Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Veri Madenciliğinde Kullanılan Kümeleme Algoritmalarının Karşılaştırılması Üzerine Bir İnceleme: Ülkelerin Beşeri Sermaye Durumlarına Göre Kümelenmesi Örneği

Ömer Faruk RENÇBER[a*]

[a] *İşletme Bölümü, İktisadi ve İdari Bilimler Fakültesi, Gaziantep Üniversitesi, Gaziantep, TÜRKİYE*

** Sorumlu yazarın e-posta adresi: ofrencber@gantep.edu.tr*

DOI : 10.29130/dubited.551531

ÖZET

Günümüz dünyasında veri madenciliği, yaşanan olayların anlaşılabilmesi, yorumlanabilmesi ve geleceğe dair tahminlerin yapılabilmesi için büyük önem arz etmektedir. Bu nedenle, istatistik teknikleri her geçen gün değişmekte ve yenilenmektedir. Özellikle, günümüzde büyük verilerin anlaşılabilmesi amacıyla makine öğrenme teknikleri sıklıkla kullanılmaktadır. Bu çalışmada, literatürde yoğun olarak kullanılan k-ortalama kümeleme algoritması çeşitlerinden klasik, bulanık ve torbalı k-ortalamalar yöntemlerinin kümeleme performanslarının karşılaştırılması amaçlanmaktadır. Bu doğrultuda veri setine ulaşılabilen 132 ülke beşeri sermaye özellikleri doğrultusunda kümelenmiştir.

Çalışmanın sonucunda, torbalı küme algoritmasının zaman açısından diğerlerinden daha yavaş olduğu ancak daha başarılı kümeleme yaptığı bulgusuna ulaşılmıştır. Benzer şekilde, bulanık k-ortalama algoritmasının klasik k-ortalamalara göre daha başarılı olduğu görülmüştür.

Anahtar Kelimeler: *K ortalamalar, Bulanık mantık, Kümeleme Analizi*

A Study on the Comparison of Clustering Algorithms Used in Data Mining

ABSTRACT

In today's world of digital, data mining is very important for predicting, interpreting and understanding of cases. Therefore, statistical techniques are been renewing and develop. Especially, machine learning techniques often use for this. In this study, it is aimed to compare of performance classic, fuzzy and bagged k-means cluster algorithm. Therefore, 132 countries which reached data was clustered according to human capital.

In the conclusion of the study, it is reached that bagged cluster algorithm is slower than other algorithms, but it has more correct classify. Similarly, the fuzzy cluster algorithm has more correct classify than the classic k-means algorithm.

Keywords: *K-means, Fuzzy Logic, Cluster Analysis*

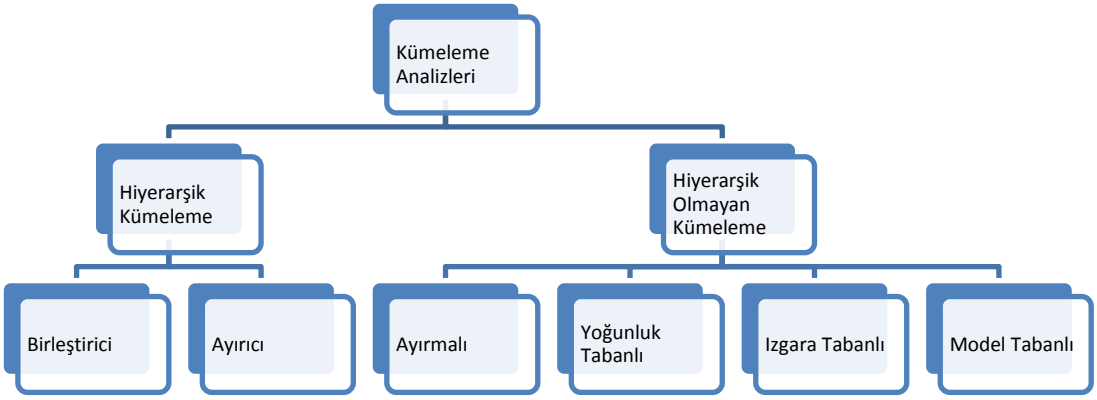
Geliş: 09/04/2019, Düzeltme: 21/05/2019, Kabul: 27/06/2019

I. GİRİŞ

Veri madenciliği, büyük miktarda verinin anlaşılması, analiz edilmesi, yorumlanması ve bilgiye dönüştürülmesi sürecini ifade eden bir kavramdır [1]. Veri madenciliğindeki temel amaç; verilerin kullanımı ile bilginin keşfedilebilmesidir. Bu alandaki yöntemler; örüntü tanıma, model oluşturma veya oluşturulan modele dayanarak tahminler yapabilme gibi amaçlar içermektedirler. Bununla birlikte günümüzde dijital verilerdeki artış, beraberinde kolay analiz edilememe ve yorumlanamama gibi sorunları da getirmektedir.

Veri madenciliğinde sınıflama ve kümeleme yöntemleri, sosyal, fen veya tıp gibi çoğu bilimsel alanda, gözlemlerin daha sade ve anlaşılır olarak yorumlanabilmesi amacıyla sıklıkla kullanılmaktadır. Bu doğrultuda uygulanan yöntemler; danışmalı veya danışmansız olmak üzere iki grupta incelenmektedir. Danışmalı tekniklerde; gözlemin ait olduğu küme ya da sınıfı bilinmektedir ancak, yöntemin gözlemleri nasıl sınıflandığının tespit edilebilmesi amacıyla uygulanmaktadır. Danışmansız öğrenmede ise, gözlemlere ait küme sayısı belli değildir. Buna göre yöntem, belirli bir metoda dayanarak kümeleme yapmaktadır.

Kümeleme analizi, veri setindeki gözlemleri grup içinde maksimum benzerlik ve gruplar arası maksimum farklılık mantığına göre değerlendirmektedir [2]. Bunun için farklı teknikler kullanılmakta olup yöntemin çeşitleri özet olarak şekil 1'deki gibidir.



Şekil 1. Kümeleme Analizi Yöntem Çeşitleri

Şekil 1'de görüldüğü üzere kümeleme analizi; hiyerarşik olup olmama durumuna göre iki grupta incelenmektedir. Bunlardan hiyerarşik kümeleme; küme içi minimum ve kümeler arası maksimum uzaklık düşüncesi ile çalışmaktadır. Birleştirici kümelemede başlangıçta, gözlemlerin her biri ayrı birer küme olarak kabul edilir. Daha sonra, birbirine en yakın özelliklere sahip gözlemler birleştirilerek kümeleme işlemi yapılır. Ayırıcı kümelemede ise başlangıçta bütün gözlemler bütün olarak bir kümede yer alır ve aşamalar halinde farklı kümelere ayrılır.

Hiyerarşik olmayan kümeleme analizinde, küme sayısı belirli olup, gözlemler bu küme sayısına göre uygun bir şekilde atanır [4]. Bu tür kümeleme analizlerinden en çok bilinenleri; k-ortalama, k-medoid, bulanık ve yığma kümeleme yöntemleridir. Bununla birlikte, yıllar içerisinde bu alanda yeni algoritmalar ve yeni yaklaşımlar ortaya atılmıştır. Bu yaklaşımlardan ayırmalı kümeleme yöntemlerinden k-ortalamalar 1967 yılında J.B.MacQueen tarafından [3], bulanık k-ortalamalar 1973 yılında Dunn tarafından [5] ve torbalı k-ortalamalar yöntemi ise 1999 yılında Leisch tarafından [6] geliştirilmiştir. Algoritmaların üçü de ortalamaya uzaklığa göre kümeleme mantığı ile çalışmaktadır. Ancak, gözlemlerin ortalamaya göre nasıl atanacağı konusunda farklılık göstermektedirler.

Bulanık k-ortalamalar yönteminde gözlemler, kümelere üyelik derecelerine göre atanır. Torbalı k-ortalamalar yönteminde ise gözlemler, bootstrap yöntemi ile alt gözlemlere ayrılır ve alt uygulamalar ile nihai kümeler belirlenir.

Literatürde k-ortalama algoritmalarının karşılaştırıldığı birçok çalışma bulunmaktadır. Örneğin, klasik k-ortalama kümeleme algoritması ile SOM (Self Organization Map) [7]; k-ortalama ile k-medoid [8]; torbalı k-ortalamalar yönteminde farklı algoritmalarının[4] ve bütünleşik kümeleme yöntemlerinin[9] karşılaştırıldığı görülmektedir.

Bu çalışmada ise, klasik, bulanık ve torbalı k-ortalama kümeleme algoritmalarının karşılaştırılması amaçlanmaktadır. Bu doğrultuda, 132 ülke beşeri sermaye özelliklerine göre üç yöntem ile kümelere ayrılmıştır. Daha sonra, bu algoritmalar ile elde edilen bulgular matematiksel olarak karşılaştırılmıştır. Bu doğrultuda çalışmanın ilk bölümünde beşeri sermaye kavramı ve kullanılan değişkenlerin tanımlarına yer verilmiştir. Ardından, yöntemler teorik olarak incelenmiş ve uygulama aşamasında 132 ülke üç yönteme göre ayrı ayrı kümelenmiştir. Son olarak, elde edilen bulgular karşılaştırılmış ve ileriki zamanlarda yapılabilecek çalışmalar için önerilerde bulunulmuştur.

II. BEŞERİ SERMAYE KAVRAMI VE ÇALIŞMADA KULLANILAN DEĞİŞKENLER

Beşeri sermaye, iktisat bilimi açısından oldukça önemli bir kavram olup geçmişi 1776 yılında Adam Smith tarafından ortaya atılan teoriye dayanmaktadır. Adam Smith, beşeri sermaye kavramını becerili insan gücü olarak tanımlamıştır [10]. Teorik anlamda iktisadi açıdan beşeri sermaye, bireylerin üretim süreci ile ilgili sahip oldukları bilgi, beceri, yetenek, bağlılık, davranış ve değerlerin ulaştığı düzey olarak ifade edilmektedir. Diğer bir tanıma göre beşeri sermaye, emeğin sahip olduğu nitelikler olarak tanımlanmıştır [12]

Makroekonomik açıdan ülkelerin sahip oldukları beşeri sermaye düzeyleri, sosyal ve ekonomik kriterler ile ölçülmektedir. İnsani gelişmişlik endeksi de denilen bu ölçümde sosyal kriterler; eğitim ve sağlık iken, ekonomik kriter ise refah standardı olarak değerlendirilmiştir. Endeks, ilk defa 1990 yılında Birleşmiş Milletler Kalkınma Programı tarafından hesaplanmıştır. Bunun yanı sıra Dünya Bankası tarafından 2017 yılında Beşeri Kalkınma Endeksi adıyla yeni bir endeks daha açıklanmıştır. Buna göre ülkeler; cinsiyete, öğrenme düzey ve durumlarına ve sağlıklı olarak hayatta kalma oranına göre karşılaştırılmıştır. Bu çalışmada ise ülkelerin beşeri sermaye durumlarını incelemek amacıyla kullanılan değişkenler ve tanımları tablo 1'deki gibidir. Değişkenlerin belirlenmesinde özellikle beşeri sermaye ve insani gelişmişlik endekslerinin ölçümünde kullanılan göstergeler dikkate alınmıştır.

Tablo 1. *Çalışmada Kullanılan Değişkenler ve Tanımları*

Kısaltma	Değişken	Tanımı
V1	Beklenen Okul Süresi	4 ile 17 yaş arası nüfusun okullaşma oranını ifade etmektedir.
V2	Karma Test Skorları	Eğitim kalitesi ölçümü olarak kullanılmaktadır.
V3	Öğrenmeye dayalı okul süresi	Test skorları ile beklenen okullaşma sürecinin birlikte değerlendirilmesi ile hesaplanmaktadır.
V4	Yetişkin Hayatta Kalma Oranı	15 ile 60 yaş arası nüfusun hayatta kalma oranını ifade etmektedir.
V5	Ortalama okullaşma süresi	Bir bireyin ömrü içerisinde ortalama okula gitme süresini ifade etmektedir.
V6	Kişi başı düşen GSYH	Ülkenin toplam üretim değerinin nüfusa oranıdır.
V7	Beşeri sermaye endeksi	Eğitim, sağlık ve gelir değişkenlerine göre hesaplanan endekstir.
V8	Beşeri Kalkınma Endeksi	Ülkelerin nitelikli insan gücü varlığını gösteren bir endekstir.

Tablo 1'de görüldüğü üzere ülkelerin beşeri sermaye durumları sekiz gösterge ile incelenmiştir. Burada kullanılan değişkenlerin ülkelerin nitelikli insan gücü ölçümünde etkili oldukları düşünülmektedir. Aynı zamanda insani gelişmişlik veya beşeri sermaye endeksi gibi uluslararası geçerliliği olan ölçümlerde ülkelerin karşılaştırılması amacıyla da kullanılmaktadır.

III. YÖNTEM

Daha önce belirtildiği üzere kümeleme analizi yöntemleri; hiyerarşik olup olmama durumuna göre iki sınıfa ayrılmaktadır. Hiyerarşik kümeleme yöntemine göre, bütün gözlemler bir kümede toplanır, daha sonra en aykırı gözlemden başlayarak birer birer farklı kümelere yerleştirilir. Hiyerarşik olmayan kümeleme analizinde ise yöntem; ardışık eşik, paralel eşik ve optimum eşik olmak üzere üç şekilde uygulanabilir. Hiyerarşik olmayan kümeleme yöntemlerinden en bilineni, k-ortalamalar analizidir. Bu bölümde ise çalışmada karşılaştırma amaçlı kullanılan klasik, bulanık ve torbalı k-ortalamalar yöntemlerinin teorik bilgilerine yer verilmiştir.

A. KLASİK K-ORTALAMALAR YÖNTEMİ

Klasik k-ortalamalar, hiyerarşik olmayan kümeleme yöntemleri içerisindedir. Buna göre, başlangıçta belli olan küme sayısına göre gözlemler kümeye yerleştirilmektedir. Bu aşamada kümeleme işlemi, küme ortalaması ile o kümeye ait gözlem değerleri arasındaki uzaklığın minimize edilmesi mantığı ile yapılmaktadır. Literatürde klasik k-ortalamalar yönteminin kullanıldığı birçok çalışma bulunmaktadır. Örneğin, Türkiye'nin dış ticaretteki durumu[14]; üniversite öğrencilerinin başarı durumları [16]; OECD ülkeleri k-ortalama yöntemi ile ekonomik özgürlüklerine göre [17] kümelemiştir.
Yöntemin uygulama aşamaları ise şu şekildedir [13];

1. Küme sayısı belirlenir,
2. Küme sayısı kadar rasgele küme merkezleri belirlenir,
3. Gözlemleri en yakın merkeze göre kümeler,
4. Her küme için ortalama merkez değeri hesaplanır,
5. Gözlemler ile küme ortalamalarına yakınlık gözden geçirilip, kümelere nihai atama yapılır.

Merkez ile gözlem arasındaki uzaklığın ölçümü için Öklid, Mahalanobis, Manhattan gibi farklı yöntemler kullanılabilir. Bu çalışmada ise Öklid tipi uzaklık fonksiyonu kullanılmış olup, denklem aşağıdaki gibidir.

$$d(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \qquad (1)$$

K-ortalamalar yönteminde temel amaç; (1) numaralı hata fonksiyonunun minimum hale getirilmesidir. Her iterasyonda bu değeri azalacağı beklenmektedir. Sonuç olarak, en küçük hatanın elde edildiği düzeyde nihai kümeleme ataması yapılmaktadır [11].

B. BULANIK K ORTALAMALAR YÖNTEMİ

Bulanık K-Ortalamalar yönteminde gözlemlerin kümeye aidiyeti klasik mantık yerine bulanık mantıkla belirlenmektedir. Klasik mantığa göre gözlemler bir kümeye aittir veya değildir diye

düşünülmektedir. Ancak bulanık mantığa göre gözlemlerin her küme için üye olma derecesi hesaplanmaktadır. Burada en yüksek üyelik derecesi hangi kümeye ait ise, gözlemler o kümeye atanır[18]. Literatürde bulanık k-ortalamalar yönteminin kullanıldığı birçok çalışma bulunmaktadır. Örneğin; illerin hayvancılık yetiştiriciliğine göre kümelenmesi [20], göç durumuna göre kümeleme[21], ülkelerin benzerliklerinin tespitinde[22] ve banka müşterilerinin gruplandırılması[23] gibi konularda bulanık k-ortalamalar yöntemi kullanıldığı görülmektedir.

Yöntemin genel kullanımı, Kaufman ve Rousseuw tarafından geliştirilen Fanny algoritması iledir. Aynı zamanda gözlemlerin kümelere olan üyelik derecelerinin toplamının 1 olması gerektiğinden bazı kısıtlar da bulunmaktadır. Buna göre algoritmanın amaç fonksiyonu ve kısıtları aşağıdaki gibidir [19];

Amaç Fonksiyonu: $C = \sum_{v=1}^{K} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2\sum_{j=1}^{n} u_{jv}^2}$ (2)

Kısıtlar: $u_{iv} \geq 0$ ve $\sum_{i=1}^{K} u_{iv} = 1$ (3)

*d(i,j):*i.ve j. birim arasındaki uzaklık
K: Toplam küme sayısı
v: Küme sayısı
u_{iv}: i. birimin v kümesine olan bilinmeyen üyeliği
u_{jv}: j. birimin v kümesine olan bilinmeyen üyeliği
n: Toplam birim sayısı

Yukarıda belirtilen (2) numaralı amaç fonksiyonunun minimum olması arzu edilen bir durumdur. Bunun için yinelemeli bir algoritma ile amaç fonksiyonu minimize edilerek katsayılar matrisine ulaşılır. Bu durumda, her bir gözlemin dört kümeye ait olma düzeyleri hesaplanır. Daha sonra bunlardan en yüksek değere sahip olana göre nihai küme belirlenir.

Yapılan kümeleme işleminin ne kadar güvenilir veya geçerli olduğunun tespit edilebilmesi için Silhouette istatistiği kullanılmaktadır[19]. Silhouette Gölge İstatistiği de denilen tekniğinde, gözlemlerin küme içi ve kümeler arası uzaklıkları karşılaştırılmaktadır. Bu değer, -1 ile +1 arasında olup, +1'e yakın olması durumunda gözlemlerin doğru kümelendiği yorumu yapılır. Bir veri setindeki kümeleme yapısının 'uygun' olarak nitelendirilebilmesi için ortalama gölge istatistiği değerlerinin 0,50 olması beklenmektedir [26].

C. TORBALI K-ORTALAMALAR YÖNTEMİ

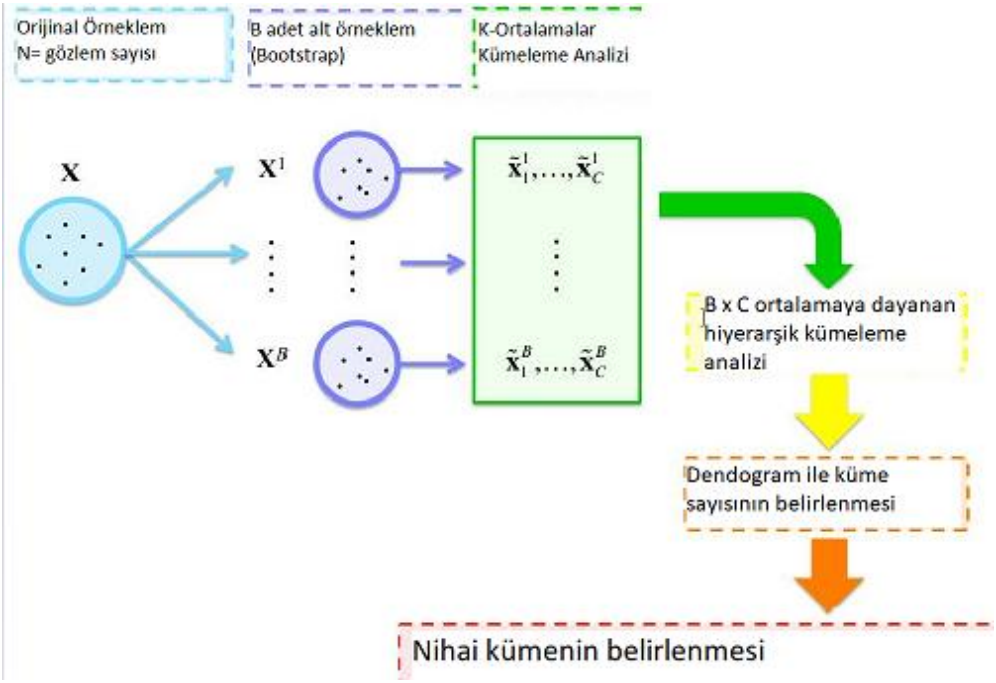
Torbalı k-ortalamalar yönteminin İngilizce karşılığı "bagged cluster" olup, açılımı "bootstrap aggregating"dir. Yöntem, hiyerarşik ve hiyerarşik olmayan yöntemlerin birleşiminden oluşan hibrit bir sisteme sahiptir. İlk defa 1999 yılında Leisch tarafından geliştirilmiştir. Literatürde torbalı k-ortalamalar yönteminin birçok alana uygulandığı görülmektedir. Örneğin, bazı çalışmalarda kümeleme yöntemi kullanarak pazar bölümlendirmesi yapmışlardır [24-25].

Yöntemin uygulama aşamaları şu şekildedir[24];

1. N birimden oluşan B adet bootstrap örneklem oluşturulur.
2. Her bir alt örneklem içerisinde kümeleme analizi uygulanır. Bu çalışmada alt örneklemlerde k-ortalamalar analizi uygulanacaktır. Burada C küme sayısı kadar ortalama belirlenir ve gözlem içerisinde kümeleme analizi yapılır.
3. B adet örneklem içerisinde C adet ortalama değer belirlenir ve bunlar matris hale getirilir.
4. Bir önceki adımda oluşturulan matris kullanılarak hiyerarşik kümeleme analizi uygulanır ve dendogram çizilerek en iyi küme sayısına göre sınıflama yapılır.

5. Bu aşamaya kadar yapılan uygulamalar neticesinde her bir gözlemin her bir kümenin merkezine olan yakınlığa göre belirlenir ve nihai sınıflama bu şekilde yapılır.

Bu aşamaların görsel olarak ifadesi şu şekildedir;



Şekil 2. *Torbalı Kümeleme Analizi Uygulama Aşamaları [25]*

Alt örneklem sayısı yöntemi uygulayan kişiye göre değişmekte olup, genellikle 10 parçaya ayrılmaktadır. K-ortalamalar yönteminin uygulamasında uzaklık fonksiyonu olarak klasik k-ortalamalarda olduğu gibi farklı çeşitler bulunmaktadır. Klasik k-ortalamalar yönteminden temel farkı ise, örneklemi daha küçük parçalara ayırması ve her biri için kendi içerisinde analiz edebilmesidir.

IV. UYGULAMA

A. *ÇALIŞMANIN AMACI VE KAPSAMI*

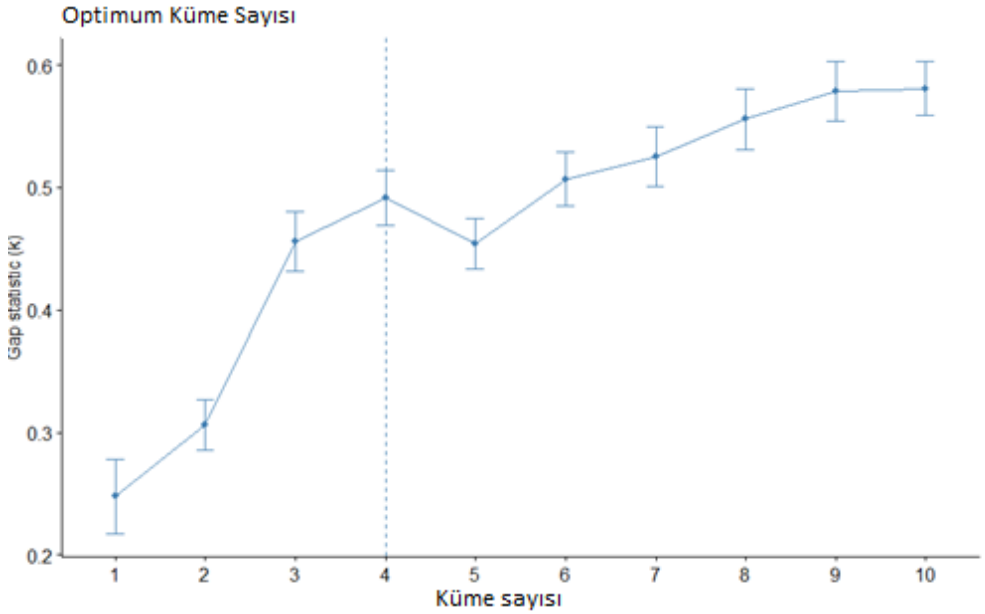
Bu çalışmada iktisadi ve istatistiki açıdan iki tür amacı bulunmaktadır. İktisadi açıdan; beşeri sermaye özelliklerine göre birbirine benzer olan ülkelerin ve nitelik anlamında ayırt edici özelliklerin tespit edilmesidir. İstatistiki açıdan ise, üç farklı k-ortalama algoritmalarının (klasik, bulanık ve torbalı) karşılaştırılması amaçlanmaktadır. Bu doğrultuda, verileri Dünya Bankası web sitesinde paylaşıma açık olan 132 ülke, beşeri sermaye göstergelerine göre klasik, bulanık ve torbalı K ortalamalar algoritmalarına göre kümelenmiştir. Daha sonra, algoritmalar elde edilen bulgulara göre karşılaştırılmıştır. Ülkelerin beşeri sermaye durumları ise 2017 yılı ve tablo1'de yer alan değişkenlere göre değerlendirilmiştir. Değişkenlerin seçiminde ise Birleşmiş Milletler tarafından yayımlanan insani gelişmişlik endeksi ve beşeri sermaye endeksi göstergeleri kullanılmıştır.

B. *OPTİMUM KÜME SAYISININ BELİRLENMESİ*

Hiyerarşik olmayan kümeleme analizi, veri setindeki gözlemleri benzerliklerine göre ve belirli küme sayısı doğrultusunda gruplandırma veya kümelere ayırma işlemlerini içermektedir. Buna göre benzerliklerine göre gözlemlerin kaç grupta toplanacağının tespit edilmesi de analizin güvenirliği açısından büyük önem arz etmektedir. Literatürde bu problem ile alakalı farklı yaklaşımlar bulunmaktadır. Buna göre grup sayısı, doğrudan ya da istatistiksel analiz ile belirlenebilir. Doğrudan yöntemler, küme içi uzaklıkların toplamının veya ortalamasının optimizasyonu; istatistiki yöntemler ise, yokluk hipotezine karşı alternatif hipotezlerin test edilmesine dayanmaktadır. Bu yöntemler; average silhouette, elbow ve gap metodu olarak adlandırılmaktadır. Bu çalışmada ise optimum küme sayısı gap istatistik yöntemine göre incelenmiştir. Gap istatistiği, küme içi gözlem değerlerinin merkez noktalar arasındaki mesafenin ölçümü ile hesaplanmaktadır. Buna göre, küme sayısı arttıkça bu değerin düşüş yaşadığı nokta optimum küme sayısı olarak ifade edilmektedir. Gap yöntemi matematiksel olarak aşağıdaki gibi ifade edilmektedir.

$$Gap_n^{(k)} = E_n(logW_k) - logW_k \qquad (4)$$

Denklemde k küme sayısını, W ise küme ortalaması ile gözlemler arasındaki mesafeyi temsil etmektedir. E ise tahmini uzaklık değerini ifade etmektedir. Burada temel amaç, noktaların tekdüze dağılımından uzak olduğu optimum küme sayısının bulunmasıdır[27]. Bu durumda elde edilen sonuç şekil 3'deki gibidir.



Şekil 3. *Optimum Küme Sayısının Belirlenmesi*

Şekil 3'e göre gap istatistik değerinin düşüş yaşadığı küme sayısı 4 olması nedeniyle optimum küme sayısının 4 olması gerektiği sonucuna ulaşılmıştır. Uluslararası düzeyde BM tarafından hesaplanan insani gelişmişlik endeksi, iş yapma endeksi gibi karşılaştırmalarda ülkeler genellikle 4 gruba ayrılmaktadır. Bu da optimum küme sayısı için elde edilen bulgunun desteklendiğini göstermektedir.

C. KLASİK K-ORTALAMALAR YÖNTEMİNE GÖRE UYGULAMA

Klasik k-ortalamalar yönteminin uygulanması ile elde edilen kümelerin merkezi değerleri tablo 2'deki gibidir.

Tablo 2. *K-Ortalamalar Yöntemine Göre Küme Merkez Değerleri*

	Değişkenler								
Küme	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**Niteleme**
1	0.353249	-0.13239	0.046172	0.104538	0.06362	0.070598	0.410097	0.327761	Yüksek
2	-0.44849	-0.85246	-0.7338	-0.68024	-0.70675	-0.72887	-0.25928	-0.29288	Düşük
3	0.902295	1.225888	1.18862	1.12541	1.142425	1.169321	0.788868	0.76916	Çok yüksek
4	-1.55394	-1.05413	-1.34099	-1.36747	-1.31526	-1.34995	-1.60287	-1.42978	Çok düşük

Tablo 2'de görüldüğü üzere ülkeler 4 kümeye ayrılmış olup, aynı zamanda kümelerin, merkezi değerlerine göre, nitelikleri de belirlenmiştir. Buna göre, beşeri sermaye açısından; 1 numaralı küme yüksek, 2 numaralı küme düşük, 3 numaralı küme çok yüksek ve 4 numaralı küme çok düşük olarak değerlendirilmiştir. Aynı zamanda ülkelerin hangi grupta olduğuna dair bulgulara, metin içerisinde çok yer kaplamaması için, Ek 1'de yer verilmiştir. Klasik k-ortalama algoritmasına göre, ülkelerden 30'u çok düşük; 23'ü düşük; 42'si yüksek ve 45'i çok yüksek olarak kümelenmiştir. Küme özelliklerine dikkat edildiği takdirde çok düşük beşeri sermayeye sahip ülkelerin çoğunlukla Afrika kıtasında, düşük gelir grubunda; çok yüksek nitelikli beşeri sermayeye sahip ülkelerin ise Avrupa veya Amerika kıtasından ve yüksek gelir grubunda oldukları görülmüştür. Düşük ve yüksek beşeri sermaye kümesindeki ülkelerin ise genellikle gelişmekte olan ülkelerden olduğu görülmektedir.

D. BULANIK K ORTALAMALAR YÖNTEMİNE GÖRE UYGULAMA

Kümeleme analizlerinde uygulanan klasik (Aristotales) mantığına göre ülkelerin kümeye aidiyeti "bir kümeye aittir veya değildir" gibi keskin bir şekilde değerlendirilir. Ancak, bulanık k-ortalamalar yöntemine göre ise gözlemin ait olduğu küme [0,1] arasında hesaplanan üyelik düzeyi ile belirlenir. Buna göre, bulanık k-ortalamalar yönteminde her bir gözlem için kümelere ait olma düzeylerini ifade eden üyelik dereceleri hesaplanmaktadır. Bu değerlerden en yüksek üyelik düzeyi hangi kümeye ait ise gözlemin o kümede olduğuna karar verilir. Çalışmada, ülkeler 4 kümeye ayrılmış olup ve kümelerin merkezi değerlerine tablo 3'de yer verilmiştir.

Tablo 3. *Bulanık Kümeleme Analizine Göre Küme Merkezleri*

	Değişkenler								
Küme	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**Niteleme**
1	0.439614	0.036664	0.183902	0.215272	0.165649	0.211642	0.484739	0.37723	Yüksek
2	-0.2562	-0.76797	-0.57565	-0.53556	-0.5428	-0.57842	-0.05974	-0.08311	Düşük
3	0.91246	1.241212	1.223915	1.226158	1.287006	1.185615	0.799102	0.835453	Çok yüksek
4	-1.49787	-1.03277	-1.31186	-1.3338	-1.28372	-1.3153	-1.55896	-1.35584	Çok düşük

Tablo 3'e göre bulanık kümeleme algoritmasında merkez değerlerinin k-ortalamalar yöntemindekine benzerlik gösterdiği dikkat çekmektedir. Buna göre 1 numaralı küme yüksek, 2 numaralı küme düşük, 3 numaralı küme çok yüksek ve 4 numaralı küme çok düşük olarak değerlendirilmiştir. Ülkelerin hangi kümede yer aldığını gösteren tablo Ek 1'de olup, 32 ülke çok düşük, 27 ülke düşük, 31 ülke yüksek ve 42 ülke çok yüksek kümede olarak tahmin edilmiştir. Sayısal olarak bakıldığı takdirde, kümelerde bulunan ülke sayıları k-ortalamalar yönteminden azda olsa farklılık göstermektedir.
Kümeleme analizinde kümeleme işleminin kararlılığı ve güvenilirliği Silhouette endeksi ile belirlenmektedir. Endeksin hesaplanma şekli aşağıdaki gibidir.

$$sil(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (5)$$

Denklemde a(i): i.birimin atandığı kümedeki tüm birimlere ortalama uzaklığı, b(i): i.birimin diğer kümelerdeki tüm birimlere ortalama uzaklıklarının minimum değeridir. Bu değer [-1,1] arasında değişmekte olup 0,5'in üzerinde olması arzu edilen bir durumdur [18]. Bu çalışmada ise Ortalama Gölge İstatistik değeri 0,7023 olarak bulunmuştur. Bu ise karar birimlerinin uygun kümelendiğini göstermektedir.

E. TORBALI K ORTALAMALAR YÖNTEMİNE GÖRE UYGULAMA

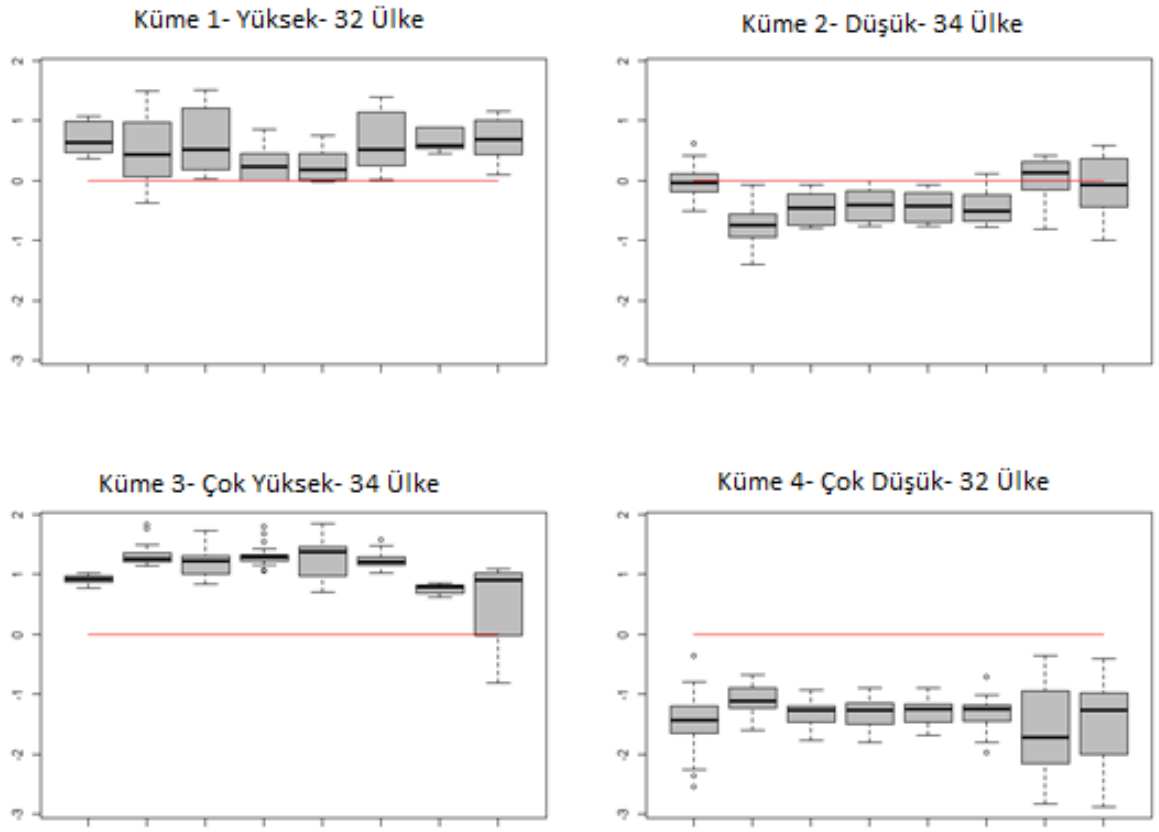
Torbalı K-ortalamalar yöntemi, hiyerarşik ve ayırmalı kümeleme algoritmalarını birlikte içeren özel bir kümeleme tekniğidir. Buna göre, ana örneklem belirli sayıda alt örnekleme ayrılmakta ve kendi içerisinde merkezler belirleyerek gözlemleri kümelere atamaktadır [6]. Bu çalışmada gözlemler, 10 alt örnekleme ayrılmıştır. Ölçüm olarak, Öklid tipi uzaklık fonksiyonu dikkate alınmıştır. Daha sonra her bir alt örneklem için gözlemler, k-ortalamalar algoritmasına göre analiz edilmiştir. Buna göre uygulama neticesinde elde edilen küme merkez değerleri aşağıda yer alan tablo 4'deki gibidir.

Tablo 4. *Torbalı Kümeleme Algoritmasına Göre Hesaplanan Küme Merkez Değerleri*

	Değişkenler								
Küme	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**Niteleme**
1	0.19	-0.32	-0.14	-0.09	-0.11	-0.11	0.19	0.17	Yüksek
2	-1.50	-0.95	-1.26	-1.26	-1.26	-1.29	-1.32	-1.24	Düşük

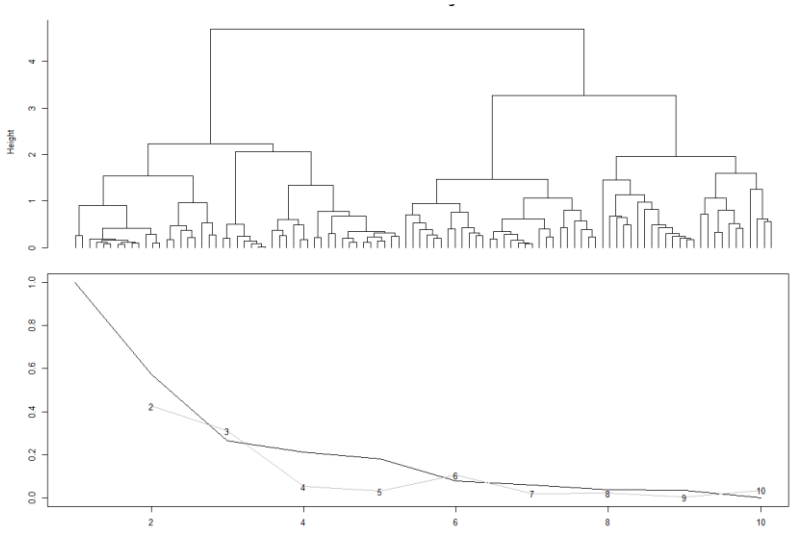
Tablo 4. *Torbalı Kümeleme Algoritmasına Göre Hesaplanan Küme Merkez Değerleri(devam)*

3	0.92	1.30	1.24	1.10	1.07	1.23	0.79	0.67	Çok Yüksek
4	-1.70	-1.34	-1.58	-1.67	-1.58	-1.51	-2.62	-2.75	Çok Düşük



Şekil 4. *Torbalı Kümeleme Yöntemine Göre Oluşan Kümelerin Özellikleri*

Tablo 4'e göre ülkeler, diğer k-ortalama algoritmaları ile benzer şekilde kümelenmiştir. Küme merkezlerine dikkat edildiği takdirde 1 numaralı küme yüksek, 2 numaralı küme düşük, 3 numaralı küme çok yüksek ve 4 numaralı küme çok düşük olarak nitelendirilmiştir. Şekil 4'de ise değişkenlere göre küme özellikleri yer almaktadır. Buna göre, şekilde yer alan düz çizgi ortalamayı, her bir kutu bir değişkeni, dört ayrı şekil ise kümeleri göstermektedir. Buna göre Şekil 4'ün tablo 4'ün görsel hali olduğu ifade edilebilir. Yani, şekil 4'deki küme özellikleri tablo 4 ile benzerlik göstermektedir. Bununla birlikte şekil 4'de görüldüğü üzere 132 ülkeden 32'si çok düşük, 34'ü düşük, 32'si yüksek ve 34'ü çok yüksek kümede olarak tahmin edilmiştir.



Şekil 5. *Torbalı Kümeleme Analizine Göre Küme Sayısı ile İlgili Görsel*

Şekil 5'de yer alan dendogram grafiğinde ülkelerin kümelere göre dağılımları görülmektedir. Alttaki grafiğe göre ise tek bağlantı kümeleme tekniği ile yapılan uygulamanın grafik hali görülmektedir. Tek bağlantı tekniği, uzaklık veya benzerlik matrisinden faydalanılarak birbirine en yakın iki gözlemin birleştirilmesi tekniğidir. Aynı zamanda bu grafiğe dayanarak optimum küme sayısı da belirlenebilmektedir. Dolayısıyla şekil 5'e göre, en keskin düşüş (gri renkli çizgi) küme sayısı 4 olduğunda gerçekleşmiştir, yani ülkelerin 4 kümede incelemenin uygun olduğu görülmektedir. Bu da çalışmada küme sayısının farklı yöntemler ile değerlendirildiğinde tutarlı olduğunu göstermektedir.

V. SONUÇ

Bu çalışmada, istatistiki açıdan; literatürde sıklıkla kullanılan k-ortalama algoritmalarından klasik, bulanık ve torbalı kümeleme yöntemlerinin karşılaştırılması amaçlanmıştır. İktisadi açıdan ise ülkelerin beşeri sermaye niteliklerine göre benzerliklerinin tespit edilmesi ve kümelenmesi amaçlanmıştır. Bu nedenle, 132 ülke sahip olduğu beşeri sermaye özelliklerine göre kümelenmiştir. Çalışmada uygulanan üç tip kümeleme algoritması genel olarak karşılaştırıldığı takdirde; merkezin değerlerindeki farklılıklar nedeniyle bazı ülkelerin ait oldukları kümelerden farklı olan tahmin edildiği görülmüştür. Sonuç olarak, üç algoritmaya göre farklı kümelerde oldukları tahmin edilen ülkeler tablo 5'deki gibidir.

Tablo 5. *Kümeleme Algoritma Tahminlerinin Karşılaştırılması*

Ülkeler	**K-Ortalama**	**Bulanık K Ort.**	**Torbalı K Ort.**
Moritanya, Yunanistan, Lüksemburg, Polonya, Portekiz, Singapur, El Salvador, Honduras, Kuveyt	Çok Yüksek	Çok Yüksek	Yüksek
İsrail	Çok Yüksek	Yüksek	Yüksek
Madagaskar, Hong Kong, Çin, Kore, Norveç	Düşük	Çok Düşük	Çok Düşük

Tablo 5. *Kümeleme Algoritma Tahminlerinin Karşılaştırılması(devam)*

Benin, Hollanda, Sırbistan, Botsvana, Gana, Tonga, Ermenistan, Bahreyn, Brezilya, Nikaragua	**Yüksek**	**Düşük**	**Düşük**
Suudi Arabistan	Yüksek	Yüksek	Çok Yüksek
Kamerun, Mali, Slovak Cumhuriyeti, İsveç, Birleşik Krallık, Lao PDR, Ukrayna	Yüksek	Yüksek	Düşük

Tablo 5'de toplamda 33 ülkenin üç algoritmadan en az birinde farklı sınıflandığı görülmektedir. Buna göre, 17 ülke klasik ve bulanık k-ortalama algoritmalarına göre aynı kümede yer alırken, torbalı k-ortalamalar yöntemine göre ise faklı kümelenmiştir. Benzer şekilde, 16 ülke bulanık ve torbalı k-ortalamalara göre aynı iken klasik k-ortalamaya göre farklı kümelenmiştir.

Tablo 6. *Yöntemlerin Çıktılarına Ait Korelasyon Analizi Sonucu*

	K-Ortalamalar	Bulanık K Ort	Bagged K Ort
K-Ortalamalar	1.000		
Bulanık K Ort	.917	1.000	
Bagged K Ort	.883	.902	1.000

Tablo 6'da görüldüğü üzere kümeleme analizinde algoritma farklılığının nihai kümelerde etkili olduğu görülmektedir. Ancak genel olarak, sağladıkları küme çıktıları açısından birbiri ile yüksek benzerlik gösterdiği ifade edilebilir. Bununla birlikte, klasik K-ortalamalar ile bulanık k-ortalamalar algoritmalarının birbirine daha çok benzediği söylenebilir.

Bununla birlikte, ülkeler ait oldukları kümelere göre her yöntem için çoklu lojistik regresyon analizi uygulanmıştır. Bu aşamada bağımlı değişken olarak ülkelerin ait oldukları kümeler, bağımsız değişken olarak ise beşeri sermaye ölçümleri alınmıştır. Analiz sonucunda; ülkelerin klasik k-ortalamalar ile %95, bulanık k-ortalamalara göre %97 ve torbalı k-ortalamalara göre ise %99 oranında doğru sınıflandırıldığı sonucuna ulaşılmıştır. Yöntemlerin uygulama süreleri karşılaştırılacak olursa; klasik ve bulanık k-ortalama algoritmaları genellikle 10 saniyeden daha kısa sürede sonuç üretebilmektedir. Ancak torbalı k-ortalama tekniğinde alt örneklem oluşturma işlemi yer alması nedeniyle süreç yaklaşık 1 dakikaya yakın olmaktadır. Bu durumda torbalı algoritmanın diğer yöntemlerden daha yavaş çalıştığı söylenebilir. İleriki zamanlarda yapılabilecek çalışmalarda iktisadi açıdan, ülkelerin farklı özelliklerine göre; istatistiki açıdan ise k-medoid gibi farklı algoritmalar ile kümelenmesi önerilebilir.

VI. KAYNAKLAR

[1] Koyuncugil, A.,ve Özgülbaş, N. "Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları," *Bilişim Teknolojileri Dergisi*, c.2, s.2, ss. 21-33, 2009.

[2] Dudoit, S., & Fridlyand, J. "A Prediction-Based Resampling Method For Estimating The Number Of Clusters in a Dataset," *Genome biology*, Vol. 3, No. 7, 2002

[3] MacQueen, "Some methods for classification and analysis of multivariate observations," The fifth Berkeley symposium on mathematical statistics and probability, vol.1, no. 14, pp. 281-29, 1967.

[4] Topchy, A., Minaei-Bidgoli, B., Jain, A. K., & Punch, W. F. "Adaptive clustering ensembles," In Proceedings of the 17th International Conference on Pattern Recognition, vol. 1, pp. 272-275, 2004

[5] Dunn, J. C. "A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters," *Journal of Cybernet*, vol. 3, no. 3, pp. 32–57, 1974

[6] Leisch, F., and Hornik, K. (1999). "Stabilization of k-means with bagged clustering," Joint Statistical Meetings, Statistical Computing Section, pp. 174-179, 1999

[7] Abbas O. "Comparisons Between Data Clustering Algorithms," *The International Arab Journal of Information Technology*, vol. 5 no.3 ss.320-235, 2008

[8] Sarıman G. "Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoid Kümeleme Algoritmalarının Karşılaştırılması," *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, c.15, s.3, ss.192-202, 2011

[9] Greene, D., Tsymbal, A., Bolshakova, N. and Cunningham, P. "Ensemble clustering in medical diagnostics," 17th IEEE Symposium on Computer-Based Medical Systems, Jyväskylä, Finland, pp. 576-581, 2004

[10] Graff Zivin, J., Hsiang, S. M., and Neidell, M. " Temperature and human capital in the short and long run," *Journal of the Association of Environmental and Resource Economists*, vol. 5, no.1, 77-10, 2018

[11] Wang, T. and Zatzick, C. D. "Human Capital Acquisition and Organizational Innovation: A Temporal Perspective," *Academy of Management Journal*, vol.62, no.1, pp.99-116, 2019

[12] Weller, I., Hymer, C. B., Nyberg, A. J. and Ebert, J. "How matching creates value: Cogs and wheels for human capital resources research," *Academy of Management Annals*, vol.13, no.1, pp.188-214, 2019

[13] Liu, X., Zhu, X., Li, M., Wang, L., Zhu, E., Liu, T., ... and Gao, W."Multiple kernel k-means with incomplete kernels," Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence IEEE transactions on pattern analysis and machine intelligence, USA, 2019

[14] Tunalı D. ve Aytekin D. "Türkiye Dış Ticaretinin Kümeleme Analizi ile İncelenmesi" *Eskişehir Osmangazi Üniversitesi İİBF Dergisi*, c. 12, s.3, ss:103-116, 2017

[15] Wang, P., Shi, H., Yang, X. and Mi, J. "Three-way k-means: integrating k-means and three-way decision". *International Journal of Machine Learning and Cybernetics*, pp.1-11, DOI: https://doi.org/10.1007/s13042-018-0901-y, 2019

[16] Ünal Y. Ekim U.ve Köklü M. "Üniversite Öğrencilerin Ortak Zorunlu Derslerdeki Başarılarının K-Means Algoritması ile İncelenmesi," *e-Journal of New World Sciences Academy*, c.6, s.1, ss.342-347, 2011

[17] Kangallı, S. G., Uyar, U. ve Buyrukoğlu, S. "OECD Ülkelerinde Ekonomik Özgürlük: Bir Kümeleme Analizi," *Journal of Alanya Faculty of Business/Alanya İşletme Fakültesi Dergisi*, c. 6, s.3, ss. 95-109, 2014

[18] Yılancı V. "Bulanık Kümeleme Analizi İle Türkiye'deki İllerin Sosyoekonomik Açıdan Sınıflandırılması" *Süleyman Demirel Üniversitesi İİBF Dergisi*, c.15, s.3, ss. 453-470, 2010

[19] Giray, S. "Ülkelerin Turizm İstatistikleri Bakımından Farklı Kümeleme Analizi Metotları ile Sınıflandırılması ve Türkiye'nin Bu Oluşumdaki Yeri," International Conference on Eurasian Economies, pp. 17-18, 2013

[20] Kılıç, İ., Lenger, Ö.F. ve Bozkurt, Z., "Bulanık Kümeleme Analizi ile Türkiye'deki İllerin Hayvancılık İstatistikleri Bakımından Sınıflandırılması," *Kocatepe Veteriner Dergisi*, c. 5, s.1,ss.21-28, 2012

[21] Sönmez, H., Er, F., "Türkiye' de İllere Göre İç Göç Hareketlerinin Modern Kümeleme Teknikleri ile İncelenmesi," *Eskişehir Osmangazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, c.20, s.1, ss.141-160, 2012

[22] Şahin, M., Hamarat, B., "G10-Avrupa Birliği ve OECD Ülkelerinin Sosyoekonomik Benzerliklerinin Fuzzy Kümeleme Analizi ile Belirlenmesi," ODTÜ Uluslararası Ekonomi Kongresi, c.6, ss.11-14, 2002

[23] Yıldız, Z., "Banka Müşterilerinin Demografik ve Sosyo-Ekonomik Özellikler Bakımından Gruplandırılmasında Kümeleme Çözümlemesi ve Bir Uygulama" Yüksek lisans tezi, Anadolu Üniversitesi, Türkiye, 2002

[24] Prayag G. Disegna M., Cohen S. A., Yan H., "Segmenting markets by bagged clustering: Young Chinese travelers to Western Europe, *Journal of Travel Research*, Vol.54, no.2, pp.234-250, 2015

[25] D'Urso, P., De Giovanni, L., Disegna, M., and Massari, R. " Bagged clustering and its application to tourism market segmentation". Expert Systems with Applications, Vol.12, pp.4944-4956, 2013

[26] Kılıç, I., and Özbeyaz, C. "Classification of Karayaka and Bafra (Chios x Karayaka B1) sheep according to body measurements by different clustering methods". *Ankara Üniv. Vet. Fak. Derg*, Vol.58, pp.203-208, 2011

[27] Cebezi Z., Yıldız F., Kayaalp T., "K-Ortalamalar Kümelemesinde Optimum K Değeri Seçilmesi" 2.Ulusal Yönetim Bilişim Sistemleri Kongresi, Erzurum, Bildiriler Kitabı, ss:231-242, 2015.