcaklık arasındaki zayıf pozitif ilişki ise modelde sınırlı bir katkı sağlayabilir. *WindDirection* (*Rüzgâr yönü*) ile sıcaklık arasında ise anlamlı bir ilişki bulunmamıştır.

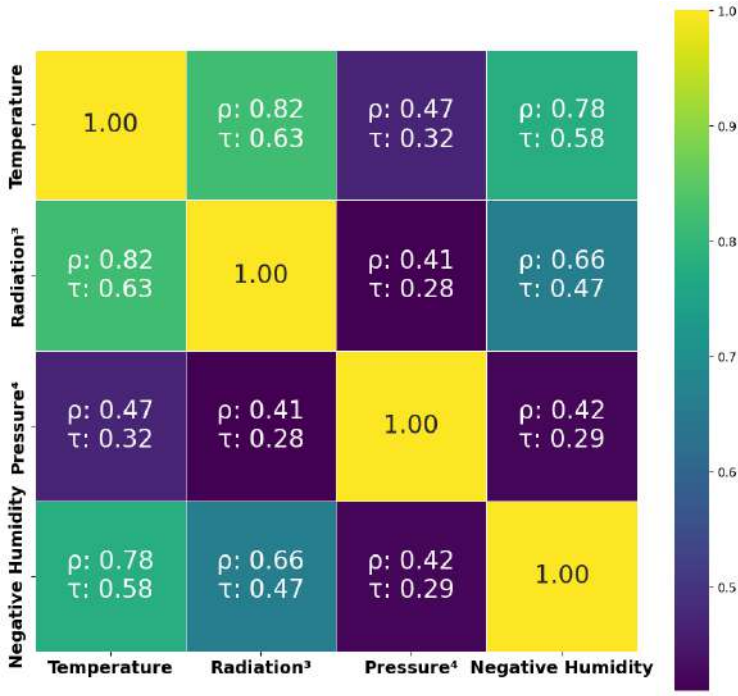
Kendall'ın tau korelasyon analizinde de, Şekil 4.9'da görüldüğü gibi, *Radiation* ile *Temperature* arasında güçlü bir pozitif ilişki bulunmakta olup bu değişken tahmin modeli için kritik öneme sahiptir. Bununla birlikte, diğer değişkenlerde Spearman'ın rho korelasyonu sonuçlarında belirgin farklılıklar gözlemlenmektedir. Örneğin, *Humidity* ile sıcaklık arasındaki negatif ilişki Kendall'ın tau korelasyonunda zayıf düzeyde belirlenirken, *WindDirection* ile *Temperature* arasında orta düzeyde negatif bir ilişki tespit edilmiştir.

Pressure ve *WindSpeed* ile sıcaklık arasındaki ilişkiler ise zayıftır. Sonuç olarak, *Radiation* dışındaki değişkenlerin modele

katılımı dikkatle ele alınmalı ve yöntemler arasındaki farklılıklar göz önünde bulundurulmalıdır.

4.2.6. Özellik Seçimi

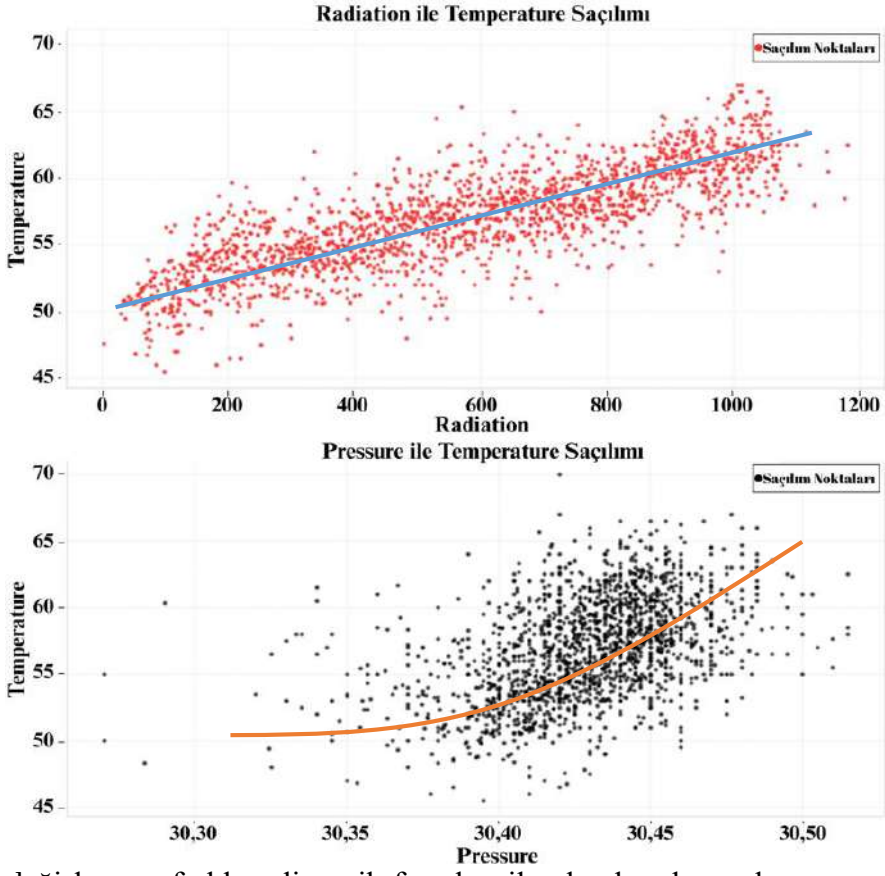
Değişken seçimi, model performansını belirleyen kritik bir süreçtir ve çok boyutlu veri yapıları içeren karmaşık bir optimizasyon problemidir. Bu çalışmada, veri dağılımına ilişkin parametrik varsayımlar gerektirmeyen Spearman'ın ρ ve Kendall'ın τ gibi sıralama tabanlı korelasyon yöntemleriyle en bilgilendirici özelliklerin seçimi sağlanmıştır (Li vd., 2017).



Şekil 4.10. Seçilen Özellikler ve Bağımlı Değişken Korelasyon Matrisi

Şekil 4.10'da yer alan Spearman'ın rho ve Kendall'ın tau korelasyonlarını incelediğimizde *Temperature* bağımlı değişkeninin en

çok doğrusal yönlü ilişkileri sırayla *Radiation*, *Humidity* ve *Pressure* ile olduğu fark edilmiştir. *Temperature* değişkeninin bu 3 bağımsız



değişken ve farklı polinomik formları ile olan korelasyonları sırasıyla 0.82, 0.78 ve 0.47 olduğu gözlemlenmiştir.

Şekil 4.11. Model Geliştirme Öncesi Olası Yüksek İlişki Göstergeleri

Şekil 4.11'de model geliştirme öncesinde yüksek korelasyon değerlerine sahip özellikler ile bağımlı değişken olan *Temperature* değişkeninin serpilme diyagramları verilmiştir. Bu veriler ışığında aralarındaki ilişkilerin yönü ve kuvveti hakkında daha fazla yorum

yapılabilir. Ek olarak en çok dikkat çeken detaylar *Temperature* bağımlı değişkeni ile *Radiation* bağımsız değişkeni arasındaki ilişkinin pozitif doğrusal yönde olduğu görülmektedir. Bu da elde edilen 0.82 korelasyon değeriyle uyum sağlamaktadır.

Temperature bağımlı değişkeni ile *Humidity* bağımsız değişkeni arasındaki negatif yönlü doğrusal ilişki olduğu Şekil 4.12'de görülmektedir. Bu öngörüyü destekler nitelikle olarak elde edilen -0.78 değeri örnek olarak verilebilir.

Temperature bağımlı değişkeni ile *Pressure* bağımsız değişkeni arasındaki ilişkinin ise doğrusal olmadığı ancak polinomik bir ilişki bulunabileceği görülmektedir. Elde edilen 0.47'lik korelasyon değeri de aralarındaki lineer ilişkinin düşük olduğunu göstermektedir.

4.3. SONUÇLAR

Çalışmanın bu aşamasında, ikinci aşamada ele alınan temel veri yapısı ve değişkenler arası ilişkiler analiz edilmiştir. Bu analizler sonucunda, *Temperature* bağımlı değişkeni üzerinde en yüksek açıklayıcılığa sahip bağımsız değişken belirlenmiştir. Ayrıca, model karmaşıklığını artırma potansiyeli taşıyan veri kümeleri modelden çıkarılmıştır. Böylece, veri önişleme süreci tamamlanmış ve elde edilen veriler model geliştirme sürecinde kullanılmıştır.

Bir Elastik-Net regresyon algoritmasında, çoklu doğrusal regresyona ek olarak L_1 yani Lasso ve L_2 yani Ridge düzeltmeleri yer alır ve bu iki düzeltme katsayıları *l1_ratio* adı verilen bir hiperparmatre ile ağırlıklandırılır. Bu ağırlıklandırmaya ek olarak Formül 4.1'de L

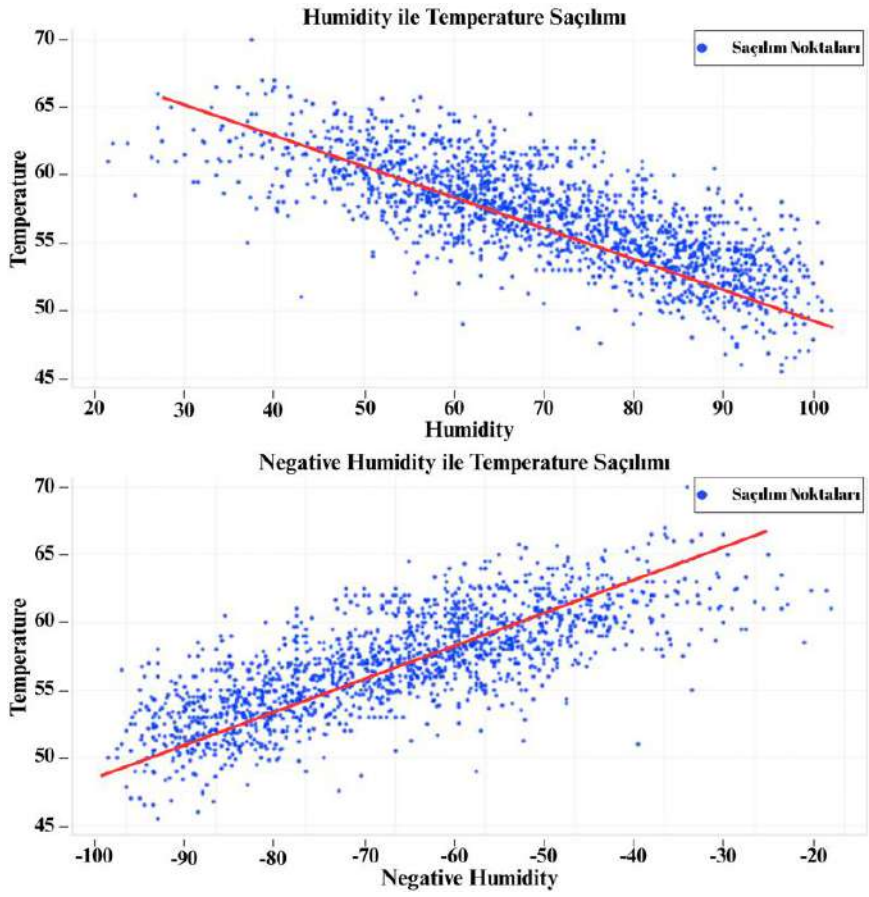
olarak ifade edilen ancak python scikit-learn kütüphanesinde a*lpha* olarak belirtilen, L_1 ve L_2 düzeltmelerinin toplamını ağırlıklandıran bir hiperparametre de mevcuttur.

Yüksek ilişkili olduğunu saptadığımız değişkenler ile kurulan modelde çoklu doğrusal regresyon bölümünde bağımlı değişken olarak *Temperature* yer alırken, bağımsız değişken olarak ise *Temperature* bağımsız değişkeniyle aralarında doğrusal ilişki olduğunu gözlemlediğimiz *Humidity* değişkeninin negatifi kullanılmıştır. Bunun temel nedeni *Humidity* ve diğer bağımsız değişkenlerin bağımlı değişken üzerindeki etkisini yorumlamada kolaylık sağlanması ve model karmaşıklığını azaltmak için gerçekleştirilmiştir.

Model ayrıca *Pressure* değişkeninin 4. kuvveti ve *Radiation* değişkeninin 3. kuvveti ile genişletilmiş olup, bu çoklu doğrusal regresyon modeline Elastik-Net regresyon parametreleri entegre edilmiştir.

Normalde bağımsız değişkenler ile bağımlı değişken arasında doğrudan klasik doğrusal bir ilişki yoktu ancak analiz ters yönlü ilişkiyi klasik doğrusal ilişkiye dönüştürecek biçimde veriler üzerinde polinomik ve negatif ile çarpma gibi işlemler gerçekleştirdik. Bunun temel nedeni bir bağımlı değişkenlerin bağımsız değişken üzerindeki saçılımları incelendiğinde ilişkilerin polinom ve parabol gibi eğriler oluşturmasıdır. Bağımlı değişken ile aralarında farklı eğimlerde eğrilerin bulunduğu bir ilişkisi bulunan değişkenlere doğrudan klasik çoklu doğrusal regresyon uygulamak hataların artarak modelin açıklanabilirliğini düşürecektir. Böylece pozitif doğrusal olmayan

ilişkileri pozitif doğrusal bir forma dönüştürmüş olduk. Şekil 4.14'deki grafikte model geliştirme öncesinde *Humidity* değişkeni ile *Temperature* değişkeni arasındaki ters yönlü ilişkiyi görmekteyiz. Bu değişkenin negatifi ile çarparak pozitif doğrusal bir ilişki formuna sokmuş olduk.



Şekil 4.12. Negatif-Pozitif Humidity Değişim Eğrisi

Korelasyon ve grafiksel dağılımlara bakıldığında *Radiation* ve *Pressure* bağımsız değişkenleri ile bağımlı değişken *Temperature*'nin arasında polinomik ilişkiler tespit edilmiştir.

$$Y_T = \beta_0 + \beta_1 X^3_{Radiaiton} + \beta_2 X^4_{Pressure} + \beta_3 X_{Negative\_Humidity} \quad (4.2)$$

4.3.1. Hiperparametre Optimizasyonu

Hedeflediğimiz Elastik-Net regresyonunun çoklu doğrusal ilişkiler modeli Formül 4.2'teki gibi modellenmiştir.

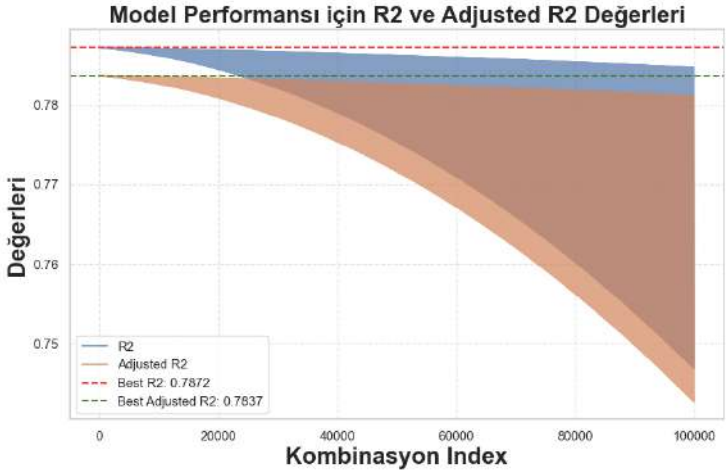
Formül 4.2'de oluşturulan tahmini modele Elastik-Net regresyonuna ait hiperparametrelerin eklenmesi fazla sayıda hiperparametre kombinasyonlarından oluşan bir seçenektir. Çünkü geliştirilen modelde *alpha* yani Elastik-Net bölümünün ağırlığı belirli aralıklar ile ağırlıklandırılabilir. Örneğin; *alpha* (L) değeri 1 değerini de alabilir 1000 değerini de alabilir. Ancak *l1_ratio* değeri yani Lasso ve Ridge ağırlıkları ise 0 ile 1 arasında değişen sayıları alırlar. Çalışmada elastik-net algoritması için uygun hiperparametre değerlerinin tespit edilebilmesi için Grid Search optimizasyonu yapılması gerekmektedir. Bunun temel nedeni kurulacak regresyon modelindeki maksimum açıklayıcılığı yakalayacak hiperparametrelerin elde edilmek istenmesidir. Böylece modelin açıklanabilirliği yani R^2 üzerinde en çok etkisi olan bu parametrelerin ağırlığı model tahmin başarısı için oldukça önemlidir.

4.3.2. Grid Search ile Hiperparamatre Optimizasyonu

Grid search yaparken *alpha* değeri için 0.001 ile 10 arasında 1000 adet rastgele değer atanırken *l1_ratio* yani Lasso mu yoksa Ridge mi daha

ağırlıklı olacak kararını veren hiperparametre için ise 0.1 ile 1 arasında rastgele 100 değer atanmıştır.

Tüm olası *alpha* ve *l1_ratio* birleşimleri ile 100.000 farklı Elastik-Net regresyon modeli oluşturulmuştur. Bu oluşan *l1_ratio* ve *alpha* modellerinden en yüksek R^2 ve R^2_α değerini veren kombinasyon model için tercih edilmiştir.



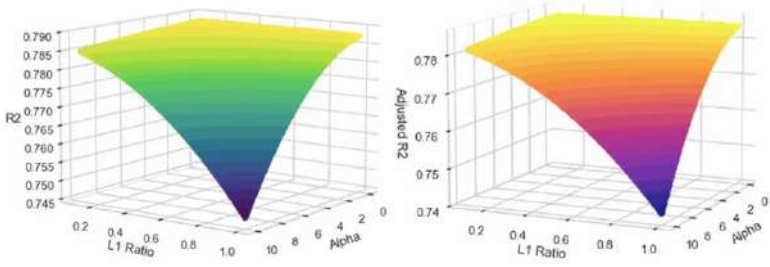
Şekil 4.13. Hiperparametre Optimizasyonu için Grid Search Çıktısı

Şekil 4.13'te açıkça görüldüğü üzere 100.000 farklı kombinasyon sonucunda kullanılan kombinasyon *l1_ratio* ve *alpha* değerine göre elde edilen en yüksek R^2 ve R^2_α değerleri sırasıyla 0,7872 ve 0,7837 olarak tespit edilmiştir.

4.3.3. R^2 - R^2_α – L (Alpha) - η (l1_ratio) İlişkisi

Model başarı ölçütleri ile model içi düzenlileştirme hiperparametreleri arasındaki ilişkilerin birbirlerine olan etkilerini gözlemlemek, modelin tahmin doğruluğunu artırmak ve bağımlı değişkenin bağımsız

değişkenler açısından açıklanabilirliğini güçlendirmek açısından büyük önem taşımaktadır.



Şekil 4.14. R^2, Ayarlanmış R^2, L (Alpha) ve η (l1_ratio) Dağılımı

Şekil 14'deki görselde R^2 ve R^2_α değerlerinin tüm birleşim *l1_ratio* ve *alpha* değerleriyle birlikte değişimini gösteren 3 boyutlu modellemesi verilmiştir. Her iki modelde de hem R^2 hem R^2_α değerinin, *l1_ratio* ve *alpha* değerlerinin arttığı durumda azaldığı gözlemlenmiştir. Bu nedenle model içerisinde düşük *alpha* ve *l1_ratio* değerleri kullanması gerekmektedir.

Determinasyon katsayının artışı *l1_ratio* ve *alpha* değerlerinin düşük olmasıyla doğru orantılı bulunmuştur. Yani *l1_ratio* ve *alpha* değeri ne kadar düşük olursa modelin başarı oranını veren R^2 artmaktadır. Elde edilen bulgulara göre *l1_ratio* ve *alpha* değerinin sırasıyla 0.001 ve 0.1 olması gerektiği sonucuna varılmıştır.

4.3.4. L (Alpha) = 0.1 & η (l1_ratio) = 0.001 Modeli:

En yüksek model başarısı için gerekli olan minimum *alpha* ve *l1_ratio* değerleri sağlandıktan sonra elde edilen değişken katsayıları Tablo 4.6'daki gibi elde edilmiştir. Tabloya göre seçilen özellikler üzerinde genel bir yorumlama yapılırsa $Radiation^3$ ve $Pressure^4$ değişkenleri

Elastik-Net modelinin düzenlileştirmeleri sonucu elimine edilerek sıfıra yakın değerlere çekilmiştir. Ancak korelasyon açısından daha düşük bir ilişkiye sahip olduğu düşünülen *Negative Humidity* değişkeni model tarafından seçilmiştir.

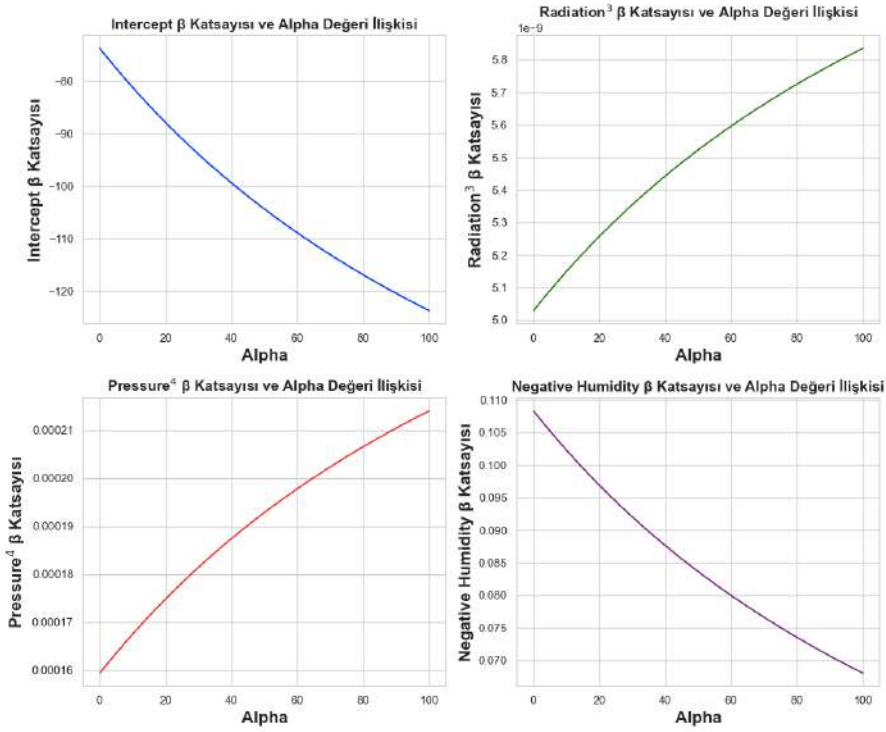
Tablo 4.6. En İyi Model Beta (β) Katsayı Değerleri

Değişken	Katsayı	Değer
Sabit Terim	β_0	-74.84
$X_{Radiation}^3$	β_1	0.000000005022555
$X_{Pressure}^4$	β_2	0.0001607242
$X_{Negative\ Humidity}$	β_3	0.10949

Elde edilen regresyon modeline ait katsayı değerleri Tablo 4.6'da verilmiştir. Bu modelin doğruluğunu değerlendirmek için hataların dağılımına ve diğer model başarı metriklerine bakmamız gerekmektedir.

4.3.5. Yüksek L (alpha) Değeri

Düşük L (*alpha*) değeri Elastik-Net'in katsayılar üzerindeki etkisini anlamak için yetersizdir. Bu nedenle daha yüksek L (*alpha*) değeri ile karşılaştırılarak test edilmelidir. L (*alpha*) = [0.1, 100] arasında rastgele 50 değer ile Elastik-Net algoritmasının *l1_ratio* ve *alpha* parametrelerinin en açıklanabilir modelin elde edilmesi için uygun değerlerinin seçilmesi için hiperparametre optimizasyonu yapıldığında, regresyon katsayılarının dağılımı aşağıdaki gibi tespit edilmiştir.



Şekil 4.15. L (Alpha) ve Beta (β) Katsayı Değerleri İlişkisi

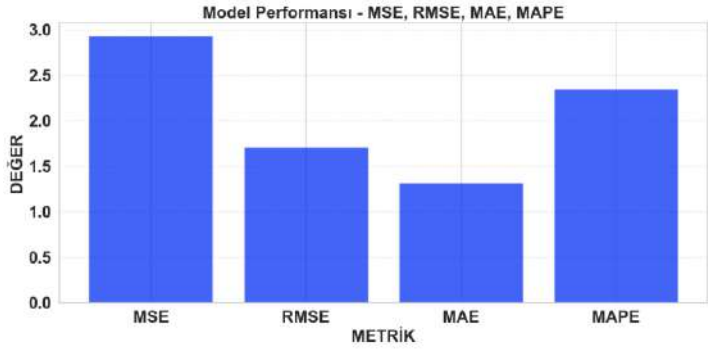
İlk modeli geliştirirken temel aldığımız 0.1 alpha değerlerinin artan bir eğilim ile optimizasyonu sonucunda bağımsız değişkenler üzerindeki etkisi Şekil 4.15'te görülmektedir. Yükselen L (*alpha*) değeri altında sabit terim ve diğer bağımsız değişkenlerin katsayılarının sıfıra yaklaşması ve küçülmesi beklenir, ancak bazı istisnai durumlarda bu durum gerçekleşmez (Zou & Hastie, 2005; Tibshirani, 2018).

Örnek olarak Şekil 4.15'te *alpha* ve katsayı ilişkileri incelendiğinde *alpha* değeri artarken düşmesi beklenen $Radiation^3$ ve $Pressure^4$ değişkenlerinin katsayılarının arttığı gözlemlenmiştir. Bunun en açıklayıcı nedenleri arasında çoklu doğrusal bağlantı problemi olabilir. Çalışmada ek olarak çoklu doğrusal bağlantı kontolü yapılmamıştır,

bunun temel nedeni kullanılan elastik-net algoritmasının çoklu doğrusal bağlantı olan verilerde aşırı öğrenmeyi önleyecek biçimde geliştirilen bir yapıya sahip olmasıdır. Sonuç olarak, *alpha* değeri arttıkça genel eğilim sıfıra yakınsama yönündedir, ancak modelin penalizasyon etkisi ve değişkenler arası ilişkilere bağlı olarak bazı katsayılar artış gösterebilir.

4.3.6. Elastik-Net Model Başarısı

Oluşturulan model için olası katsayı artışı nedenleri arasından çoklu doğrusal bağlantı olması ihtimali en yüksektir. Bunun temel nedeni bağımsız değişkenler arası yüksek korelasyonun olmasıdır. Bunlara ek olarak *Negative Humidity* katsayısının değeri ise beklendiği gibi alpha ile azalan bir ilişki göstermiştir.

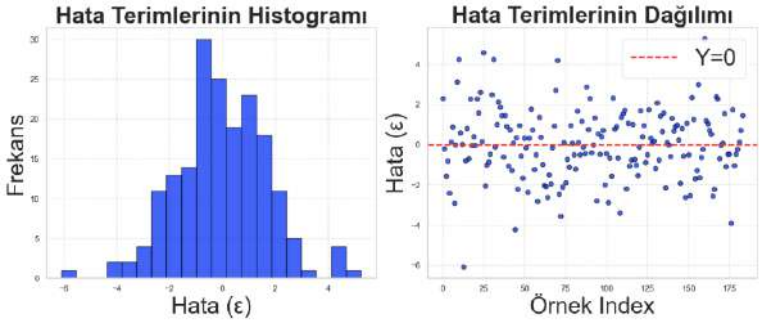


Şekil 4.16. En İyi Model Performans Metrikleri

Şekil 4.16'daki grafikte yer alan performans metrikleri, modelin tahmin doğruluğunu detaylı biçimde analiz etmemize olanak tanımaktadır. MSE değeri (~3.0), büyük hata sapmalarının model performansı üzerinde belirgin bir etkisi olduğunu göstermektedir. Bununla birlikte, RMSE (~2.5) metriği, hataların büyüklüğünü tahmin

biriminde ifade ettiği için daha yorumlanabilir bir ölçüt sağlamaktadır. MAE (~2.0) değeri, hata boyutlarının daha düşük bir seviyede olduğunu ve modelin genel anlamda istikrarlı tahminler ürettiğini göstermektedir. Benzer biçimde modelin yüzdelik mutlak hatası MAPE incelendiğinde ~%2.8 oranı, iyi bir oran göstermiştir yine de özellikle gerçek değerlerin düşük olduğu durumlarda, modelin yüzdesel hata oranının yüksek olabileceği anlamına da gelebilir.

Model başarı metriklerine ek olarak hata terimlerinin dağılımlarını da incelemek oldukça önemlidir. Çünkü hataların normal dağılıma uygun olması, modelin tahminlerinde tarafsız olduğunu ve doğrusal regresyonun geçerli olduğunu gösterir. Rastgele dağılmış hatalar, bağımsız değişkenlerin tüm bilgi içeriğini yansıttığını ve modelde eksik bir değişken olmadığı anlamına gelebilir. Eğer hatalar normal dağılmıyor veya sistematik bir desen gösteriyorsa, bu modelin yanlış kurulduğunu veya iyileştirilmesi gerektiğini ifade eder.



Şekil 4.17. En İyi Model Hata Terimleri Histogram ve Dağılımı

Şekil 4.17'deki histogram ve serpilme grafiklerine göre oldukça normalliğe yakın bir dağılım sergilediği yorumunda bulunabiliriz. Ancak bunu test etmek için Shapiro-Wilk testinden faydalanmamız gerekecektir.

Hipotezler:

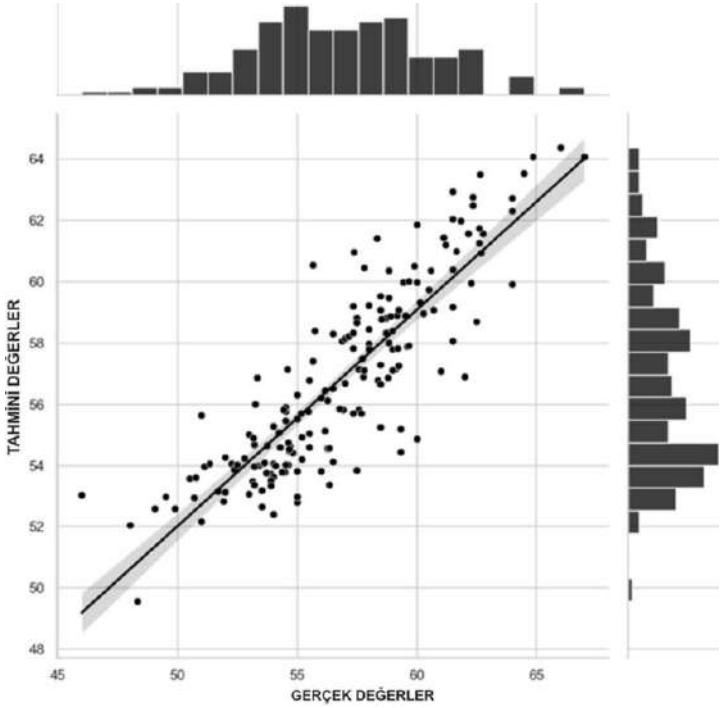
H_0: Hata terimleri normal dağılıma uygundur.

H_1: Hata terimleri normal dağılıma uygun değildir.

Elde ettiğimiz test sonucuna göre p-değeri 0.302 olarak bulunmuştur. Bu değer $\alpha = 0.05$ değerinden büyük olduğu için H_0 hipotezi reddedilemez. Dolayısıyla hata terimlerinin dağılımının normalliğe uygun olduğunu söyleyebiliriz.

Şekil 4.18. Gerçek ve Tahmin Edilen Değer Örtüşme Eğrileri

Şekil 4.18'de yer alan grafik gerçek değerler ile tahmin edilen *Temperature* değerlerinin birbirine olan yakınlığını gösteren grafiklerini içermektedir. Bu grafik çıktıları da model başarı parametrelerin uygunluk göstermekte olup değerler birbirine oldukça yakındır. Ancak bazı noktalarda model başarısı düştüğü gibi bu grafiğe de yansımıştır. Şekil 18'deki grafikte belirli aralıklarda kırmızı grafik ile siyah grafiğin birbirinden uzaklaştığı açıkça görülmektedir.



Şekil 4.19. Tahmin ve Gerçek Değerlerinin Dağılımı ve Regresyon Eğrisi

Şekil 4.19'a baktığımızda ise tahmin ve gerçek noktalarının kesişimlerinin oldukça regresyon eğrisine yaklaştığını ve çoğunun üzerinde toplandığını görebiliriz. Buna ek olarak eğrinin *x* ekseni ile

yaptığı açının 45 dereceye de oldukça yakınsadığını ve bu da gerçek değerlerin tahmin edilen değere oldukça yaklaştığını bize göstermiştir.

4.4. TARTIŞMA

Çalışmada, zaman serisi verilerinden kesitler arası yüksek korelasyon, aynı derecede durağanlık ve trend eğiliminin olmaması varsayımlarına dayalı olarak kesitsel veriler elde edilmiştir. Bu süreçte, denetimsiz kümeleme yöntemlerinden olan DBSCAN algoritması kullanılarak anomaliler tespit edilmiş ve analizden çıkartılmıştır.

Elastic-Net modelinde bağımsız değişkenler arasında yüksek korelasyon bulunması çoklu doğrusal bağlantı sorununa neden olmuştur. Modelde kullanılan düşük Elastik-Net ağırlığı L (0.001) değeri ve *l1_ratio* (0.1), Lasso'nun değişken seçici etkisini sınırlandırmış ve Ridge'in katsayı küçültme etkisini ön plana çıkarmıştır. Bu nedenle, özellikle *Radiation*^{3} ve *Pressure*^{4} gibi yüksek dereceli polinomik terimlerin katsayıları aşırı küçülmüştür. *Negative Humidity*'nin daha yüksek bir katsayı alması (0.109), hedef değişken *Temperature* ile daha güçlü bir ilişkiye sahip olduğunu ve diğer değişkenlerin elenerek bu değişken ile yeni bir model kurulabileceğini göstermektedir. Bu yapı, düzenlileştirmenin etkisiyle aşırı öğrenmeyi engellemiş ve modeli genelleştirmeye uygun hale getirmiştir.

Formül 4.2 ile tanımladığımız model, Lasso ve Ridge regresyon teknikleri, anormal değerlerin çıkarılmasından sonra güçlü ve dengeli bir model haline gelmiştir. Lasso regresyon, yalnızca en önemli özellikleri seçerken, Ridge regresyon çoklu doğrusal bağlantı

sorunlarını çözmüştür. Modelin genel başarısı, doğrusal bağımsız değişkenlerle kurulan bir regresyon modelinde düzenlileştirme ne kadar etkili olduğunu göstermektedir.

Elastik-Net katsayısı L (*alpha*) ile bağımsız değişkenlerin katsayıları arasındaki ilişki beklenenin aksine çoklu doğrusal bağlantıdan dolayı artan eğilimde gözlemlenmiştir. Ancak bunların yanı sıra yaklaşık olarak %79 R^2 hesaplanırken %78 ayarlanmış R^2 hesaplanmıştır. Modelin iyi bir başarı gösterdiği yorumu yapılabilir.

Elastik-Net regresyon modelinde hata terimlerinin dağılımı da normal dağılıma sahip olduğu testler ile ispatlanmış olup bu da model çıktılarının yorumlarındaki varsayımlarımızı doğrular nitelik taşımaktadır.

Hiperparametre optimizasyonu ise yüksek açıklanabilirlik odaklı olduğu için R^2'nin olduğu optimizasyon işlemi gerçekleştirilmiştir. Böylece varsayımlarımız altında durağan ve trendsiz verilerde ortalama bazlı yaklaşımların uygulanabilirliğini Formül 4.2 ile gösterilen regresyon modelinde genelleştirilebilir bir model üreterek test etmiş olduk.

KAYNAKÇA

Bhattacharjee, M., Banerjee, M., & Michailidis, G. (2019). Change Point Estimation in Panel Data with Temporal and Cross-sectional Dependence. arXiv. https://doi.org/10.48550/arXiv.1904.11101

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall, and Spearman correlations. Psychometrika, 65(1), 23-28. https://doi.org/10.1007/BF02294183

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). Wiley. https://doi.org/10.1002/9781118619193

Dette, H., & Heinrichs, F. (2020). A Distribution Free Test for Changes in the Trend Function of Locally Stationary Processes. arXiv. https://doi.org/10.48550/arXiv.2005.11132

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366), 427-431. https://doi.org/10.2307/2286348

Djenouri, Y., Belhadi, A., & Lin, J. C.-W. (2022). Recurrent neural network with density-based clustering for group pattern detection in energy systems. Sustainable Energy Technologies and Assessments, 53, Article 102308. https://doi.org/10.1016/j.seta.2022.102308

Dumler, D., Faatz, S., Friedrich, M. & Döpper, F. (2023). Automatic time series segmentation and clustering for process monitoring in series production. Procedia CIRP, 120, 336-341. https://doi.org/10.1016/j.procir.2023.06.103

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2016). The Gaussian Graphical Model in Cross-sectional and Time-series Data. arXiv. https://doi.org/10.48550/arXiv.1609.04156

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), 226-231. https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. International Journal of Endocrinology and Metabolism, 10(2), 486-489. https://doi.org/10.5812/ijem.3505

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts. https://otexts.com/fpp3/

Jaén-Vargas M, Reyes Leiva KM, Fernandes F, Barroso Gonçalves S, Tavares Silva M, Lopes DS, Serrano Olmedo JJ. 2022. Effects of sliding window variation in

the performance of acceleration-based human activity recognition using deep learning models. PeerJ Computer Science 8:e1052 https://doi.org/10.7717/peerj-cs.1052

Kirchgässner, G., Wolters, J., & Hassler, U. (2013). Introduction to modern time series analysis (2nd ed.). Springer. https://doi.org/10.1007/978-3-642-33436-8

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics, 54(1-3), 159-178. https://doi.org/10.1016/0304-4076(92)90104-Y

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM Computing Surveys, 50(6), 1-45. https://doi.org/10.1145/3136625

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. The American Statistician, 54(3), 217-224. https://doi.org/10.1080/00031305.2000.10474549

Lütkepohl, H. (2005). New introduction to multiple time series analysis. Springer. https://doi.org/10.1007/978-3-540-27752-1

Moy AJ, Cato KD, Withall J, Kim EY, Tatonetti N, Rossetti SC. Using Time Series Clustering to Segment and Infer Emergency Department Nursing Shifts from Electronic Health Record Log Files. AMIA Annu Symp Proc. 2023 Apr 29;2022:805-814. PMID: 37128367; PMCID: PMC10148355.

Phillips, P. C. B., & Perron, P. (1988). Testing for a unit root in time series regression. Biometrika, 75(2), 335-346. https://doi.org/10.1093/biomet/75.2.335

Pinaria, R. (2021). Güneş Radyasyonu Tahmini [Veri seti]. Kaggle. 17 Aralık 2024 tarihinde erişildi: https://www.kaggle.com/code/rickypinaria/solar-radiation-prediction-using-linear-regression/input

Puth, M.-T., Neuhäuser, M., & Ruxton, G. D. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. Animal Behaviour, 102, 77-84. https://doi.org/10.1016/j.anbehav.2015.01.010

Tibshirani, R. (2011). Regression shrinkage and selection via the Lasso: A retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3), 273-282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

Tibshirani, R. (2018). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Yamashita Rios de Sousa, A. M., Takayasu, H., & Takayasu, M. (2020). Segmentation of time series in up- and down-trends using the epsilon-tau procedure with application to USD/JPY foreign exchange market data. PLOS ONE, 15(9), e0239494. https://doi.org/10.1371/journal.pone.0239494

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

ISBN: 978-625-5923-59-2

DOI: https://doi.org/10.5281/zenodo.15599950