Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1581168

Open Access Journal
e-ISSN: 2619 – 8991

Araştırma Makalesi (Research Article)
Cilt 8 - Sayı 1: 161-171 / Ocak 2025
(Volume 8 - Issue 1: 161-171 / January 2025)

ÜLKELERİN İNSANİ GELİŞMİŞLİK ÖLÇÜSÜNE GÖRE SEVİYELERİNİN BELİRLENMESİNDE KÜMELEME ALGORİTMALARININ KULLANILMASINA İLİŞKİN BİR UYGULAMA

Latife Sinem SARUL[1]*

[1]*İstanbul University, School of Business, Department of Quantitative Techniques, 34322, İstanbul, Türkiye*

Özet: Temel yaşam standartlarına erişim, eğitim ve sağlıklı yaşama süresi gibi üç temel gösterge dikkate alınarak oluşturulan İnsani Gelişme İndeksi (Human Development Index) ilk olarak 1990 yılında Birleşmiş Milletler Kalkınma Programı tarafından ortaya konulmuştur. Belirtilen temel göstergeler dikkate alınarak tüm dünya ülkelerinin İnsani Gelişme İndeksi hesaplanmakta, oluşturulmuş olduğu ilk yıldan itibaren her yıl düzenli olarak Birleşmiş Milletler tarafından kamuoyu bilgisine sunulmaktadır. Bu çalışmada veri madenciliği kapsamında kümeleme analizi teknikleri detaylı olarak incelenmiş ve bu teknikler kullanılarak İnsani Gelişme İndeksine göre ülkelerin gruplaması yapılmıştır. Elde edilen sonuçlar Birleşmiş Milletler Kalkınma Programı tarafından açıklanan listeye göre karşılaştırılarak yapılan analizlerin bu alanda uygulanabilirliği tartışılmıştır.

Anahtar kelimeler: Veri madenciliği, Kümeleme algoritmaları, İnsani Gelişme İndeksi

An Application on the Use of Clustering Algorithms in Determining the Levels of Countries According to their Human Development Index

Abstract: The Human Development Index, which is created by considering three basic indicators such as access to basic living standards, education and healthy life expectancy, was first created by the United Nations Development Program in 1990. The Human Development Index of all countries in the world is calculated by taking into account the specified basic indicators and is regularly presented to the public by the United Nations every year since its creation. In this study, cluster analysis techniques within the scope of data mining are examined in detail and countries are grouped according to the Human Development Index using these techniques. The results obtained are compared according to the list announced by the United Nations Development Program and the applicability of the analyses in this field was discussed.

Keywords: Data mining, Clustering algorithms, Human Development Index

Sorumlu yazar (Corresponding author): *İstanbul University, School of Business, Department of Quantitative Techniques, 34322, İstanbul, Türkiye*
E mail: lasinem@istanbul.edu.tr (L. S. SARUL)
Latife Sinem SARUL https://orcid.org/0000-0001-7013-3755

Gönderi: 07 Kasım 2024 **Received:** November 07, 2024
Kabul: 10 Aralık 2024 **Accepted:** December 10, 2024
Yayınlanma: 15 Ocak 2025 **Published:** January 15, 2025

Cite as: Sarul LS. 2025. An application on the use of clustering algorithms in determining the levels of countries according to their Human Development Index. BSJ Eng Sci, 8(1): 161-171.

1. Giriş

İnsani Gelişme İndeksi, Sağlıklı Yaşam İndeksi, Eğitime Ulaşılabilirlik İndeksi, Temel Yaşam Düzeyine Erişebilirlik İndeksinden oluşan üç temel faktörün normalleştirilmiş değerlerinin geometrik ortalaması alınarak elde edilen bir ölçüdür. İnsani Gelişme İndeksini oluşturan üç temel faktörden Sağlıklı Yaşam İndeksi, sağlıklı doğan bir bireyin beklenen yaşam süresi ile ölçülmektedir. İkinci temel faktör olan Eğitim İndeksi, 25 yaş ve üzeri bireylerin ortalama eğitim süresi ve okula başlama yaşındaki çocuklar için beklenen eğitim alma süresi ile ölçülmektedir. Son olarak Temel Düzeyde Yaşam Standartlarına Erişme İndeksi ise kişi başına düşen gayri safi milli gelir değeri ile ölçülmektedir (HDI Reports, 2024). Birleşmiş Milletler Kalkınma Programı tarafından açıklanan rapora göre ülkeler insani gelişmişlik düzeylerine göre dört gruba ayrılmaktadır. Birleşmiş Milletler Kalkınma Programı tarafından açıklanan 2023-2024 yılı raporuna göre en yüksek İnsani Gelişmişlik düzeyine sahip ilk beş ülke sırasıyla İsviçre, Norveç, İzlanda, Hong Kong- Çin (SAR), Danimarka olarak açıklanmıştır. Açıklanan 191 ülke içerisinde Türkiye 45. sırada yer almaktadır. Bu noktada ülkelerin gelişmişlik düzeylerine göre gruplanmasında sınır değerlerinin belirlenmesi ile ilgili literatürde yapılan pek çok çalışma bulunmaktadır. Bu alanda yapılan en son çalışmalardan biri olan Wang vd. (2023) tarafından yapılmış çalışmada sınır değerlerinin belirlenmesinde denetimsiz makine öğrenmesi tekniklerinden K-Means ve K-Medoids algoritmaları kullanılmıştır. Bu çalışmada ülkelerin insani gelişmişlik düzeylerine göre düşük, orta ve yüksek olarak üç düzeyde gruplanması önerilmiştir.

Bu çalışmada ise hiyerarşik olmayan yöntemlerden K-Means Algoritması ve K-Medoids Algoritması, hiyerarşik

yöntemlerden En Yakın Komşuluk Algoritması (Single Linkage), En Uzak Komşu Algoritması (Complete Linkage), Ortalama Bağlantı Yöntemi(Average Linkage), Ward Yöntemi ve yoğunluk bazlı yöntemlerden DBSCAN (Density based Spatial Clustering of Applications with Noise) Algoritmasının kullanıldığı 39 farklı model oluşturulmuştur. Elde edilen sonuçlar, Birleşmiş Milletler Kalkınma Programı tarafından yapılmış gerçek sınıflamalarla karşılaştırılmıştır. Hiyerarşik ve Hiyerarşik Olmayan Yöntemlerin birarada incelenmiş olması bakımından bu çalışmanın literatüre önemli katkı sağlayacağı düşünülmektedir.

2. Literatür

Literatürde Kümeleme Algoritmaları ve İnsani Gelişim İndeksi üzerine yapılmış pek çok çalışma yer almaktadır. Bu kısımda kümeleme algoritmaları ile yapılan çalışmalar ve ayrıca kümeleme algoritmaları ile İnsani Gelişim İndeksinin birlikte incelendiği çalışmalardan bazılarına yer verilmiştir. Bu anlamda Ezugwu vd. (2022) veri madenciliği çerçevesinde kümeleme algoritmaları kullanımına ilişkin kapsamlı bir çalışma yapmışlardır. Bu çalışmada pek çok farklı alanda kullanılmakta olan kümeleme algoritmaları ve ileri düzey teknikler açıklanmış ayrıca detaylı bir literatür taraması yapılmıştır. Buna göre Ezugwu vd. (2022) küme sayısının önceden belirlenmesinin gerekliliğinin kümeleme algoritmalarının uygulanmasında önemli bir sorun olduğunu tespit etmişler bu nedenle de küme sayısının önceden belirlenmesini gerektirmeyen algoritmaların diğer algoritmalara göre daha çok tercih edildiğini ifade etmişlerdir. Literatürde yapılan çalışmalara bakıldığında diğer algoritmalara kıyasla K-Means algoritmasının önemli ölçüde fazla kullanılmakta olduğu yapılan grafik analizlerle de gösterilmiştir.

Veri Madenciliği çerçevesinde Kümeleme Algoritmalarının kullanımına ilişkin bir diğer çalışma Neha ve Vidyavathi (2015) tarafından yapılmıştır. Bu çalışmada kümeleme algoritmaları, Hiyerarşik, Hiyerarşik Olmayan, Yoğunluk Bazlı ve Izgara Bazlı Yöntemler olarak dört başlık altında incelenmiştir. Bu çalışmada da K-Means algoritmasının daha iyi sonuçlar üretmesi ve daha hızlı uygulanabilir olması nedeniyle diğer algoritmalara kıyasla daha fazla kullanılmakta olduğu ifade edilmiştir. Ayrıca kümeleme algoritmalarının suç tespiti, eğitim gibi alanlarda kullanımının önemine dikkat çekmişlerdir.

Shah ve Nair (2015) yine veri madenciliği çerçevesinde kümeleme algoritmalarını inceledikleri bir çalışma yapmışlardır. Bu çalışmada Kümeleme Algoritmaları Hiyerarşik, Hiyerarşik Olmayan, Yoğunluk Bazlı ve Izgara Bazlı olarak dört başlıkta incelenmiştir. Hiyerarşik kümeleme teknikleri başlığı altında K-Means ve K-Medoids (PAM ve CLARA) Algoritmaları, Hiyerarşik Teknikler başlığı altında Birleştirici (BIRCH ve CHAMELEON) ve Ayrıştırıcı Yöntemler incelenmiştir. Yoğunluk Bazlı Yöntemler başlığı altında literatürde en çok kullanılan algoritmalar DBSCAN ve DENCLUE Algoritmaları, Izgara Bazlı Yöntemler başlığı altında ise STING ve CLIQUE Algoritmalarından bahsedilmiş bu yöntemlerin birbirlerine göre üstün ve üstün olmayan yönleri vurgulanmıştır. Buna göre K-Means Algoritmasının büyük veri setlerinde kullanıma uygun olduğu, gürültü ve aykırı değerlere karşı duyarlı olduğu belirtilmiştir. Bununla birlikte hiyerarşik tekniklerden biri olan BIRCH Algoritmasının da büyük veri setleri için en uygun seçimlerden biri olabileceği belirtilmiştir. DBSCAN ve DENCLUE Algoritmalarının ise büyük veri setleri için çok uygun olmamakla birlikte veri seti içindeki kümelenmelerin konveks olmaması halinde iyi sonuçlar verdiğini belirtmişlerdir.

Kameshwaran ve Malarvizhi (2014) çalışmalarında veri madenciliği çerçevesinde kümeleme tekniklerini incelemişlerdir. Bu çalışmada kümeleme tekniklerinin Dağılım Modelleri, Merkez Modelleri, Yoğunluk Modelleri, Alt Uzay Modelleri, Grup Modelleri ve Grafik Bazlı Modeller olarak sınıflanabileceği gibi bir başka yaklaşım olarak da Hiyerarşik, Hiyerarşik Olmayan, Izgara Bazlı, Yoğunluk Bazlı ve Grafik Bazlı olarak da gruplanabileceğini belirtmişler ve bu yöntemlere ilişkin açıklamalara yer vermişlerdir.

Rocha vd. (2021) çalışmasında Kümeleme ve K-Means Algoritmalarını kullanarak İnsani Gelişim İndeksine göre Peru'da bir uygulama yapmışlardır. Buna göre, yaşam beklentisi, eğitime erişebilirlik ve gelir düzeyi açısından insani gelişmişliklerine göre Peru dört bölgeye ayrılmıştır. Optimal bölge sayısının belirlenmesinde Elbow Yöntemi ve Temel Bileşenler Analizi kullanılmıştır. Çalışmada, K-Means ve diğer kümeleme algoritmalarının tutarlı sonuç verdiği sonucuna varılmıştır.

Nurhasanah vd. (2021) çalışmasında K-Means Algoritmasını kullanarak İnsani Gelişim İndeksine bağlı olarak Endonezya'nın şehirlerini düşük, orta, yüksek ve çok yüksek gelişim düzeyinde dört bölgeye ayırmışlardır. K-Means Algoritması çok değişkenli istatistik tekniklerden biri olarak ele alınmış, değişkenler arasında çoklu doğrusal bağlantı olup olmaması durumu VIF (Variance Inflation Factor) değerlerine bakılarak incelenmiştir.

Muttaqin (2022) çalışmasında Endonezya'nın Sumatra adasının ilçe ve şehirlerini çok değişkenli istatistik analizi tekniklerinden biri olan kümeleme analizini kullanarak İnsani Gelişmişlik İndeksine göre yüksek, orta ve düşük olarak üç kategoriye ayırmışlardır. Bu çalışmada oluşturulan üç kümenin birbirlerinden anlamlı bir şekilde farklı olup olmadığını test etmek için Anova Analizi yapılmıştır.

3. Kümeleme Algoritmaları

Kümeleme Algoritmaları Çok Değişkenli İstatistik Teknikler başlığı altında ve Veri Madenciliği Teknikleri başlığı altında incelenen pek çok çalışmada kullanılmış ve kullanılmakta olan oldukça temel bir konudur. Bu çalışmada Veri Madenciliği Teknikleri başlığı altında kümeleme algoritmalarının incelenmesi hedeflenmiştir. Kümeleme algoritmaları temel olarak Hiyerarşik (Hierarchical Methods) ve Hiyerarşik Olmayan

Yöntemler (Partitioning Methods) olarak iki ana başlık altında incelenebilmektedir (Pujari, 2001; Kurnaz vd, 2022). Bununla birlikte daha ileri düzey gelişmiş tekniklerle birlikte, Hiyerarşik Olmayan Yöntemler/ Bölümleme Yöntemleri (Partitioning Methods), Hiyerarşik Yöntemler (Hierarchical Methods), Yoğunluk Bazlı Yöntemler (Density Based Methods) ve Izgara Bazlı Yöntemler (Grid Based Methods) olarak dört başlıkta da incelenebilmektedir (Han vd., 2023). Bu ayrımı ileri düzey tekniklerin eklenmesiyle artırmak mümkündür (Ezugwu vd., 2022).

3.1. Hiyerarşik Olmayan Yöntemler (Partitioning Techniques)

Bölümlemeli Yöntemler veya diğer adıyla Hiyerarşik Olmayan Yöntemler öncelikle "k" parametresinin araştırmacı tarafından önceden belirlenmesini gerektirmektedir. Bu yöntemler içerisinde literatürde en çok karşımıza çıkan algoritma K-Means Algoritması olmakla birlikte K-Medoids ve K-Modes Algoritmaları da bu bölüm altında incelenmektedir (Pujari vd., 2001; Han, 2023).

3.1.1. K-Means Algoritması

K-Means Algoritması, MacQueen tarafından 1967 yılında geliştirilmiş olup daha sonra farklı algoritmalar da üretilmiştir (MacQueen, 1967; Morissette and Chartier 2013). K-means Algoritması kavramsal olarak sezgisel ve uygulanabilirliği kolay bir yöntem olarak literatürde çok sayıda uygulamada karşımıza çıkmaktadır. Ezugwu vd. (2022)'de yapılan çalışmaya göre, K-Means Algoritması diğer kümeleme algoritmalarına kıyasla literatürde en çok kullanılan algoritma olarak tespit edilmiştir.

Merkez bazlı algoritmalar olarak da ifade edilen K-Means Algoritması, veri setini oluşturan gözlemleri, birbirleriyle ortak bir eleman içermeyen k sayıda kümeye ayırmaktadır. Burada "k" parametresinin belirlenmesi veri setinin özelliklerine göre araştırmacının öngörüsüne dayanmaktadır. "k" parametresinin belirlenmesiyle ilgili olarak Calinski-Harabasz index, Gap Statistic gibi yöntemler de literatürde kullanılan bazı yöntemlerdir (Han, 2023). K-Means Algoritması sadece nümerik değerler olduğunda ve veri seti içindeki kümelerin konveks yapıda olması durumunda kullanılabilmektedir (Kameshwaran ve Malarvizhi K, 2014). Bu algoritmada, her bir kümenin merkezi, kümeyi oluşturan gözlemlerin aritmetik ortalaması alınarak hesaplanmaktadır. Her bir kümeyi temsil eden küme merkezi ile gözlemler arasındaki uzaklıklar genellikle Öklid Algoritması ile hesaplanmaktadır. Ancak uzaklık ölçüsü olarak Öklid Algoritmasının dışında, Manhattan Mesafesi, Minkowski Mesafesi gibi diğer uzaklık ölçüleri de kullanılabilmektedir (Han, 2023). K-Means Algoritmasında, her bir kümeyi oluşturan gözlemlerin ait oldukları küme merkezlerine olan mesafelerinin toplamını bir diğer ifadeyle küme içi varyanslarını en küçük yapacak gözlemlerin biraraya getirilmesi hedeflenmektedir.

K-Means Algoritması İterasyonlar

1. Küme sayısının (k) belirlenmesi
2. Verisetindeki tüm gözlemlerin her birinin kendisine en yakın uzaklıktaki kümeye atanması
3. Küme merkezlerinin hesaplanması

$$M_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik}$$

4. Her bir gözlemin küme merkezlerine uzaklıklarının hesaplanması
5. Küme içi varyanslarının hesaplanması

$$e_i^2 = \sum_{i=1}^{n_k}(x_{ik} - M_k)^2$$

Bu iterasyonlar ikinci aşamadan başlayarak küme içi varyasyon en küçük kalacak biçimde gözlemlerin yer değiştirmesi ile devam etmektedir. Küme içi varyasyonu en küçük olduğunda iterasyon sonlandırılmaktadır. Burada k, küme sayısını, n, veri seti içindeki gözlem sayısını, x_{ik}, k. kümeye ait i. gözlemi, M_k, k. kümenin merkezini, e_i^2, hata karelerini ifade etmektedir.

Gözlem sayısının büyüklüğü, küme sayısının fazla olması gibi değişkenliklere bağlı olarak algoritmanın sonuca ulaşma süresi de değişmektedir. Bu anlamda büyük veri setlerine de uygulanabilir bir yöntemdir. Ancak veri seti içerisinde aykırı gözlemlerin olması küme merkezlerini etkileyeceğinden iterasyon süresi de buna bağlı olarak değişmektedir (Han, 2023). Bununla birlikte, veri setini oluşturan gözlemlerin konveks olarak gruplanabildiği ve lineer olarak ayrılabildiği durumlarda oldukça kullanışlı bir yöntem olmakla birlikte, belirtilen durumların dışında kullanıldığında etkinliğini yitirmektedir (Han, 2023).

3.1.2. K-Medoids Algoritması

K-Medoids Algoritması veya bir diğer adı ile PAM (Partitioning Around Medoids) Algoritması, 1990 yılında Kaufman ve Rousseeuw (1990) tarafından geliştirilmiştir. K-Medoids Algoritması merkez bazlı yöntemlerden biri olup aslında, K-Means Algoritmasının aykırı değerler ve gürültü içeren veri setleri için daha iyileştirilmiş bir hali olarak literatürde kullanılmakta olan bir yöntemdir. K-Medoids Algoritmasının K-Means Algoritmasından farkı oluşturulan kümelerin merkezi olarak, kümeyi oluşturan gözlemlerin aritmetik ortalamasından oluşan bir merkez seçmek yerine, kümenin içinden bir gözlem noktasının merkez olarak belirlenmesidir. Algoritma, seçilen merkez nokta ile diğer gözlemler arasındaki mesafelerin toplamı minimum olana kadar devam etmektedir (Han, 2023).

PAM Algoritması İterasyonlar

1. Küme sayısının (k) belirlenmesi
2. Verisetindeki tüm gözlemlerin her birinin kendisine en yakın uzaklıktaki kümeye atanması
3. Oluşturulan her bir küme için küme merkezi olabilecek bir gözlemin belirlenmesi
4. Her bir gözlemin küme merkezlerine uzaklıklarının hesaplanması
5. Küme içi varyanslarının hesaplanması

$$e_i^2 = \sum_{i=1}^{n_k}(x_{ik} - M_k)^2$$

Bu iterasyonlar üçüncü aşamadan başlayarak küme içi varyasyon en küçük kalacak biçimde gözlemlerin yer değiştirmesi ile devam etmektedir. Küme içi varyasyonu en küçük olduğunda iterasyon sonlandırılmaktadır. PAM Algoritması çok büyük veri setlerinde iterasyon sayısı artacağından çok elverişli bir yöntem olmayabilir. Bu durumda bu soruna çözüm olarak geliştirilen CLARA Algoritması kullanılabilir (Kassambara, 2017).

K-Medoids Algoritması başlığı altında incelenen bir diğer Algoritma ise PAM Algoritmasının büyük veri setleri için geliştirilmiş hali olan CLARA (Clustering Large Applications) algoritmasıdır (Kassambara, 2017).

3.1.3. CLARA Algoritması (Clustering Large Applications)

CLARA Algoritması, Kaufman ve Rousseeuw tarafından 1990 yılında geliştirilen PAM algoritmasının büyük veri setleri için geliştirilmiş bir halidir (Kaufman ve Rousseeuw, 1990). CLARA Algoritmasında PAM Algoritmasından farklı olarak bütün veri seti içinde k farklı küme oluşturmak yerine veri seti tesadüfi olarak belirlenmiş belirli sayıdaki gözlem kümelerine ayrılarak küme merkezleri belirlenir.

CLARA Algoritması İterasyonlar

1. Veri setini oluşturan gözlemler belirli sayıdaki birimden oluşan alt kümelere tesadüfi olarak ayrılır.
2. Bundan sonraki aşamada oluşturulan her bir alt kümeye yukarıda ifade edilen PAM Algoritmasının aşamaları uygulanır.
3. Her bir küme içi hata kareleri toplamı(varyansların)nın hesaplanması

$$e_i^2 = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2$$

3.2. Hiyerarşik Yöntemler (Hierarchical Techniques)

Hiyerarşik yöntemler genel olarak birleştirici (aşağıdan yukarıya) ve ayrıştırıcı/bölücü (yukarıdan aşağıya) yöntemler olarak iki başlık altında incelenmektedir. Birleştirici yöntemler veri setindeki tek bir gözlem biriminden başlayarak benzer özellikteki ikinci gözlemin eklenmesiyle aşağıdan yukarıya doğru ilerleyen bir yapıdır. Ayrıştırıcı/ Bölücü Yöntemler ise bütün veri setini oluşturan gözlemleri tek bir veri seti olarak kabul ederek başlayan ve benzer özellikteki gözlemleri bir araya getirmek için en uygun küme sayısına göre gözlemleri gruplara ayıran bir yapıdır (Berkhin, 2002). Hiyerarşik Yöntemlerde bir gözlem herhangi bir kümeye atandıktan sonra başka bir kümeye geçirilmesi durumu söz konusu olmamaktadır (Han, 2023). Bununla birlikte hiyerarşik olmayan yöntemlerden farklı olarak küme sayısının önceden araştırmacı tarafından belirlenmesi gerekmemektedir. Bu durum araştırmacılar açısından bir kolaylık gibi görünse de verisetinin hiyerarşik bir yapıya sahip olmaması durumunda hiyerarşik yöntemlerin seçilmesi konusunda dikkatli olunması gerekmektedir (Everitt, 2011). Hiyerarşik Yöntemlerde küme sayısının belirlenmesinde Dendrogram adı verilen ağaç diyagramından yararlanılmaktadır (James vd., 2013).

3.2.1. Birleştirici (Agglomerative) Yöntemler

Birleştirici Yöntemler, Hiyerarşik Yöntemler içerisinde literatürde en çok kullanılan yöntemler olarak görülmektedir. En Yakın Komşuluk Algoritması (Single Linkage), En Uzak Komşu Algoritması (Complete Linkage), Ortalama Bağlantı Yöntemi (Group Average Linkage), Ward Yöntemi (Ward's Method) ise en fazla bilinen birleştirici algoritmalardır. Bu algoritmaların dışında literatürde var olan diğer Birleştirici Algoritmalar ise, Merkezi Bağlantı Yöntemi (Centroid Linkage), Ağırlıklı Ortalama Bağlantı Yöntemi (Weighted Average Linkage), Median Bağlantı Yöntemi (Median Linkage) olarak sıralanabilir (Everit, 2011). Hiyerarşik Yöntemler mesafeye dayalı yöntemler olup gözlemler arasındaki mesafenin hesaplanmasında sıklıkla Öklid Mesafesi kullanılmaktadir. Bunun dışında Manhattan Mesafesi, Minkowski Mesafesi, City Block Mesafesi gibi diğer mesafe ölçüleri de kullanılabilmektedir (Everit, 2011; Bramer, 2016).

En Yakın Komşuluk Algoritması (Single Linkage)

En Yakın Komşuluk Algoritmasında temel yaklaşım gözlemler arasındaki mesafenin en küçük olması şeklindedir. Başlangıçta her bir gözlem tek başına bir küme olarak kabul edilir. İkinci aşamada ise bütün gözlemler arasındaki mesafeler yukarıda bahsedilen mesafe ölçüleri kullanılarak hesaplanır. Aralarındaki mesafenin en küçük olduğu iki gözlem birleştirilerek yeni bir küme olarak kabul edilir bundan sonraki aşamada yine var olan kümeler arasındaki mesafeler hesaplanarak kümeler birleştirilir. Bu işlem bütün gözlemler tek bir kümede birleşene kadar devam ettirilir (Bramer, 2016).

En Yakın Komşuluk Algoritması İterasyonlar

1. Öncelikle gözlemler arasındaki uzaklıklar hesaplanır.
2. Birbirine en benzer özellikteki gözlemleri birleştirebilmek için $min\, d(i,j)$ belirlenir.
3. Minimum uzaklığa sahip gözlemler birleştirilerek yeni bir küme elde edilir. Bu duruma göre uzaklıklar yeniden hesaplanır.
4. Birden fazla gözlem değerine sahip iki küme söz konusu olduğunda, iki farklı küme içinden alınan gözlemler arasında birbirine en yakın olanların uzaklığı iki kümenin birbirine olan uzaklığı olarak kabul edilir.

$$d_{A\to B} = \min_{\substack{\forall i\in A \\ j\in B}} d_{i,j}$$

5. Bütün gözlem birimleri tek bir kümede birleşene kadar iterasyonlar devam ettirilir.

Burada $d(i,j)$; i. ve j. gözlemler arasındaki mesafeyi ifade etmektedir (Özkan, 2013).

En Uzak Komşu Algoritması (Complete Linkage)

En Uzak Komşu Algoritması, En Yakın Komşu Algoritmasından farklı olarak küme içindeki farklılığın maksimum olması esasına dayanır (James, 2013). Burada da algoritmanın işleyişi En Yakın Komşu Algoritmasında ifade edildiği gibi birbirlerine en yakın olan gözlemlerin birleştirilmesi şeklindedir (Bramer, 2016).

En Uzak Komşuluk Algoritması İterasyonlar

1. Öncelikle gözlemler arasındaki uzaklıklar belirlenir.
2. Daha sonra uzaklıklar dikkate alınarak $min\, d(i,j)$ belirlenir. (En Yakın Komşu Algoritmasında olduğu gibi)
3. Minimum uzaklığa sahip gözlemler birleştirilerek yeni bir küme elde edilir. Bu duruma göre uzaklıklar yeniden hesaplanır.
4. Birden fazla gözlem değerine sahip iki küme söz konusu olduğunda, iki farklı küme içinden alınan gözlemler arasında birbirine en uzak gözlemler arasındaki mesafe iki kümenin birbirine olan uzaklığı olarak kabul edilir.
   $d_{A\to B} = \max_{\substack{\forall i\in A \\ j\in B}} d_{i,j}$
5. Bütün gözlem birimleri tek bir kümede birleşene kadar iterasyonlar devam ettirilir.

Burada $d(i,j)$; i. ve j. gözlemler arasındaki mesafeyi ifade etmektedir (Özkan, 2013).

Ortalama Bağlantı Yöntemi (Average Linkage)

En Yakın Komşu Algoritması (Single Linkage) ve En Uzak Komşu Algoritmalarında (Complete Linkage) mesafelerin hesaplanmasında minimum ve maksimum değerlerin kullanılmasından dolayı aykırı değerlere daha duyarlıdır. Ortalama Bağlantı Yönteminde gözlemler arasındaki mesafeler hesaplanırken gözlem değerlerinin ortalaması hesaplanır. Bu durum aykırı değerlerin var olduğu veri setlerinde daha iyi sonuç alınmasına olanak sağlamaktadır (Han, 2023).

Ortalama Bağlantı Algoritması İterasyonlar

1. Öncelikle gözlemler arasındaki uzaklıklar belirlenir.
2. Daha sonra uzaklıklar dikkate alınarak $min\, d(i,j)$ belirlenir. (En Yakın Komşu Algoritmasında olduğu gibi)
3. Minimum uzaklığa sahip gözlemlerden oluşan yeni bir küme elde edilir. Bu duruma göre uzaklıklar yeniden hesaplanır.
4. Birden fazla gözlem değerine sahip iki küme söz konusu olduğunda iki kümenin birbirine olan uzaklığı hesaplanırken kümeyi oluşturan gözlem mesafelerinin ortalaması alınır (Everitt, 2011).
   $d_{A\to B} = \operatorname*{ort}_{\substack{\forall i\in A \\ j\in B}} d_{i,j}$
5. Bütün gözlem birimleri tek bir kümede birleşene kadar iterasyonlar devam ettirilir.

Burada $d(i,j)$; i. ve j. gözlemler arasındaki mesafeyi ifade etmektedir.

Ward Yöntemi (Ward's Method)

Ward Yöntemi 1963 yılında Ward tarafından geliştirilmiştir (Ward, 1963). Daha önce bahsettiğimiz hiyerarşik algoritmalarda gözlemler arası mesafeye dayanan bir yaklaşım mevcut iken Ward Algoritması K-Means yöntemine benzer şekilde grup içi varyanslarını en küçük yapacak gözlemleri biraraya getirmeyi hedeflemektedir. Bu nedenle bu yöntemde elde edilen küme büyüklüklerinin yaklaşık olarak eşit olma eğiliminde olduğu görülmektedir (Hair vd., 2014).

3.2.2. Ayrıştırıcı (Divisive) Yöntemler

Hiyerarşik Kümeleme Yöntemleri başlığı altnda incelenen bir diğer grup kümeleme algoritması ise Ayrıştırıcı Yöntemler olarak incelenmektedir. Ayrıştırıcı Yöntemler, Birleştirici Yöntemlerden farklı olarak başlangıçta incelenen veri setini tek bir küme olarak kabul ederek daha sonra birbirinden farklı özelliklerine göre veri setini gruplara ayırmaktadır. Ancak bu yöntem Ayrıştırıcı Yöntemlere göre uygulanabilirliği daha zor olduğundan daha az tercih edilmektedir.

3.3. Yoğunluk Bazlı Yöntemler (Density Based Methods)

Yoğunluk Bazlı Yöntemler, veri seti içindeki gruplanmaların konveks yapıda olmadığı durumlarda kümeleri ayrıştırmada oldukça başarılıdır. DBSCAN Algoritması Yoğunluk Bazlı Yöntemler içerisinde en çok literatürde karşımıza çıkan algoritmalardan birisidir. Yoğunluk Bazlı Yöntemlerin en önemli avantajı, analiz öncesinde küme sayının belirlenmesinin gerekmemesi, gürültü içeren veri setlerinde kullanılabiliyor olması ve aykırı değerlerin belirlenmesinde başarılı olmasıdır (Kameshwaran, 2014). Ancak boyut nedeniyle çok büyük boyutlu veri setlerinde kullanımı elverişli değildir (Ezugwu, 2022).

3.4. Izgara Bazlı Yöntemler (Grid Based Methods)

Çok büyük boyutlu veri setlerinde Yoğunluk Bazlı Yöntemlerin etkili bir çözüm sunamamasına karşılık, Izgara Bazlı Yöntemler veri setini küçük parçalara ayırarak incelemekte bu nedenle de daha hızlı bir çözüm sunabilmektedir (Han, vd., 2023). Izgara bazlı yöntemlerde veri seti ızgara yapısına benzer şekilde küçük parçalara ayrılmakta ve her bir parçanın yoğunluğu hesaplanmaktadır. Daha önceden belirlenen bir eşik değere göre her bir parçanın yoğunluğuna karar verilmektedir.

3. Bulgular

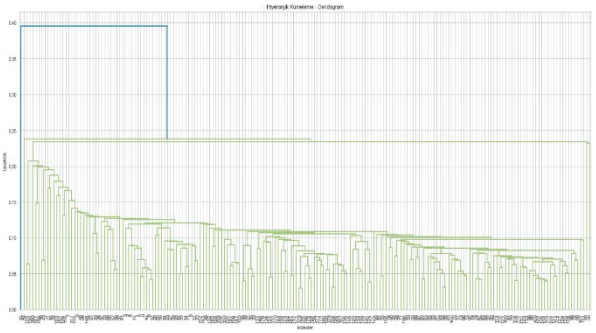
Bu çalışmada hiyerarşik olmayan yöntemlerden K-Means Algoritması ve K-Medoids Algoritması, hiyerarşik yöntemlerden En Yakın Komşuluk Algoritması (Single Linkage), En Uzak Komşu Algoritması (Complete Linkage), Ortalama Bağlantı Yöntemi (Average Linkage), Ward Yöntemi ve Yoğunluk Bazlı Yöntemlerden DBSCAN Algoritmasının kullanıldığı 55 farklı model oluşturulmuştur. Ancak Silhouette Skoruna göre anlamlı 39 model çalışmada sunulmuştur. Ayrıca metodoloji kısmında açıklanan CLARA Algoritması ve Izgara Bazlı Yöntemler verisetinin boyutu nedeniyle analize dahil edilmemiştir. Elde edilen sonuçlar bu veri seti üzerinden Birleşmiş Milletler Kalkınma Programı tarafından yapılmış gerçek sınıflamalarla karşılaştırılmıştır. Çalışma Birleşmiş Milletler Kalkınma Programı tarafından açıklanan en güncel veri seti üzerinde yapılmıştır. Verisetindeki değerler normalize edildikten sonra sırasıyla En Yakın Komşu Algoritması (Single Linkage), Average Linkage, En Uzak Komşu Algoritması (Complete Linkage) ve Ward Yöntemleri için dendrogramlar elde

edilmiştir. Elde edilen diyagram sonuçlarına göre 4 veya 5 kümelemenin uygun olacağı tespit edilmiştir.

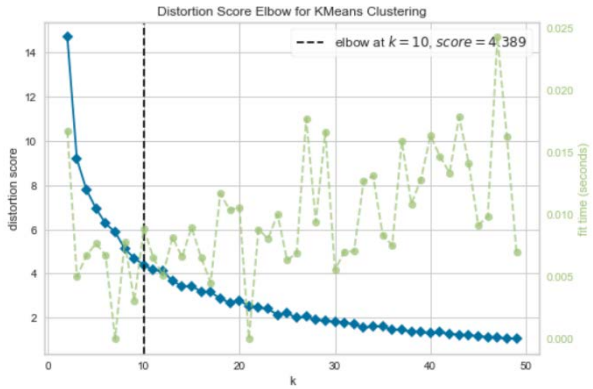
Tablo 1. Silhouette skoruna göre model sıralaması

No	Model	Standardizasyon	k değeri	Silhouette Skoru
1	Ward(k=6)	MinMax	6	0.79
2	Completelink(k=6)	MinMax	6	0.78
3	Averagelink(k=6)	MinMax	6	0.77
4	Ward(k=4)	MinMax	4	0.77
5	Ward(k=5)	MinMax	5	0.77
6	Completelink(k=5)	MinMax	5	0.76
7	Averagelink(k=4)	MinMax	4	0.74
8	Averagelink(k=5)	MinMax	5	0.74
9	Completelink(k=4)	MinMax	4	0.74
10	Kmedoids(k=4)	MinMax	4	0.71
11	Completelink(k=4)	StdScaler	4	0.6
12	Completelink(k=5)	StdScaler	5	0.6
13	Singlelink(k=4)	MinMax	4	0.54
14	Singlelink(k=4)	StdScaler	4	0.54
15	Singlelink(k=5)	MinMax	5	0.53
16	Singlelink(k=6)	MinMax	6	0.53
17	Averagelink(k=4)	StdScaler	4	0.5
18	Averagelink(k=5)	StdScaler	5	0.5
19	Ward(k=4)	StdScaler	4	0.5
20	Averagelink(k=6)	StdScaler	6	0.49
21	Kmedoids(k=4)	StdScaler	4	0.48
22	Completelink(k=6)	StdScaler	6	0.48
23	Ward(k=6)	StdScaler	6	0.46
24	Ward(k=5)	StdScaler	5	0.38
25	Kmeans(k=4)	MinMax	4	0.33
26	Kmedoids(k=5)	MinMax	5	0.32
27	Kmeans(k=4)	StdScaler	4	0.31
28	Kmeans(k=8)	MinMax	8	0.31
29	Kmeans(k=5)	StdScaler	5	0.3
30	Kmeans(k=7)	StdScaler	7	0.29
31	Kmeans(k=5)	MinMax	5	0.28
32	Kmeans(k=6)	MinMax	6	0.28
33	Kmeans(k=8)	StdScaler	8	0.28
34	Dbscan(k=6)	StdScaler	6	0.28
35	Singlelink(k=6)	StdScaler	6	0.27
36	Kmeans(k=7)	MinMax	7	0.26
37	Kmedoids(k=5)	StdScaler	5	0.24
38	Kmeans(k=6)	StdScaler	6	0.23
39	Singlelink(k=5)	StdScaler	5	0.17

K-Means ve K-Medoids Algoritmalarını uygulamadan önce küme sayısının belirlenmesinde Elbow Yöntemi kullanılmış ancak yapılan farklı denemelere göre k değerinin 10 ve üzerinde olması gerektiği sonucu elde edilmiştir. Şekil 1' de Hiyerarşik Kümeleme Algoritmaları için yapılan dendrogram sonucu görülmektedir. Buna göre 4 veya 5 kümelenmenin uygun olacağı görülmektedir. Şekil 2' de ise Hiyerarşik Olmayan Kümeleme Algoritmaları için küme sayısının belirlenmesinde kullanılan Elbow Grafiği görülmektedir.
Küme sayısı Birleşmiş Milletlerin öngördüğü şekilde 4 küme olacak şekilde ve bununla birlikte Dendrogram ve Elbow Yöntemlerinin sonuçları dikkate alınarak 4 ile 10 arasında farklı küme sayıları için de belirtilen tüm yöntemler uygulanmıştır. Oluşturulan kümelerin uygunluğu için Silhouette Skoru değerleri elde edilmiştir. Silhouette Skoru değerine göre modeller en yüksek skoru veren modelden en düşük skoru veren modele doğru Tablo 1' de verilmiştir.



Şekil1. Hiyerarşik kümeleme için dendrogram grafiği.



Şekil2. Hiyerarşik olmayan kümeleme için Elbow grafiği

Elde edilen sonuçlara göre en iyi kümelenme Ward tekniği ile k küme sayısının 6 olması durumunda elde edilmiştir. Ancak Birleşmiş Milletlerin öngördüğü şekilde 4 küme olmaya zorladığımızda da yine en iyi sonuç Ward Tekniği ile elde edilmiştir. Tablo 2' de Ward Tekniği 6 küme için yapıldığında elde edilen kümeleme sonucu verilmiştir.
Tablo 2, Tablo3, Tablo4, Tablo5, Tablo6 ve Tablo7' de Ward Tekniği 6 küme için yapıldığında elde edilen kümeleme sonuçları her küme için ayrı ayrı tablolar ile verilmiştir. Bu tablolar Ward Tekniği 6 küme için yapıldığında elde edilen sonuçları ve Birleşmiş Milletler tarafından yapılan sıralamayı da içerecek biçimde düzenlenmiştir.
Tablo 8, Tablo 9, Tablo 10 ve Tablo 11 ise Birleşmiş Milletlerin öngördüğü biçimde 4 küme oluşturulduğunda elde edilen sonuçları her bir küme ayrı bir tablo olacak şekilde göstermektedir. 4 küme için de en iyi sonuç Ward Tekniği ile elde edilmiştir.

Tablo2. Ward Tekniğine göre oluşturulan k=6 küme için Küme 1 Sonuçları

Küme No	Ülke No	BMI Sıralaması
1	İsviçre	1
1	Norveç	2
1	İzlanda	3
1	Hong Kong, Çin (SAR)	4
1	Danimarka	5
1	İsveç	5
1	İrlanda	7
1	Almanya	7
1	Singapur	9
1	Avustralya	10
1	Hollanda	10
1	Finlandiya	12
1	Belçika	12
1	Lihtenştayn	12
1	Birleşik Krallık	15
1	Yeni Zelanda	16
1	Birleşik Arap Emirlikleri	17
1	Kanada	18
1	Kore (Cumhuriyeti)	19
1	Lüksemburg	20
1	Amerika Birleşik Devletleri	20
1	Slovenya	22
1	Avusturya	22
1	Japonya	24
1	İsrail	25
1	Malta	25
1	İspanya	27
1	Fransa	28
1	Kıbrıs	29
1	İtalya	30
1	Estonya	31
1	Çekya	32
1	Yunanistan	33
1	Bahreyn	34
1	Andorra	35
1	Polonya	36
1	Litvanya	37
1	Letonya	37
1	Hırvatistan	39
1	Suudi Arabistan	40
1	Katar	40
1	Portekiz	42
1	San Marino	43
1	Şili	44
1	Slovakya	45
1	Macaristan	47
1	Arjantin	48
1	Kuveyt	49
1	Karadağ	50
1	Brunei Darussalam	55

Tablolar incelendiğinde, Ward Tekniği ile elde edilen sonuçların Birleşmiş Milletler tarafından açıklanan listeyle uyumlu olduğu görülmektedir. Bu durum uygulanan kümeleme algoritmalarının kullanılmasının pratikte önemli bir katkı sağlayacağını göstermektedir.

Tablo 3. Ward Tekniğine göre oluşturulan k=6 küme için Küme 2 Sonuçları

Küme No	Ülke No	BMI Sıralaması
2	Fas	120
2	Butan	125
2	Hindistan	134
2	Guatemala	136
2	Lao Demokratik Halk Cumhuriyeti	139
2	Sao Tome ve Principe	141
2	Namibya	142
2	Esvatini (Krallık)	142
2	Myanmar	144
2	Nepal	146
2	Kenya	146
2	Kamboçya	148
2	Kongo	149
2	Kongo (Demokratik Cumhuriyeti)	149
2	Angola	150
2	Kamerun	151
2	Komorlar	152
2	Zambiya	153
2	Doğu Timor	155
2	Solomon Adaları	156
2	Suriye Arap Cumhuriyeti	157
2	Zimbabve	159
2	Nijerya	161
2	Togo	163
2	Tanzanya (Birleşik Cumhuriyeti)	167
2	Lesoto	168
2	Malawi	172

Tablo 4. Ward Tekniğine göre oluşturulan k=6 küme için Küme 3 Sonuçları

Küme No	Ülke No	BMI Sıralaması
3	Ekvator Ginesi	133
3	Papua Yeni Gine	157
3	Haiti	158
3	Uganda	159
3	Ruanda	161
3	Moritanya	164
3	Pakistan	164
3	Fildişi Sahili	166
3	Senegal	169
3	Sudan	170
3	Cibuti	171
3	Benin	173
3	Gambiya	174
3	Eritre	175
3	Etiyopya	176
3	Madagaskar	177
3	Liberya	177
3	Gine-Bissau	179
3	Gine	181
3	Afganistan	182
3	Mozambik	183
3	Sierra Leone	184
3	Burkina Faso	185
3	Yemen	186
3	Burundi	187
3	Mali	188
3	Nijer	189
3	Çad	189
3	Orta Afrika Cumhuriyeti	191
3	Güney Sudan	192

Tablo 5. Ward Tekniğine göre oluşturulan k=6 küme için Küme 4 Sonuçları

Küme No	Ülke No	BMI Sıralaması
4	Romanya	53
4	Rusya Federasyonu	56
4	Bahamalar	57
4	Panama	57
4	Umman	59
4	Trinidad ve Tobago	60
4	Gürcistan	60
4	Malezya	63
4	Sırbistan	65
4	Kazakistan	67
4	Seyşeller	67
4	Belarus	69
4	Bulgaristan	70
4	Palau	71
4	Mauritius	72
4	Arnavutluk	74
4	Ermenistan	76
4	Sri Lanka	78
4	İran (İslam Cumhuriyeti)	78
4	Bosna Hersek	80
4	Saint Vincent ve Grenadinler	81
4	Kuzey Makedonya	83
4	Küba	85
4	Moldova (Cumhuriyeti)	86
4	Azerbaycan	89
4	Türkmenistan	94
4	Tonga	98
4	Ürdün	99
4	Ukrayna	100
4	Fiji	104
4	Özbekistan	106
4	Lübnan	109
4	Filistin Devleti	111
4	Samoa	116
4	Kırgızistan	117
4	Venezuela (Bolivar Cumhuriyeti)	119
4	Tacikistan	126

Tablo 6. Ward Tekniğine göre oluşturulan k=6 küme için Küme 5 Sonuçları

Küme No	Ülke No	BMI Sıralaması
5	Libya	92
5	Guyana	95
5	Dominika	97
5	Paraguay	102
5	Marshall Adaları	102
5	Mısır	105
5	Vietnam	107
5	Aziz Lucia	108
5	Güney Afrika	110
5	Endonezya	112
5	Filipinler	113
5	Botsvana	114
5	Jamaika	115
5	Belize	118
5	Bolivya (Çokuluslu Devlet)	120
5	Gabon	123
5	Surinam	124
5	El Salvador	127
5	Irak	128
5	Bangladeş	129
5	Nikaragua	130
5	Yeşil Burun Adaları	131
5	Tuvalu	132
5	Mikronezya (Federal Devletleri)	135
5	Kiribati	137
5	Honduras	138
5	Vanuatu	140
5	Gana	145

Tablo 7. Ward Tekniğine göre oluşturulan k=6 küme için Küme 6 Sonuçları

Küme No	Ülke No	BMI Sıralaması
6	Türkiye	45
6	Saint Kitts ve Nevis	51
6	Uruguay	52
6	Antigua ve Barbuda	54
6	Barbados	62
6	Kosta Rika	64
6	Tayland	66
6	Grenada	73
6	Çin	75
6	Meksika	77
6	Dominik Cumhuriyeti	82
6	Ekvador	83
6	Peru	87
6	Maldivler	87
6	Brezilya	89
6	Kolombiya	91
6	Cezayir	93
6	Moğolistan	96
6	Tunus	101

Tablo 8. Ward Tekniğine göre oluşturulan k=4 küme için Küme 1 Sonuçları

Küme No	Ülke No	BMI Sıralaması
1	Fas	120
1	Butan	125
1	Ekvator Ginesi	133
1	Hindistan	134
1	Guatemala	136
1	Lao Demokratik Halk Cumhuriyeti	139
1	Sao Tome ve Principe	141
1	Namibya	142
1	Esvatini (Krallık)	142
1	Myanmar	144
1	Nepal	146
1	Kenya	146
1	Kamboçya	148
1	Kongo	149
1	Kongo (Demokratik Cumhuriyeti)	149
1	Angola	150
1	Kamerun	151
1	Komorlar	152
1	Zambiya	153
1	Doğu Timor	155
1	Solomon Adaları	156
1	Suriye Arap Cumhuriyeti	157
1	Papua Yeni Gine	157

1	Haiti	158
1	Zimbabve	159
1	Uganda	159
1	Nijerya	161
1	Ruanda	161
1	Togo	163
1	Moritanya	164
1	Pakistan	164
1	Fildişi Sahili	166
1	Tanzanya (Birleşik Cumhuriyeti)	167
1	Lesoto	168
1	Senegal	169
1	Sudan	170
1	Cibuti	171
1	Malawi	172
1	Benin	173
1	Gambiya	174
1	Eritre	175
1	Etiyopya	176
1	Madagaskar	177
1	Liberya	177
1	Gine-Bissau	179
1	Gine	181
1	Afganistan	182
1	Mozambik	183
1	Sierra Leone	184
1	Burkina Faso	185
1	Yemen	186
1	Burundi	187
1	Mali	188
1	Nijer	189
1	Çad	189
1	Orta Afrika Cumhuriyeti	191
1	Güney Sudan	192

Tablo 9. Ward Tekniğine göre oluşturulan k=4 küme için Küme 2 Sonuçları

Küme No	Ülke	BMI Sıralaması
2	Türkiye	45
2	Saint Kitts ve Nevis	51
2	Uruguay	52
2	Romanya	53
2	Antigua ve Barbuda	54
2	Rusya Federasyonu	56
2	Bahamalar	57
2	Panama	57
2	Umman	59
2	Trinidad ve Tobago	60
2	Gürcistan	60
2	Barbados	62
2	Malezya	63
2	Kosta Rika	64
2	Sırbistan	65
2	Tayland	66
2	Kazakistan	67
2	Seyşeller	67
2	Belarus	69
2	Bulgaristan	70
2	Palau	71
2	Mauritius	72
2	Grenada	73
2	Arnavutluk	74
2	Çin	75
2	Ermenistan	76
2	Meksika	77
2	Sri Lanka	78
2	İran (İslam Cumhuriyeti)	78
2	Bosna Hersek	80
2	Saint Vincent ve Grenadinler	81
2	Dominik Cumhuriyeti	82
2	Kuzey Makedonya	83
2	Ekvador	83
2	Küba	85
2	Moldova (Cumhuriyeti of)	86
2	Peru	87
2	Maldivler	87
2	Brezilya	89
2	Azerbaycan	89
2	Kolombiya	91
2	Cezayir	93
2	Türkmenistan	94
2	Moğolistan	96
2	Tonga	98
2	Ürdün	99
2	Ukrayna	100
2	Tunus	101
2	Fiji	104
2	Özbekistan	106
2	Lübnan	109
2	Filistin, Devlet	111
2	Samoa	116
2	Kırgızistan	117
2	Venezuela (Bolivar Cumhuriyeti)	119
2	Tacikistan	126

Tablo 10. Ward Tekniğine göre oluşturulan k=4 küme için Küme 3 Sonuçları

Küme No	Ülke	BMI Sıralaması
3	İsviçre	1
3	Norveç	2
3	İzlanda	3
3	Hong Kong, Çin (SAR)	4
3	Danimarka	5
3	İsveç	5
3	İrlanda	7
3	Almanya	7
3	Singapur	9
3	Avustralya	10
3	Hollanda	10
3	Finlandiya	12
3	Belçika	12
3	Lihtenştayn	12
3	Birleşik Krallık	15
3	Yeni Zelanda	16
3	Birleşik Arap Emirlikleri	17
3	Kanada	18
3	Kore (Cumhuriyeti)	19
3	Lüksemburg	20
3	Amerika Birleşik Devletleri	20
3	Slovenya	22
3	Avusturya	22
3	Japonya	24
3	İsrail	25
3	Malta	25
3	İspanya	27
3	Fransa	28

3	Kıbrıs	29
3	İtalya	30
3	Estonya	31
3	Çekya	32
3	Yunanistan	33
3	Bahreyn	34
3	Andorra	35
3	Polonya	36
3	Litvanya	37
3	Letonya	37
3	Hırvatistan	39
3	Suudi Arabistan	40
3	Katar	40
3	Portekiz	42
3	San Marino	43
3	Şili	44
3	Slovakya	45
3	Macaristan	47
3	Arjantin	48
3	Kuveyt	49
3	Karadağ	50
3	Brunei Darussalam	55

Tablo 11. Ward Tekniğine göre oluşturulan k=4 küme için Küme 4 Sonuçları

Küme No	Ülke	BMI Sıralaması
4	Libya	92
4	Guyana	95
4	Dominika	97
4	Paraguay	102
4	Marshall Adaları	102
4	Mısır	105
4	Vietnam	107
4	Saint Lucia	108
4	Güney Afrika	110
4	Endonezya	112
4	Filipinler	113
4	Botsvana	114
4	Jamaika	115
4	Belize	118
4	Bolivya (Çokuluslu Devlet)	120
4	Gabon	123
4	Surinam	124
4	El Salvador	127
4	Irak	128
4	Bangladeş	129
4	Nikaragua	130
4	Yeşil Burun Adaları	131
4	Tuvalu	132
4	Mikronezya (Federal Devletleri)	135
4	Kiribati	137
4	Honduras	138
4	Vanuatu	140
4	Gana	145

4. Sonuçlar ve Tartışma

Birleşmiş Milletler Kalkınma Programı tarafından ilk olarak 1990 yılında oluşturulmuş ve bu yıldan itibaren her yıl düzenli olarak sunulmakta olan İnsani Gelişim İndeksi sağlıklı yaşam, eğitim ve temel yaşam standartlarına erişim gibi üç temel göstergeye dayanmaktadır.

Bu çalışmada denetimsiz makine öğrenmesi tekniklerinden kümeleme algoritmaları başlığı altında incelenen Hiyerarşik ve Hiyerarşik Olmayan Yöntemler açıklanmış uygulamada Hiyerarşik Olmayan Yöntemlerden K-Means Algoritması ve K-Medoids Algoritması, Hiyerarşik Yöntemlerden En Yakın Komşuluk Algoritması (Single Linkage), En Uzak Komşu Algoritması (Complete Linkage), Ortalama Bağlantı Yöntemi (Average Linkage), Ward Yöntemi ve Yoğunluk Bazlı Yöntemlerden DBSCAN Algoritması kullanılmıştır. Uygulanan bu yöntemlerde k sayısının belirlenmesinde Hiyerarşik Yöntemler için Dendrogram Grafikleri ve Hiyerarşik Olmayan Yöntemler için Elbow Grafiklerinden yararlanılmıştır. Bu analizler dikkate alınarak k sayısının 4 ile 10 arasında değiştiği farklı durumlar için 39 farklı model oluşturulmuştur. Oluştulan modeller elde edilen kümelerin uygunluğunun değerlendirilmesinde kullanılan Silhoutte İndeksi değerlerine göre en yüksekten en düşük Silhoutte değerine sahip modele göre sıralanmıştır. Buna göre Hiyerarşik Tekniklerden Ward Tekniğine göre 6 küme şeklinde yapılan gruplamalar en yüksek Silhoutte skor değerini vermiştir. Ancak model Birlemiş Milletler Kalkınma programının öngördüğü biçimde 4 küme olarak yapıldığında elde edilen kümelenmeler de tablo halinde verilmiştir. Elde edilen sonuçlar, Birleşmiş Milletler Kalkınma programı tarafından yapılmış gerçek sınıflamalarla karşılaştırıldığında sonuçların tutarlı olduğu gözlenmektedir. Bu durum uygulanan kümeleme algoritmalarının pratikte de kullanılabileceğini göstermekte, uzman kişiler için yol gösterici olabileceğine de işaret etmektedir. Bu çalışmada pek çok algoritmanın bir arada incelenerek araştırmacılar için detaylı bir analiz sunması hedeflenmiştir. Bu kapsamda, yapılan çalışmanın, kullanılan yöntemlerin avantajları ve dezavantajları konusunda derinlemesine incelemeler yapması bakımından literatüre önemli bir katkı sağlayacağı düşünülmektedir. Bununla birlikte bu çalışmada ülkelerin gelişmişlik düzeylerine göre ayrımlarının kesin sınırlarla belirlenmesi yerine, makine öğrenmesi tekniklerinden kümeleme algoritmaları kullanılarak değişen koşullara göre gruplamaların yapılabileceği yönünde bir öneri sunulmaktadır.

Katkı Oranı Beyanı

Yazarın katkı yüzdeleri aşağıda verilmiştir. Yazar makaleyi incelemiş ve onaylamıştır.

	L.S.S.
K	100
T	100
Y	100
VTI	100
VAY	100
KT	100
YZ	100
KI	100
GR	100
PY	100
FA	100

K= kavram, T= tasarım, Y= yönetim, VTI= veri toplama ve/veya işleme, VAY= veri analizi ve/veya yorumlama, KT= kaynak tarama, YZ= Yazım, KI= kritik inceleme, GR= gönderim ve revizyon, PY= proje yönetimi, FA= fon alımı.

Çatışma Beyanı

Yazar bu çalışmada hiçbir çıkar ilişkisi olmadığını beyan etmektedir.

Etik Onay Beyanı

Bu araştırmada hayvanlar ve insanlar üzerinde herhangi bir çalışma yapılmadığı için etik kurul onayı alınmamıştır.

Kaynaklar

Berkhin P. 2002. A survey of clustering data mining techniques, In Grouping multidimensional data: Recent advances in clustering. Springer, Berlin, Germany, pp: 25-71.

Bramer M. 2016, Principles of data mining, 3rd. ed. Springer, London,UK, pp: 221-238.

Everitt BS. 2011. Cluster Analysis. Wiley, Londok, UK, ss: 15-110.

Ezugwu AE, Ikotun AM, Oyelade OO, Abualigah L, Agushaka JO, Eke CI, Akinyelu AA. 2022. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Eng Appl Artif Intel, 110: 104743.

Hair JF, Black WC, Babin BJ, Anderson RE. 2014. Multivariate statistical analysis. Pearson, New York, US, pp: 415-474.

Han J, Kamber M, Pei J. 2023. Data mining concepts and techniques. Morgan Kaufmann Publications, San Francisco, US, pp: 379-425.

HDI Reports. 2024. Human development index. URL: https://hdr.undp.org/data-center/human-development-index#/indicies/HDI (accesed date: March 11, 2024).

James G, Witten D, Hastie T, Tibshirani R. 2013. An Introduction to Statistical Learning with Applications in R. Springer, New York, US, pp: 516-542.

Kameshwaran K, Malarvizhi K. 2014. Survey on clustering techniques in data mining. Int J Comput Sci Info Technol, 5(2): 2272-2276.

Kassambara A. 2017. Practical guide to clustering analysis in R, unsupervised machine learning. STHDA, Marseille, France, pp: 17-185.

Kaufman L, Rousseeuw P. 1990. Finding groups in data: an introduction to cluster analysis. John Wiley& Sons, New York, USA, pp: 163.

Kurnaz B, Yüksel HM, Önder H, Tırınk C. 2022. 3-D Classification of agricultural areas of Turkey using mammalian livestock existence. BSJ Agri, 5(3): 311-313.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. Fifth Berkeley Symposium on Mathematical Statistics and Probability, December 27, Berkeley, US, pp: 281-297.

Morissette L, Chartier S. 2013. The k-means clustering technique: General considerations and implementation in Mathematica. Tutor Quantit Meth Psychol, 9(1): 15-24.

Muttaqin MFJ. 2022. Cluster analysis using k-means method to classify sumatera regency and city based on human development ındex ındicator. Nas Offic Stat, 2022: 967-976.

Neha D, Vidyavathi BM. 2015. A survey on applications of data mining using clustering techniques. Int J Comput Appl, 126(2): 7-12.

Nurhasanah N, Salwa N, Ornila L, Hasan A, Mardhani M. 2021. Classifying regencies and cities on human development index dimensions: Application of K-Means cluster analysis. J Sains Sosio Humaniora, 5(2): 759-765.

Özkan Y. 2013. Veri madenciliği yöntemleri. Papatya Yayınclık, İstanbul, Türkiye, ss: 131-156.

Pujari AK. 2001. Clustering techniques in data mining- A survey. JETE J Res, 47(1&2): 19-28.

Rocha JLM, Zela MAC, Torres NIV, Medina GS. 2021. Analogy of the application of clustering and K-means techniques for the approximation of values of human development indicators. Int J Adv Comput Sci Appl, 12(9): 526-532.

Shah M, Nair S. 2015. A survey of data mining clustering algorithms. Int J Comput Appl, 128(1): 1-5.

Wang H, Feil JH, Yu X. 2023. Let the data speak about the cut-off values for multidimensional index: Classification of human development index with machine learning, Socio-Econ Plan Sci, 87:101523.

Ward JH. 1963. Hierarchical grouping to optimize an objective function. J Amer Stat Assoc, 5(301): 236-244.