HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı

Eğitimde Ölçme ve Değerlendirme Programı

ULUSLARARASI ÖĞRENCİ DEĞERLENDİRME PROGRAMI 2015 VERİLERİNİN VERİ MADENCİLİĞİNDE KÜMELEME YÖNTEMLERİYLE İNCELENMESİ

Mehmet Taha ESER

Doktora Tezi

Ankara, 2019

Liderlik, araştırma, inovasyon, kaliteli eğitim ve değişim ile

Daha ileriye... En İyiye...

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı

Eğitimde Ölçme ve Değerlendirme Programı

ULUSLARARASI ÖĞRENCİ DEĞERLENDİRME PROGRAMI 2015 VERİLERİNİN VERİ MADENCİLİĞİNDE KÜMELEME YÖNTEMLERİYLE İNCELENMESİ

EXAMINATION OF THE PROGRAM FOR INTERNATIONAL STUDENT ASSESSMENT 2015 DATA BY CLUSTERING METHODS IN DATA MINING

Mehmet Taha ESER

Doktora Tezi

Ankara, 2019

Kabul ve Onay

Eğitim Bilimleri Enstitüsü Müdürlüğüne,

Mehmet Taha Eser'in hazırladığı "Uluslararası Öğrenci Değerlendirme Programı 2015 Verilerinin Veri Madenciliğinde Kümeleme Yöntemleriyle İncelenmesi " başlıklı bu çalışma jürimiz tarafından **Eğitim Bilimleri Ana Bilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalında Doktora Tezi** olarak kabul edilmiştir.

Jüri Başkanı	Prof. Dr., Nuri DOĞAN	İmza
Jüri Üyesi (Danışman)	Dr. Öğr. Üyesi, Derya ÇOBANOĞLU AKTAN	İmza
Jüri Üyesi	Prof. Dr., Neşe GÜLER	İmza
Jüri Üyesi	Doç. Dr., Burcu ATAR	İmza
Jüri Üyesi	Doç. Dr., Hakan ATAR	İmza
İkinci Tez Danışmanı	Prof. Dr., Cem OKTAY GÜZELLER	Enstitü Yönetim Kurulunun 06/12/2016 Tarihli ve 2016-47/02 sayılı kararı.

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim, Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca yukarıdaki jüri üyeleri tarafından 20/06/2019 tarihinde uygun görülmüş ve Enstitü Yönetim Kurulunca ..... / ..... / ..... tarihinde kabul edilmiştir.

Prof. Dr. Ali Ekber ŞAHİN

Eğitim Bilimleri Enstitüsü Müdürü

i

Öz

Bu çalışmada veri madenciliğine dayalı kümeleme yöntemlerinden Kohonen'in Öz Örgütlemeli Harita Yöntemi, K-Ortalamalar ve İki Aşamalı Kümeleme Yöntemi yardımıyla PISA verilerine dayalı olarak ele alınan değişkenlere göre elde edilen sonuçların incelenmesi amaçlanmıştır. Bu amaç kapsamında PISA 2015 öğrenci anketinde yer alan fen bilgisi öğretimine ilişkin alt boyutlar ile olası fen başarı puan ortalaması girdi olarak kullanıldığında öğrencilerin farklı yöntemlere göre kaç kümeye ayrıştığı, her bir kümenin nasıl tanımlandığı ve bu kümelere ayrışmada etkili olan değişkenlerin belirlenmesi hedeflenmiştir. Çalışma kapsamında Slovenya hariç OECD üyesi ülkelerinin öğrencilerine sistematik örnekleme uygulanmış ve sonuç olarak 9870 öğrenci üzerinden analizler gerçekleştirilmiştir. Çalışmada kullanılan girdi değişkeni sayısı, fen bilgisi öğretiminin dört alt boyutuna ilişkin faktör puanları ortalaması ve olası fen başarı puanları ortalaması olmak üzere beş olarak belirlenmiştir. Çalışma sonucunda Kohonen ve K-Ortalamalar Yöntemleriyle belirlenen ideal küme sayısının dört; İki Aşamalı Kümeleme Yöntemiyle belirlenen ideal küme sayısının ise iki olduğu belirlenmiştir. Kohonen'in Öz Örgütlemeli Harita Yöntemi ile sorgulama temelli fen bilgisi öğretimi; K-Ortalamalar Yöntemi ile ise öğretmen merkezli fen bilgisi öğretimi olmak üzere iki yöntem kapsamında en başarılı öğrencilerin yer aldığı kümelerin oluşmasında farklı değişkenlerin etkili olduğu belirlenmiştir. Çalışma sonucunda genel anlamda Kohonen'in Öz Örgütlemeli Harita ve K-Ortalamalar Yöntemlerinden elde edilen sonuçların benzerlik gösterirken, İki Aşamalı Kümeleme Analizinden elde edilen sonuçların farklılık gösterdiği belirlenmiştir. Çalışma sonucunda kümeleme analizinde araştırmacıların farklı yöntemlerle elde ettikleri sonuçları rapor etmeleri önerilmektedir. Aynı zamanda çalışma sonucunda zengin çıktılar elde edilebilmesi sebebiyle kümeleme analizlerinin R programı kullanılarak yapılması önerilmiştir.

Anahtar sözcükler: veri madenciliği, kümeleme, öz örgütlemeli harita, k-ortalamalar, iki aşamalı kümeleme, r.

Abstract

In this study, it is aimed to investigate the results obtained from the methods based on PISA 2015 data with the help of Self-Organizing Map, K-Means and Two-Stage Clustering Method. For this purpose, it is aimed to determine how many clusters of students are divided according to different methods, how each cluster is defined and the variables that are effective in these clusters. In the scope of the study, systematic sampling was applied to the students of OECD member countries and as a result analyzes were conducted on 9870 students. The number of input variables used in the study was determined to be five, namely the average of factor scores for four science teaching sub-dimensions and the average of ten plausible values in science . As a result of the study, the ideal number of clusters determined by Self-Organizing Map and K-Means Methods is four and the ideal number of clusters determined by two-stage clustering method was determined to be two. It was determined that different variables were effective in the formation of clusters with the most successful students within the scope of two methods. As a result of the study, it was found that the results obtained from Self-Organizing Map and K-Means Methods were similar in general. As a result of the study, it is recommended that researchers report the results obtained by different methods in clustering analysis. At the same time, clustering analysis was proposed by using R program because of rich results.

Keywords: data mining, clustering, self organizing map, k-means, two-step cluster, r.

Teşekkür

Ölçme ve değerlendirme bilim dalında eğitime başladığım ilk günden bu yana bilgisini ve emeğini esirgemeden önerileri ve fikirleriyle beni yönlendiren, tez sürecinin tüm aşamalarında yanımda olan ve kafama takılan her soru için çekinmeden kapısını çalabildiğim, her koşulda anlayışlı tavrı ve yardımseverliği için değerli danışmanım Dr. Öğr. Üyesi Derya ÇOBANOĞLU AKTAN'a içten teşekkürlerimi sunarım.

Doktora öğrenimim boyunca her zaman desteklerini hissettiğim ayrıca görüş ve önerileriyle de bu çalışmaya katkıda bulunan ve aynı zamanda ikinci danışmanın olan değerli hocam Prof. Dr. Cem Oktay GÜZELLER'e sonsuz teşekkürlerimi sunarım.

Ölçme ve değerlendirmeye alanında eğitime başladığım ilk günden bu yana bizlerden bilgi, destek ve emeklerini esirgemeyen doktora eğitimim boyunca aldığım dersler aracılığıyla bilgisinden ve deneyiminden faydalandığım değerli hocalarım Prof. Dr. Selahattin GELBAL'a, Prof. Dr. Hülya KELECİOĞLU'na ve Prof. Dr. Nuri DOĞAN'a teşekkür ederim.

Tezimi tamamlayabilmem için her zaman bana destek olan değerli arkadaşım Gökhan AKSU'ya sonsuz teşekkürlerimi sunarım.

İçindekiler

Tablolar Dizini

Şekiller Dizini

Simgeler ve Kısaltmalar Dizini

ABK: Akaike Bilgi Kriteri

BBK: Bayes Bilgi Kriteri

EVM: Eğitsel Veri Madenciliği

OECD: The Organization for Economic Cooperation and Development

Bölüm 1
Giriş

Bu bölümde; araştırmanın temelini oluşturan problem durumu, araştırmanın amacı ve önemi, problem cümlesi, alt problemler ve sınırlılıklar yer almaktadır.

Problem Durumu

“Bilgi çağında yaşıyoruz” popüler bir deyim; ancak, veri çağı içerisinde yaşamaktayız. Terabayt veya petabayt (1024 terabayt) gibi birimlerle ölçülebilecek veriler bilgisayar ağımıza, World Wide Web'e (WWW) ve her gün iş, toplum, bilim ve mühendislik, tıp ve günlük hayatın hemen hemen her yönünden çeşitli veri depolama cihazlarına aktarılmaktadır. Mevcut veri hacmindeki bu artış, toplumdaki bilgisayar kullanımının, güçlü veri toplama ve depolama araçlarının hızla gelişmesinin bir sonucudur. Dünya çapında şirketler, satış işlemleri, hisse senedi alım satım kayıtları, ürün tanımları, satış promosyonları, şirket profilleri, performans ve müşteri geribildirimi gibi devasa veri setleri üretmektedir. Bilimsel uygulamalar, uzaktan algılama, süreç ölçümü, bilimsel deneyler, sistem performansı, mühendislik gözlemleri ve çevre gözetimi gibi faktörler aracılığıyla sürekli olarak petabaytlar seviyesinde veri üretilmektedir. Küresel omurga telekomünikasyon ağları her gün onlarca petabyte veri trafiği taşır. Arama motorları tarafından desteklenen milyarlarca web araması her gün petabaytlarca veriyi işlemektedir. Topluluklar ve sosyal medya, dijital fotoğraf ve videolar, bloglar, web toplulukları ve çeşitli sosyal ağlar kısa sürede önemli veri kaynakları haline gelmiştir. Büyük miktarda veri üreten kaynakların listesi her geçen gün artmakta ve bu duruma bağlı olarak bilgisayar teknolojisinde gelişim anlamında büyük sıçramalar gerçekleşmektedir.

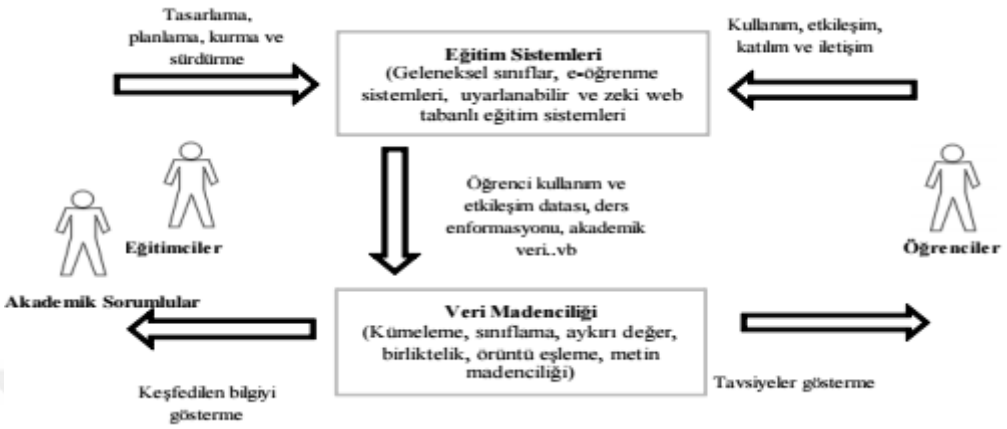
İnternet erişiminin kolaylaşması, bilgisayar teknolojisinin ilerlemesi ve veri depolamaya yardımcı sistemlerin kullanımının artması farklı türde verilerin kaydedilmesi ve verilerin hem çok ucuz hem de çok kolay bir şekilde saklanmasına büyük katkı sağlamıştır. Günümüzde bilgi teknolojileri alanında yaşanan hızlı gelişmeler sonucunda veri ve gizli bilgi konusundaki hızlı artış çok büyük boyutlara ulaşmıştır (Sarıman, 2011; Şentürk, 2006). Hızlı bir şekilde artış gösteren bu çok büyük boyutlardaki verilerin saklanması konusunda senelerdir kullanılmakta olan veritabanları tek başına yeterli olmamaya başlamış ve veri ambarı kavramı ortaya çıkmıştır (Inmon ve Hackathorn, 1994; Oğuzlar, 2004). Söz konusu veriler, ancak

belirli bir amaç doğrultusunda kullanıldığında anlamlı hale gelmeye başlamaktadır (Kalikov, 2006; Koç ve Karabatak, 2012; Özbay, 2015). Verimli bilgileri büyük miktarlarda otomatik bir şekilde ortaya çıkarmak ve bu bilgileri organize bilgiye dönüştürmek için güçlü ve çok yönlü araçlara ihtiyaç vardır. Bu gereklilik, veri madenciliğinin doğmasına yol açmıştır. Depolanmış olan çok büyük veri setlerine ilişkin parametreler arasındaki ilişkinin keşfedilmesi ve gizli örüntünün ortaya çıkarılması noktasında devreye veri madenciliği teknikleri girmektedir. "Veri madenciliği" terimi 1980'li yılları sonunda ortaya çıkmıştır ve verilerdeki *ilginç örüntüleri* çıkarmaya çalışan faaliyetleri tanımlamaktadır. 1980'lerden bu yana veri madenciliği ve bilgi keşfi hem akademik dünyada hem de endüstride en sıcak gündem konularından biri olmuştur. Veri madenciliği, büyük miktardaki tarihsel veriler içinde gizli olan çok değerli ticari ve bilimsel istihbaratın elde edilmesine imkan vermektedir (Jain ve Dubes, 1998; Wu, 2012).

Mühendislik, eğitim, pazarlama, tıp, maliye ve spor gibi alanlarda veri madenciliğine ilişkin pek çok çalışma bulunmaktadır. Veri madenciliği teknikleri karar vericiler için problem çözmede alternatif çözüm sağlama becerisinin belirli alanlarda ortaya çıkmasına yardımcı olmaktadır. Eğitim alanı, veri madenciliği teknikleri açısından düşünüldüğünde çok geniş verilere ulaşılabilecek alanlardandır. Eğitim alanında veri madenciliği teknikleri kullanarak keşif verilerine Eğitsel Veri Madenciliği (EVM) denir. Bu alan eğitim verilerinden gizli bilgileri keşfetmek amacıyla bir örüntünün çıkarılması ile ilgilidir (Ramaswami ve Bhaskaran, 2010; Romero, Ventura, Espejo ve Hervas, 2008; Tair ve El-Halees, 2012).

Eğitsel veri madenciliği. EVM içinde büyük verilerin bulunması nedeniyle analizi çok zor ve imkansız olan büyük boyutlu eğitim verilerine ait büyük koleksiyonlarda örüntüleri tespit etme amacıyla bilgisayarda analiz yöntemleri geliştirme, araştırma ve uygulamayla ilgilidir (Romero ve Ventura, 2013). Aynı zamanda EVM, çok büyük verilerin içerisinde yer alan ve kıymetli olan bilgiyi ortaya çıkartmak ve bu yolla gelecek ile ilgili çıkarımları betimlemeyi sağlayan bağıntı ve kuralların programlamalar aracılığı ile aranmasına yardımcı olan yöntemler bütünüdür (Kayri, 2008). EVM son yıllarda önem kazanan bir araştırma alanıdır ve eğitimde araştırma problemlerinin çözülmesi için bualanlarda ortaya çıkan özgün veri türlerinin analiz edilmesini amaçlar (Baker ve Yacef, 2009). Bu yönüyle EVM önemli eğitim sorularını ele almak için eğitim ortamlarından elde edilen spesifik veri türlerine veri

madenciliği (VM) tekniklerinin uygulanması olarak da tanımlanabilir. Romero ve Ventura (2007), anlam olarak EVM'nin iteratif bir hipotez meydana getirme, test etme, test geliştirme döngüsüne karşılık gelen bir kavram olduğunu belirtmişlerdir. Şekil 1'de eğitimde veri madenciliğine ilişkin döngü yer almaktadır.



Şekil 1. Eğitsel veri madenciliğine ilişkin döngü (Romero ve Ventura, 2007)

EVM, öğrenimi ve öğretimi destekleyen her türlü bilgi sistemince üretilen verileri analiz etmektedir (geleneksel ve modern öğretim formları ve yöntemlerini sağlayan okullar, kolejler, üniversiteler ve diğer akademik veya profesyonel kurumlar ile ilişkili veriler). Bu veriler bir eğitim sitemi içerisinde yer alan öğrencilerin etkileşimleriyle sınırlı değildir (örneğin testlerdeki ve interaktif alıştırmalardaki katkı) fakat işbirliği içinde olunan öğrencilerden gelen verileri (örneğin mesaj ile gerçekleştirilen sohbet), idari verileri (okul, okul bölgesi, öğretmen), demografik verileri (örneğin cinsiyet, yaş, okul sınıfları), öğrenci duyuşsal özelliklerine ilişkin (örneğin motivasyon, duygusal durumlar) veri türlerini içermektedir (Baker ve Yacef, 2009; Merceron ve Yacef, 2007; Witten ve Frank, 2011). Bu veri türleri, farklı yapılara sahip olabilmektedir. Bu verilerde birden çok hiyerarşik düzey (konu, atama, soru düzeyleri) gibi tipik özellikler, bağlam (belirli bir tarihte belirli bir zamanda belirli bir soruyla karşılaşan belirli bir sınıftaki belirli bir öğrenci), ince taneli (farklı analizleri kolaylaştırmak için farklı çözünürlükte verilerin kaydedilmesi, örneğin her 20 saniyede verilerin kaydedilmesi) ve boylamsal (çoğu veriler uzun bir süre boyunca pek çok oturumda kaydedilir, örneğin dönem boyunca veya yıl boyu süren kurslar boyunca) olma gibi tipik özellikler söz konusudur (Romero ve Ventura, 2007; Siemens ve Baker, 2012).

EVM, bilgi getirme, önerici sistemler, görsel veri analitikleri, alan güdümlü veri madenciliği, sosyal ağ analizi, psikopedagoji, bilişsel psikoloji, psikometri vb.

disiplinleri içeren disiplinler arası bir alandır. Evm üç ana alanın kombinasyonu olarak düşünülebilir. Bu üç ana alan, bilgisayar bilimi, eğitim ve istatistiktir. Şekil 2'de bu üç ana alanı ve ana alanlarla ilgili olan yan alanları gösteren ilgili veri madenciliğine ilişkin şema yer almaktadır. Bu üç ana alanın kesişimi; bilgisayar tabanlı eğitim, veri madenciliği ve makine öğrenimi, öğrenme analitikleri (ÖA) gibi EVM ile yakından ilişkili diğer yan alanları meydana getirmektedir.

Şekil 2. Eğitsel veri madenciliği ile ilgili temel ve yan alanlar (Romero ve Ventura, 2013).

Şekil 1'de görülen alanların arasında EVM ile en ilgili olan *Saha Akademi Analitiği* olarak da bilinen *Öğrenme Analitikleri*'dir. Öğrenme analitiği veri güdümlü karar verme üzerine ve Öğrenme analitiklerinin teknik ve sosyal/pedagojik boyutlarının entegre edilmesine odaklanır. EVM kapsamında genel olarak veriler kullanılarak yeni örüntüler keşfedilmesi amacıyla yeni algoritmalar ve/veya modeller geliştirilmesine rağmen, öğrenme analitikleri ile eğitim sistemleri üzerinde bilinen kestirim modelleri uygulanmaktadır. Bu bağlamda öğrenme analitikleri, öğrenmenin ve öğrenmenin meydana geldiği ortamın anlaşılması, optimize edilmesi amacıyla öğrenciler ve öğrencilere ilişkin bağlamlar hakkında verilerin ölçülmesi, toplanması, analiz ve rapor edilmesi olarak tanımlanabilir. Öğrenme analitikleri ve EVM pek çok ortak özelliğe, benzer amaçlara ve çıkarlara sahip olmasına rağmen, her iki kavram arasında önemli farklılıklar da bulunmaktadır. Öğrenme analitiklerinde en çok kullanılan istatistiksel teknikler sosyal ağ analizi (social network analysis), duygu analizi (sentiment analysis), görselleştirme (visualization), söylem çözümlemesi (discourse analysis), kavram analizi ve anlamlandırma modelleri (sense-making models)'dir. EVM

kapsamında en çok kullanılan teknikler sınıflama, kümeleme, Bayezyen modelleme, ilişki madenciliği (relationship mining) ve model yoluyla keşiftir. Öğrenme analitiklerinin kökenini anlamsal ağ (semantic web), zeki program (intelligent curriculum) ve sistemik aracılıklar (systemic interventions) oluştururken, EVM ise eğitimde yazılım, öğrenci modellemesi ve eğitime ilişkin çıktıların öngörülmesi ile ilişkilidir. Öğrenme analitikleri verilerin ve sonuçların açıklanması üzerine odaklanmışken, EVM daha çok veri madenciliği tekniklerinin kullanımının tanımlanmasına ve karşılaştırılmasına odaklanmıştır (Baepler ve Murdoch, 2010; Romero ve Ventura, 2013; Siemens ve Baker, 2012).

EVM ile ilgili birçok teknik bulunmaktadır. Bu teknikler yordama, kümeleme, ilişki madenciliği, model keşfi olarak sınıflandırılmaktadır. Baker (2010), Bienkowski, Feng ve Means (2012) ve daha sonra Romero ve Ventura (2007) bu sınıflamayı genişletmiştir. Baker (2010), Bienkowski, Feng ve Means (2012) ve daha sonra Romero ve Ventura (2007)'nın gerçekleştirdiği sınıflama genişletme çalışmalarını Baker ve Yacef (2009) ve AlMazroui (2013) yordama (prediction), kümeleme (clustering), uç değer tespiti (outlier detecting), ilişki madenciliği (relationship mining), sosyal ağ analizi (social network analysis), süreç analizi (process mining), metin madenciliği (text mining) şeklinde alanyazında genel anlamda kabul görmüş bir forma sokmuşlardır. Aşağıda, bu tekniklere ilişkin kısa açıklamalar yer almaktadır.

Yordama. Amaç, veri noktalarını bir başka ilgili veri değerinin açıklama düzeyine göre tanımlamaktır. Yordamanın gelecekteki olaylarla ilintili olması gerekli değildir, ve kullanılan değişkenler bilinmemektedir. Yordama metotlarının türleri, sınıflandırma (tahmin edilen değişken bir kategorik değer olduğunda), regresyon (tahmin edilen değişken sürekli bir değer olduğunda) ya da yoğunluk tahmini (tahmin edilen değişken olasılık yoğunluk fonksiyonu)'dir. Akademik başarı ve davranışların yordanması, yordama metodu kapsamında EVM'ye örnek olarak gösterilebilir (AlMazroui, 2013; Baker, 2010; Baker ve Yacef, 2009).

Kümeleme. Doğal bir şekilde bir araya toplanmış ve bütün bir veri setini kategorilere ayırmak için kullanılabilen örnekleri bulmayı ifade eder. Bu yöntemde tipik olarak, benzer örneklerin nasıl olduğuna karar vermek için bazı mesafe ölçümleri kullanılır. Kümeler belirlendikten sonra, yeni örnekler en yakın kümelenmeyi belirleyerek sınıflandırılabilir. EVM'de kümeleme genellikle öğrencileri öğrenme

kalıpları veya bilişsel stratejileri temelinde gruplamak için kullanılabilir (AlMazroui, 2013; Baker ve Yacef, 2009; Bienkowski, Feng ve Means, 2012).

İlişki madenciliği. Veri setindeki değişkenler arasındaki ilişkileri bulmak ve bu ilişkilerin daha sonra kullanılmasını sağlamak için kodlanmasını içerir. Birliktelik kuralı madenciliği (değişkenler arasındaki herhangi bir ilişki), sıralı örüntü madenciliği (değişkenler arasındaki zamansal çağrışımlar), korelasyon madenciliği (değişkenler arasındaki doğrusal korelasyonlar), rastgele veri madenciliği (değişkenler arasındaki tesadüfi ilişkiler) bu kapsamda incelenir. EVM kapsamında ilişki madenciliğine örnek olarak öğrencilerin çevrimiçi aktiviteleri ile final notları arasındaki ilişkinin tanımlanması ya da öğrencilerin problem çözme aktivitelerinin modellenmesi örnek gösterilebilir (AlMazroui, 2013; Baker ve Yacef, 2009).

Model keşfi. Verilerin, bir insanın özelliklerini hızlı bir şekilde tanımlamasına veya sınıflandırmasına olanak tanıyacak şekilde tasvir edilmesini sağlayan bir tekniktir. Bu yaklaşım, yararlı bilgileri vurgulamak ve karar vermeyi desteklemek için özetleme, görselleştirme ve etkileşimli ara yüzleri kullanır. Bir yandan, küresel veri karakteristiklerini elde etmek ve öğrenenlerin davranışları hakkında özet ve raporlar elde etmek için eğitim verilerinden tanımlayıcı istatistikler elde etmek nispeten kolaydır. Öğrencilerin etkinlik sıralarının görselleştirilmesi, öğrenme ortamı kullanım örüntülerini anlamaya yardımcı olur. Model keşfinin amacı, tahmin veya ilişki madenciliği gibi daha ileri düzey analizlerde bir bileşen olarak bir fenomenin geçerliği yüksek bir modelini (tahmin, kümeleme veya bilgi mühendisliğinin kullanılması) kullanmaktır. Öğrencinin davranışları ve özellikleri arasındaki ilişkileri tanımlamak için kullanılabilir (AlMazroui, 2013; Baker, 2010; Bienkowski, Feng ve Means, 2012).

Uç değer tespiti. Amacı, bir veri setindeki uç verilerin belirlenmesi ve bu verilerin veri setinden hariç tutulması işlemidir. Bir uç değer, verideki diğer değerlerden genellikle daha büyük veya daha küçük olan farklı bir gözlemdir (veya ölçümdür). EVM kapsamında, öğrencinin veya eğitimcinin eylemlerinde veya davranışlarında, düzensiz öğrenme süreçlerinde sapmaları tespit etmek ve öğrenme güçlüğü olan öğrencileri tespit etmek için kullanılabilir (AlMazroui, 2013; Baker ve Yacef, 2009).

Sosyal Ağ Analizi. Amaç, bireysel nitelikler veya özellikler yerine bireyler arasındaki ilişkileri incelemektir. Düğümlerden (ağ içindeki bireysel aktörleri temsil eden) ve/veya

bağlantılardan (dostluk, işbirliğine dayalı ilişkiler vb.) oluşan yapıları ağ teorisi kapsamında inceler.

Metin madenciliği. Metin madenciliğinin amacı, ilgili metine ilişkin yapılandırılmamış (metinsel) bilgileri işlemek, metinden anlamlı sayısal indeksler çıkarmak ve böylece çeşitli veri madenciliği (istatistik ve makine öğrenimi) algoritmalarına erişilebilen metinde yer alan bilgileri sağlamaktır. Metin madenciliği kapsamında, dokümanlar içerisindeki kelimelere ilişkin özetler elde edilebilir. Metin madenciliği ile kelimeleri, dokümanlarda kullanılan kelimelerin kümelerini ya da belgeleri analiz edebilir, belgeler arasındaki benzerlikleri belirleyebilir ya da kelimelerin diğer değişkenlerle nasıl ilişkili olduklarını analiz edebilirsiniz (AlMazroui, 2013; Baker ve Yacef, 2009).

Bu tez çalışması kapsamında, veri madenciliğinin temel konularından bir tanesi olan *Kümeleme* yer almaktadır. Kümeleme analizi objeleri obje gruplarına (kümeler) bölerek verilerin içyüzünü açıklamaktadır. Sınıf etiketleri gibi harici bilgi kullanmadıklarından, kümeleme analizine makine öğrenimi ve örüntü tanıma gibi bazı geleneksel alanlarda gözetimsiz-danışmasız öğrenme (unsupervised learning) adı verilmektedir (Atiya, 1990; Öztemel, 2006; Pan, Shen ve Liu, 2013; Reill, Wang ve Rutherford, 2005; Xu ve Wunsch, 2005).

Veri madenciliğinde kümeleme. Kümeleme analizi, gruplandırma yapmak amacıyla var olan çok sayıdaki istatistiksel yöntemlerden bir tanesidir. Küme birbirine benzer ya da yakın ögelerin oluşturduğu topluluk olarak tanımlanmaktadır (Alpar, 2011). Kümeleme analizi birçok matematiksel yöntemi içinde bulunduran ve hangi objelerin özelliklerine göre diğer objelerle aynı kümede yer alacağını belirlemeye çalışan yöntemler topluluğudur (Romesburg, 2004). Kümeleme analizi objeleri sınıflama amacıyla kullanılan istatistiksel bir yöntemdir ve diğer istatistiksel yöntemlerin aksine evrende önemli farklılıklar olduğuna dair önsel varsayımları bulunmamaktadır (Klösgen ve Zytkow, 2002; Punj ve Stewart, 1983; Vellido, Castro ve Nebot; 2010).

Basit olarak ifade edildiğinde, veri madenciliğinde kümeleme *N* adet veri maddesinin her birini *K* adet olası kümelerden birine atamak olarak tanımlanabilir. Bu tanımın *sert* belirlemeler yerine sonucun küme üyeliğinin bir ölçüsü veya olasılığı olduğu bazı bulanık veya olasılıkçı kümeleme yöntemleri durumunda eksik kaldığı unutmamalıdır. Bu belirleme çoğu standart örnekte örneğin noktalar arasında bir Öklitçi mesafe gibi

benzerlik ölçülerini uygulamanın sonucu olabilir. Veri yoğunluğunun ötesinde küme şeklini ve boyutunu da dikkate almamız gerektiği düşünüldüğünde benzerliğin kümelemede genellikle karışık bir kavram olduğu gerçeği göz ardı edilmemelidir (Tatlıdil, 1992; Vellido ve diğ., 2010).

Kümeleme yöntemleri genel olarak hiyerarşik ve hiyerarşik olmayan olarak ikiye ayrılmaktadır. Hiyerarşik yöntemler kendi içinde birleştirici yöntemler (tek bağlantı, tam bağlantı, ortalama bağlantı, Ward's yöntemi ve merkezileştirme yöntemi) ve ayırıcı yöntemler (bölünmüş ortalamalar ve otomatik etkileşim belirleme) olarak ikiye ayrılmaktadır. Hiyerarşik olmayan yöntemler ise K-Ortalamalar Yöntemi, metoid parçalama yöntemi, yığma yöntemi ve bulanık kümeleme yöntemi olmak üzere dörde ayrılmaktadır (Gürsoy, 2009).

Hiyerarşik kümeleme yöntemleri, küme yapısının farklı ve genellikle iç içe yapısal düzeylerde göründüğünü varsaymaktadır. Hiyerarşik olmayan kümeleme yöntemleri ise tüm kümeler için tek bir ortak düzeyi değerlendirmekte ve bu nedenle hiyerarşik kümelemenin özel bir örneği olarak düşünülmektedir. Hiyerarşik olmayan yöntemler arasında belki de en yaygın olanı K-Ortalamalardır. Son yıllarda K-Ortalamalar tabanlı birçok yöntem geliştirilmiştir. Bunlar arasında, Bulanık c-ortalamaları gibi kesin küme üyeliklerini belirlemekten kaçınan bulanık versiyonlar veya küme sayısının analiz öncesinde bilindiği K-Ortalamalar gibi hiyerarşik işlemeye dayalı yöntemler yer almaktadır (Dunn, 1973; Macqueen, 1967; Mclachlan ve Basford, 1988).

Kümeleme tekniklerine ilişkin farklı sınıfların olması, kümeleme analizi ile optimizasyonun amaçlanması, kümeleme yöntemlerinin doğası gereği stokastik (olasılıkçı teknikler) veya keşifsel (çoğunlukla algoritmik) olması analizde kullanılan objektif fonksiyon ile ilişkilidir. Kümelemeye ilişkin keşifsel yöntemler sayıca çoktur ve bu yöntemler köken açısından birbirinden çok farklı (teorik ve uygulama açısından) ve çok çeşitlidir. Kümeleme yöntemleri içerisinde en başarılı yöntemlerden bir tanesi farklı formları bulunan Kohonen'in Öz Örgütlemeli Harita Yöntemidir. Bu modelin kökenlerini sinir bilim sahasının inceleme alanı içerisinde yer alan Yapay Sinir Ağları (YSA) oluşturmaktadır. Bu model eş zamanlı veri kümeleme ile görselleştirme için son derece başarılı bir araçtır. Kohonen'in Öz Örgütlemeli Harita Yöntemi'ne olasılıkçı bir alternatif olarak Oluşturulmuş Topografik Harita yöntemi gösterilmektedir ve bu model Kohonen'in Öz Örgütlemeli Harita Yöntemi işlevselliğinin korunması ve

yöntemin çoğu kısıtlamalarından kaçınılması amacıyla kullanılmaktadır (Bishop, Svensem ve Williams, 1998; Hore Hall ve Goldgof, 2009; Kohonen, 2001; Oja, Kaski ve Kohonen, 2003). Öte yandan Kohonen'in Öz Örgütlemeli Harita Yöntemine ilişkin çıktıların yorumlanmasının kolay olması bu yöntemi Oluşturulmuş Topografik Harita yöntemine göre daha popüler hale getirmiştir.

Sınıflandırma ve kümeleme, nesneleri bir veya daha fazla özellik ile gruplar halinde karakterize eden iki tür öğrenme yöntemidir. Bu iki kavram her ne kadar benzer anlamlara sahip gibi görünse de veri madenciliği bağlamında aralarında fark vardır. Kümeleme, kümelemede hedef bir değişken söz konusu olmadığından sınıflandırmadan farklıdır (Larose ve Larose, 2012). Kümeleme metodu denetimsiz sınıflandırma; sınıflandırma metodu ise denetimli sınıflandırma altında incelenmektedir. Kümeleme sürecinin değerlendirilmesi doğrudan bir şekilde gerçekleştirilmemektedir. Bir kümeleme analizine ilişkin sonuçların, özellikle de ideal küme sayısının kaç olduğuna ilişkin karar verme aşaması geçerliği genellikle araştırmacının sezgisel bakış açısına bağlıdır. Bu aşamada elde edilen sonuçların geçerlik düzeyinin belirlenmesi zor olabilmektedir. Sınıflandırmada ise değerlendirme, genel olarak test setlerinde mevcut sınıf bilgileri bazında gerçekleştirilir (Anil ve Dubes, 1988; Vellido ve diğ., 2010). Sonuç olarak denetimli öğrenmede sınıflama yaparken önceden belirlenmiş özellikler esas alınırken; kümeleme yaparken belirlenen özellik veya özelliklere göre benzer örneklerin bir arada olup olmadığı incelenmektedir.

Kullanılacak olan kümeleme tekniğinin seçimi ile ilgili pek çok ölçüt söz konusudur. Bunlardan birisi de heterojenliktir. Farklı veriler farklı yöntemler gerektirir ve genellikle veriler eş zamanlı olarak farklı formlardadırlar. Bu duruma örnek olarak veri akışlarının analizi ve kademe verileri gösterilebilir. Bütün yöntemler büyük veri setleri ile işlem yapmaya uygun olmadığı için kümeleme yöntemin belirlenmesinde veri boyutu da önemlidir. Çok büyük veri setleri söz konusu olduğunda komşu araştırması, veri özetleme, dağıtılmış hesaplama, adım adım kümeleme ve örnekleme tabanlı yöntemlerin kullanılması önerilmektedir. Bu yöntemlerin dışında bazen, değerlendirilmesi gereken her obje için semantik ilişkilerin söz konusu olduğu çok büyük yapılandırılmış verileri veya grafiksel verileri analiz etmemiz gerekebilmektedir. Bu tip örnekler ile karşı karşıya kalındığında grafik kümeleme

yöntemleri kullanılabilir (Andrews ve Fox, 2007; Hore ve diğ., 2009; Tsuda ve Kudo, 2006).

Kümeleme yöntemleri bir bütün olarak değerlendirildiğinde, birimlerin hangi sınıflarda yer aldığını belirlemede farklı yöntemlerin olduğu ve kullanılan yönteme göre elde edilecek sonuçların farklılık göstereceği belirlenmiştir. Çalışma kapsamında k-ortalamalar, yapay sinir ağını temel almış Kohonen'in Öz Örgütlemeli Harita Yöntemi ve İki Aşamalı Kümeleme Analizi ile elde edilen sonuçların benzer ve farklı yönleri belirlenmiştir. Bunun yanında elde edilen kümeler için küme profilleri belirlenerek kümelerde yer alan bireylerin hangi duyuşsal özelliklerinin daha etkili olduğu ortaya çıkarılmıştır.

Araştırmanın Amacı ve Önemi

Alanyazın incelendiğinde çok fazla kümeleme yönteminin olduğu göze çarpmaktadır. Bu kümeleme yöntemleri içerisinde K-Ortalamalar ve hiyerarşik kümeleme yöntemi araştırmacıların sıklıkla başvurduğu kümeleme yöntemleridir. Bu yöntemler içerisinde K-Ortalamalar Yöntemi büyük veri setleri söz konusu olduğunda çok avantajlı ve en popüler yöntem olarak kabul edilmektedir. K-Ortalamalar Yöntemi çok hızlı ve kaliteli kümeler üreten bir yöntem olarak kabul edilmektedir. Aynı zamanda K-Ortalamalar Yöntemi ile elde edilen çıktılar çok kolay yorumlanabilmektedir ve yöntem gürültülü verilere (noisy data) karşı dayanıklı bir yöntemdir (Genolini ve Falissard, 2010; Usami, 2014). Araştırmacıların kullandığı fakat az bilinen kümeleme yöntemleri de söz konusudur. İki aşamalı kümeleme analizi bunlardan bir tanesidir. İki Aşamalı Kümeleme Analizi ile çok büyük veri setleri analiz edilebilmektedir. Ayrıca iki Aşamalı Kümeleme Analizi ile kategorik ve sürekli veriler hem aynı anda hem de ayrı ayrı analiz edebilmektedir. Aynı zamanda bu analiz ile ideal küme sayısının ne olabileceğine ilişkin sonuçlar elde edilebilmektedir. İki Aşamalı Kümeleme Analizinin araştırmacılar tarafından çok fazla bilinen bir kümeleme analizi türü olmadığı düşünülmektedir (Garson, 2014). Bunun yanında kümeleme analizine ilişkin öğrenme sürecinin yapay sinir ağı temelli olarak gerçekleştirildiği kümeleme yöntemleri de mevcuttur. Kohonen'in Öz Örgütlemeli Harita Yönteminde nitelikli bir öğrenme süreci sonucunda kümeler elde edilmektedir. Bu aşamada, öğrenme süreci sonucunda kümelere ilişkin farklı haritalar oluşturulmaktadır. Aynı zamanda yöntem büyük veri setleri söz konusu olduğunda rahatlıkla kullanılabilmektedir (Kohonen, 2014). Araştırma kapsamında model incelenmesi anlamında, k-ortalamalar, İki Aşamalı

Kümeleme Analizi ve Kohonen'in Öz Örgütlemeli Harita Yöntemi kullanılmıştır. Çalışmada ayrıca küme profilleri belirlenerek elde edilen kümelerde yer alan öğrencilerin hangi duyuşsal özelliklerinin kümelerin oluşmasında daha fazla etkiye sahip olduğu ortaya çıkarılmıştır. Bu sayede çalışma kapsamında ele alınan değişkenlerin birbirlerine kıyasla farklı ülkelerdeki öğrencileri kümelere ayırmada ne düzeyde etkili oldukları belirlenmiştir. Alanyazın incelendiğinde, çalışma kapsamında ele alınan kümeleme yöntemlerinden biri olan Kohonen'in Öz Örgütlemeli Harita Yönteminin kullanıldığı eğitim bilimleri ile ilgili herhangi bir çalışmaya rastlanamamıştır. Fakat PISA kapsamında ölçülen özelliklerin diğer kümeleme yöntemleri ile incelendiği çeşitli araştırmalar mevcuttur (Acar, 2012; Akın ve Eren, 2012; Aksu, Güzeller ve Eser, 2017; Kjærnsli ve Lie, 2011; Linnakyla ve Malin, 2008). Linnakyla ve Malin (2008) Öklid mesafesi temelli hiyerarşik kümeleme yöntemini, Kjærnsli ve Lie (2011) hiyerarşik kümeleme yöntemini, Acar (2012) hiyerarşik olmayan K-Ortalamalar Yöntemini, Akın ve Eren (2012) hiyerarşik kümeleme ve İki Aşamalı Kümeleme Analizini, Aksu, Güzeller ve Eser (2017) K-Ortalamalar Yöntemi ve hiyerarşik kümeleme yöntemini kullanmışlardır. Alınyazında yer alan çalışmalar daha çok eğitim alanında veri madenciliği tekniklerinin kullanıldığı çalışmalar, akademik başarı ve başarısızlığın kestirimi ile bunları etkileyen faktörlerin belirlenmesi (Baker, Growda ve Corbett, 2011; Bırtıl, 2012; Bilen, Hotaman, Aşkın ve Büyüklü, 2014; Kanakana ve Olanrewaju, 2011; Taşdemir, 2012; Taylan ve Karagözoğlu, 2009; Tiwari, Singh ve Vimal, 2013;) eğitimde veri madenciliği alanında yapılan çalışmaları inceleme, tanıtma ve bu tür çalışmaların önemini vurgulama (Ali, 2013; Bhise, Thorat ve Supekar, 2013; Barahate, 2012; Çırak ve Çokluk, 2013; Kumar ve Vijayalakshimi, 2013; Romero ve Ventura, 2007; Romero ve Ventura, 2013; Sharma ve Singh, 2013; Siemens ve Baker, 2012) üzerine yoğunlaşmış durumdadır.

Bu çalışmanın temel amacı 2015 PISA'ya katılım gösteren OECD ülkeleri öğrencilerini fen bilgisi öğretimini etkileyen faktörlere ve fene ilişkin olası başarı puanları (plausible values) ortalamasına göre farklı kümeleme yöntemleri ile modellemek ve elde edilen modelleri incelemektir. Fen öğretimini etkileyen faktörler ve fen başarısı göz önünde bulundurulduğunda oluşan kümelerin özelliklerinin neler olduğunun belirlenmesi ise çalışmanın bir diğer amacı olarak düşünülebilir. Ayrıca çalışma kapsamında, katılımcıların ülke bazında kümelere dağılım, fen bilgisi öğretim

yöntemleri ve fen başarısı anlamında ülkeler arasındaki benzerlikler ve farklılıkların belirlenmesi ilişkin yorum yapılabilecek çıktılar elde edilmiştir.

Bu çalışmanın Kohonen'in Öz Örgütlemeli Harita Yöntemi, K-Ortalamalar Yöntemi ve İki Aşamalı Kümeleme Yöntemi kullanılarak oluşturulan modellerin incelenmesi bakımından önemli olduğu düşünülmektedir. Aynı zamanda çalışmanın PISA öğrenci anketinde yer alan fen bilgisi öğretimine ilişkin maddelere verilen cevaplara göre bireylerin kümelere ayrılmasında değişkenlerin ne düzeyde etkili olduğunun belirlenmesi anlamında alanyazında ilk çalışma olma özelliğine sahip olacağı düşünülmektedir.

Alan yazında PISA sınavı sonucunda elde edilen veriler büyük veri olarak kabul edilmekte ve bu sınavdan elde edilen sonuçlar birçok farklı kurum ve kuruluş tarafından büyük önem taşımaktadır. Eğitimde politika yapıcılar ve uygulayıcılar tarafından önemli sonuçları olan bu sınavdan elde edilen veriler eğitimde veri madenciliği yöntemlerinin PISA veri setleri üzerinde kullanılabileceğine ilişkin bir göstergedir. Çalışma kapsamında veri seti olarak kullanılan PISA sınavı, Ekonomik İşbirliği ve Kalkınma Teşkilatı (Organisation for Economic Co-operation and Development-OECD) tarafından her üç yılda bir 15 yaş grubu öğrencilerin bilgi ve becerilerinin uluslararası anlamda değerlendirildiği dünyanın en kapsamlı tarama araştırmasıdır. PISA ile 15 yaş grubu öğrencilerinin topluma ve yaşam boyu öğrenmeye etkin katılım için uygun olduğu düşünülen ''gerçek hayat'' görevleri üzerindeki performansları değerlendirilir.

15 yaş grubu öğrencilerin sahip olduğu bilimsel prensip ve teorilerin çoğu okulda öğretilmektedir. Diğer alanlarda olduğu gibi fen dersinin okullarda öğretilme şekli sadece öğrencilerin fende başarılı olup olmamalarını değil; aynı zamanda ileri düzey eğitimde ve kariyer planlaması sürecinde yer almak isteyenleri de etkileyebilir. Dünya genelinde fen ile ilgili istihdamda beklenen büyüme ve öğrencilerin okul sebepli fene yönelik ilgilerindeki azalma göz önüne alındığında bazı öğrencilerin neden fen ile ilgili kariyer yapmaya diğer alanlardan daha fazla ilgi duyduklarının incelenmesi daha da önem kazanmıştır. Bu durum, okulda fen öğrenme fırsatını, laboratuvar uygulamalarını, fen öğretmenleri ve fen faaliyetleri gibi fen alanına sunulan kaynakları ve fenin okulda öğretilme yollarını ayrıntılı bir şekilde analiz etme ihtiyacını doğurmuştur. Bu nedenle fen öğretimini etkileyen faktörlerle ilgili çalışmalar önemli görülmektedir. Son 30 yılda fen öğretimini etkileyen faktörler üzerinde çeşitli

çalışmalar yapılmıştır (Langdon, McKittrick, Beede, Khan and Doms, 2011; Vedder-Weiss ve Fortus, 2011). Bu nedenle çalışma kapsamında kümeleme analizi ile model oluşturmada PISA 2015 veri setinde yer alan ve fen başarısında etkili olduğu belirlenen öğretmen yönetimindeki öğretim, algılanan geri bildirim, uyarlanabilir öğretim ve sorgulama temelli öğretim alt boyutlarına ilişkin öğrencilerin verdiği cevaplar ve fene ilişkin olası başarı puanları ortalaması girdi değişkenleri olarak kullanılmıştır.

Araştırmanın fen bilgisi öğretim yöntemlerine ve olası başarı puan ortalamasına ilişkin bireylerin kümelere ayrılması ve oluşan her bir kümenin nasıl tanımlandığı hakkında fikir sahibi olunması anlamında büyük önem arz ettiği düşünülmektedir. Araştırmanın bu bakımdan ilgili alanyazına ve gelecekte gerçekleştirilecek olan çalışmalara büyük ölçüde katkı sağlayacağı tahmin edilmektedir. Gerçekleştirilen alanyazın taraması sonucunda, öğrencilerin fen öğretim yöntemleri ve fen başarısı göz önünde bulundurularak kurulmuş farklı kümeleme modellerinin incelendiği, model sonuçlarının kendi içerisinde değerlendirildiği, tanımlanan kümelerin oluşmasında en etkili olan faktörlerin birey ve ülke temelli yorumlandığı herhangi bir çalışmaya rastlanmamıştır. Bu çalışma klasik kümeleme yöntemlerinin yanında veri madenciliğine dayalı kümeleme yöntemlerinin aynı anda kullanılması ve elde edilen sonuçların incelenmesi yönünden önem arz etmektedir.

Çalışma kapsamında kümeleme analizi kullanma sebeplerinden bir tanesi sınava giren ve bu sınav sonuçlarına dayalı olarak karar alanlar için karmaşık veri yapısı içerisinden daha anlaşılır yapılar keşfedilerek daha anlaşılır sonuçlar elde edilebilmesidir. Kümeleme analizi kullanmanın bir diğer sebebi ise sağlık bilimlerinde büyük veriler yardımıyla yüz ifadelerinden suç profilleri belirlenmeye çalışılması örneğinde olduğu gibi eğitim alanında da EVM yardımıyla büyük yapıdaki veriyi ele alınan değişkenler bakımından daha küçük ve daha anlamlı yapılara ayrıştırarak yorumlanabilir sonuçlar elde edilebilmesidir.

Araştırma Problemi

Bu araştırmanın amacı, 2015 PISA'ya katılım gösteren OECD örneklemindeki öğrencilerin fen bilgisi öğretimini etkileyen faktörlere ve fene ilişkin olası başarı puanları (plausible values) ortalamasına göre k-ortalamalar, İki Aşamalı Kümeleme Analizi ve Kohonen'in Öz Örgütlemeli Harita Yöntemi ile modellemek ve oluşturulan

modelleri incelemektir. Araştırma kapsamında ele alınan değişkenler PISA 2015 öğrenci anketinde yer alan fen öğretimine ilişkin maddeler ve öğrencilerin olası başarı puanı ortalamalarıdır. Bu amaç doğrultusunda aşağıdaki araştırma sorusuna cevap aranmıştır:

1. PISA 2015 öğrenci anketine katılan ve OECD örnekleminde yer alan öğrenciler fen bilgisi öğretimine ilişkin maddeler ve olası başarı puan ortalaması göz önünde bulundurularak nasıl kümelenmektedirler?

Alt problemler. K-ortalamalar yöntemine göre öğrenciler nasıl kümelenmektedir?

İki Aşamalı Kümeleme Yöntemine göre öğrenciler nasıl kümelenmektedir?

Kohonen'in Öz Örgütlemeli Harita Yöntemine göre öğrenciler nasıl kümelenmektedir?

Sayıltılar

Öğrencilerin PISA 2015 öğrenci anketinde yer alan ve araştırma kapsamında kullanılan maddeleri içtenlikle cevapladıkları varsayılmıştır.

Sınırlılıklar

Bu araştırma değişken anlamında, PISA 2015 örnekleminde yer alan öğrencilerin öğrenci anketinden seçilen fen bilgisi öğretimi stratejilerine ilişkin maddeler ve olası fen başarı puan ortalamasına ilişkin maddelerle sınırlıdır.

1. Yapılan araştırma 2015 yılında gerçekleştirilen PISA sonuçları ile sınırlandırılmıştır.

2. Bu araştırma model incelenmesi anlamında üç farklı kümeleme yöntemi ile sınırlandırılmıştır.

3. Bu araştırma sistematik örnekleme sonucunda 9870 OECD ülkesi öğrencisi ile sınırlandırılmıştır.

Bölüm 2
Araştırmanın Kuramsal Temeli ve İlgili Araştırmalar

Bu bölümde; araştırma kapsamında kullanılan k-ortalamalar, İki Aşamalı Kümeleme Analizi ve Kohonen'in Öz Örgütlemeli Harita Yönteminin kuramsal temelinden ve ilgili araştırmalardan bahsedilmiştir.

Araştırmanın Kuramsal Temeli

Kümeleme algoritmaları. Kümeleme analizindeki en eski araştırmalar 1894'e kadar gitmektedir. Bu tarihte Karl Pearson iki adet tek değişkenli bileşenin etkileşim parametrelerini belirlemek için moment eşleme yöntemini kullanmıştır. O tarihten bu yana kümeleme analizi için yeni kümeleme algoritmalarının tasarlanması konusunda çok büyük çabalar gösterilmiştir. Milligan (1996) kümeleme analizinin zorluklarının aşağıdaki üç hususta toplandığını belirtmiştir. (1) Kümeleme temel anlamda hiç bitmeyen bir süreçtir.(2) Kümeleme için yaygın şekilde kabul edilmiş herhangi bir teori söz konusu değildir. (3) İdeal küme sayısının belirlenmesi özneldir ve bu öznellik veri özellikleri ve kullanıcıların anlayışlarıyla yakından ilişkilidir. Bu üç husus, alanyazında çok fazla kümeleme yönteminin bulunmasının ve kümeleme problemlerinin bazı buluşsal yöntemlerle çözülebileceğinin sebebi olarak düşünülebilir (Jain ve Dubes, 1988; Kaufman ve Rousseeuw; 1990; Kleinberg, 2002).

Her bir kümeleme analizi kapsamında bir kümeleme algoritması kullanılmaktadır. Kümeleme analizi sonucunda oluşan küme sayısı ve küme kalitesinde ilgili kümeleme analizinde kullanılan algoritma büyük pay sahibidir. Kümeleme algoritmaları; *Prototip Tabanlı Algoritmalar*, *Yoğunluk Tabanlı Algoritmalar*, *Grafik Tabanlı Algoritmalar*, *Hibrit Algoritmalar*dır (Anderberg, 1973; Berkhin, 2002; Jain ve Dubes, 1988; Kaufman ve Rousseeuw; 1990; Kleinberg, 2002; Mirkin, 1996) başlıkları altında incelenmektedir. Aşağıdaki bölümde, kümeleme analizlerinde kullanılan bu algoritmalardan bahsedilmektedir.

İlk örnek (prototip) tabanlı algoritmalar. Prototip tabanlı algoritma sonucu oluşan her bir küme prototipler etrafında toplanan veri objeleri ile meydana gelmektedir. Yapay sinir ağları temelli bir eğitim sürecine sahip olan Kohonen'in Öz Örgütlemeli Harita ve Bulanık C ortalama yöntemi prototip tabanlı algoritmalar kapsamında değerlendirilmektedir. Kohonen'in Öz Örgütlemeli Harita Yöntemi kapsamında kullanılan algoritma veri objelerinin özelliklerini korumak için bir

komşuluk fonksiyonu kullanmaktadır. Çalışma kapsamında kullanılan veri analizi yöntemlerinden olan Kohonen'in Öz Örgütlemeli Harita Yöntemi prototip tabanlı kümeleme teknikleri kapsamında değerlendirilmektedir (Anderberg, 1973; Berkhin, 2002).

Yoğunluk tabanlı algoritmalar. Bu tür algoritmalar ile bir küme düşük yoğunluk bölgeleri (eleman sayısının az olduğu bölgeler) ile çevrili olan yoğun bir veri bölgesi olarak ele alınmaktadır. Bu algoritmalar genellikle kümeler iç içe geçmiş (bir elemanın birden fazla kümeye ait olma durumu) olduğunda veya gürültülü veriler (veri girişi veya veri toplanması esnasında oluşan sistem dışı hatalar) olduğunda kullanılmaktadır. Yoğunluk tabanlı algoritmalardan en çok kullanılanları Density-Based Spatial Clustering of Applications with Noise (DBSCAN) ve Density-based Clustering (DENCLUE) algoritmalarıdır. DBSCAN, Öklit yoğunluğuna (öklit uzaklığının temel alınarak objelerin veya canlıların yoğunlaştığı bölgeler) dayanarak veri objelerini sırasıyla çekirdek noktalara, sınır noktalarına ve gürültüye böler ve sonra kümeleri doğal olarak bulur. DENCLUE bir olasılık yoğunluk fonksiyonunu kernel fonksiyonuna dayalı olarak tanımlamaktadır. Çok boyutlu veriler söz konusu olduğunda, yoğunluk bilgisi sadece özellikler alt uzayında geçerlidir (Kleinberg, 2002; Mirkin, 1996).

Grafik tabanlı algoritmalar. Grafik tabanlı algoritmalar kapsamında veri objeleri düğümlenmekte ve iki obje arasındaki mesafe iki düğümü bağlayan mesafenin ağırlığı olarak düşünülerek bir grafik oluşturulmaktadır. Jarvis-Patrick algoritması (JP) her veri objesi için paylaşılan en yakın komşuları tanımlayan ve sonra kümeleri elde etmek için grafiği seyrekleştiren tipik grafik tabanlı algoritmadır. Son yıllarda spektral kümeleme bu alanda önemli bir konu olmuştur. Burada veriler çeşitli grafik türleri ile temsil edilebilir ve daha sonra doğrusal cebir grafikler üzerinde tanımlı optimizasyon problemlerini çözmek için kullanılır. Alanyazında grafik tabanlı algoritmalara Normalleştirilmiş Kesitler ve MinMaxCut gibi birçok spektral kümeleme algoritmaları önerilmiştir (Anderberg, 1973; Berkhin, 2002; Jain ve Dubes, 1988).

Hibrit algoritmalar. Kümelemeye ilişkin tek bir algoritmanın kullanımı sonucunda oluşan eksiklikleri aşmak için iki veya daha fazla kümeleme algoritmasının birlikte kullanıldığı hibrit algoritmalar önerilmektedir. Chameleon tipik bir hibrit algoritmadır ve ilk olarak verileri pek çok küçük bileşene ayırmak için grafik tabanlı bir algoritma kullanır, son olarak ise nihai kümeleri elde etmek için hiyerarşik

kümeleme analizini kullanır. Çalışmada kullanılan İki Aşamalı Kümeleme Analizi, hiyerarşik olmayan kümeleme tekniklerinden “k Ortalamalar” ve hiyerarşik tekniklerden olan “Ward’ın En Küçük Varyans” tekniğinin birleştirilmesi ile oluşan hibrit kümeleme teknikleri içerisinde yer almaktadır (Ceylan, Gürsev ve Bulkan, 2017).

Yukarıdaki algoritmalara ek olarak veri madenciliği kapsamında, büyük verilerdeki örüntü kalıplarını keşfetmek için birçok farklı türde öğrenme algoritması kullanılmaktadır. Bu algoritmalar, denetimli öğrenme ve denetimsiz öğrenme adı altında iki farklı kategori kapsamında değerlendirilmektedir. Denetimli ve denetimsiz öğrenme yöntemlerinin her ikisinde de; istenen çıktı göz önünde bulundurulduğunda girdi ve çıktıya ilişkin en iyi sonuçların elde edilebilmesi; veri setine ilişkin örtük ve çok kolay belirlenemeyen yapıların keşfedilmesi, veri setinin doğal yapısının belirlenmesi ve farklı öğrenme yöntemlerinin karşılaştırılabilmesi büyük önem taşımaktadır.

Denetimli öğrenme. Denetimli öğrenme alanyazında (Pan, Shen ve Liu, 2013) denetimsiz öğrenmeye göre daha sık kullanılmaktadır. Denetimli öğrenme doğrusal ve lojistik regresyon, çok düzeyli sınıflandırma ve destek vektörleri gibi algoritmaları içermektedir. Denetimli öğrenmede, veri madencisi kullandığı algoritmaya ne gibi sonuçlar ortaya çıkabileceğini öğretmesine ilişkin rehberlik yapmaktadır. Bu durum, bir çocuğun matematik öğretmeninden aritmetiği öğrenmesine benzer. Denetimli öğrenme, algoritmanın olası çıktılarının bilinmesini ve algoritmayı eğitmek için kullanılan verilerin doğru cevaplarla etiketlenmesini gerektirir. Bir sınıflandırma algoritmasının herhangi bir hayvan türüyle uygun şekilde etiketlenmiş ve bir takım belirleyici özellikler taşıyan bir veri kümesi üzerinde eğitildikten sonra hayvanları tanımlamayı öğrenmesi bu duruma örnek olarak gösterilebilir (Kotsiantis, Zaharakis ve Pintelas, 2006).

Denetimsiz öğrenme. Denetimsiz öğrenme, psikolojiden mühendisliğe kadar çok farklı yaklaşımlar söz konusu olduğunda ele alınabilecek derin bir kavramdır. Genellikle ‘’Öğretmen olmadan öğrenme’’ olarak adlandırılmaktadır. Bu öğrenme şekli, gerçek yapay zeka olarak adlandırılan kavram ile çok yakından ilişkilidir. Bu durum, bir bilgisayarın süreç boyunca rehberlik görevini üstlenmesi yani karmaşık süreç ve kalıpların tanımlanması konusunda kendi kendine öğrenebileceği fikrine dayanmaktadır. Denetimsiz öğrenme, kuramsal anlamda karmaşık olmayan ve basit vakalar söz konusu olduğunda karmaşık gibi görünse de, araştırmacıların normalde

başa çıkamayacağı problemleri çözmek için çok faydalı olmaktadır. Denetimsiz öğrenmenin sonucu genellikle, gözlem verilerinin yeni bir açıklaması ya da temsili şeklindedir. Böylece gelecekte daha iyi cevaplar veya kararlar verilebilmektedir. Bu öğrenme biçimi kapsamındaki modeller için doğru veya yanlış cevap gibi bir durum söz konusu değildir. Örneğin söz konusu olan kümeleme modelleri ise hangi modelin veri setin için daha uygun olduğu ve daha faydalı sonuçlar vereceği, veri setine ilişkin ilginç gruplaşmalar ve bu gruplaşmalara ilişkin açıklamalar sonucunda belirlenmektedir. Denetimsiz öğrenme genel olarak, kümeleme algoritmalarını, temel bileşenler analizini, ilişkilendirme (association) kurallarını içermektedir (Barlow, 1989; Becker ve Plumbley, 1996)

Çalışma kapsamında gerçekleştirilen Kohonen'in Öz Örgütlemeli Harita Yöntemi, K-Ortalamalar Kümeleme Analizi ve İki Aşamalı Kümeleme Analizi denetimsiz öğrenme kapsamında incelenmektedir ve aşağıda bu analizlerden bahsedilmiştir.

Kohonen'in Öz Örgütlemeli Harita Yöntemi

Modelin tanımı. Yazılım, donanım ve ölçme araçlarındaki teknolojik ilerlemelere bağlı olarak, araştırmacılar giderek çok çeşitli büyüklüklerde ve boyutlarda olan veri kümelerini toplayabilmekte ve analiz edebilmektedirler. Büyük veriler söz konusu olduğunda, görselliğin çok önemli bir rol oynadığı boyut indirgeme yöntemlerine ilişkin sonuçların yorumlanması daha da zorlaşmaktadır. Ayrıca, desen tanıma ve diğer yeteneklerimizi daha iyi kullanabilmemiz için iki boyutlu ve anlamlı eşleştirmelerin yer aldığı yöntemlere ihtiyaç vardır. Yüksek boyutlu bir veri setinin iki boyuta indirgenmesine ilişkin birkaç yöntem bulunmaktadır. Bu yöntemlerden en çok kullanılanı temel bileşenler analizidir. Temel bileşenler analizi kapsamında, ikiden daha fazla boyut söz konusu olduğunda görselleştirme bir sorun olarak devam etmekte ve analiz sonuçlarının daha iyi görsellerle desteklenmesi ihtiyacı doğmaktadır. Üstelik, temel bileşenler analizi temelde nesnelerin nasıl karşılaştırılacağı hakkında bilgi vermemektedir (standart Öklid mesafe ölçüsü her zaman optimal farklılıklar konusunda bilgi vermez). Bu noktada uzaklık ve benzerlik matrislerini temel alan yöntemlerin daha yararlı yöntemler olduğu düşünülmektedir. Uzaklık ve benzerlik matrislerini temel alan yöntemler, büyük veri setlerini analiz etmede ve eldeki veriler için uygun bir mesafe fonksiyonu seçerek en bilgilendirici veriler üzerine yoğunlaşma konusunda diğer yöntemlere göre daha başarılıdır. Uzaklık matrisinin iki boyutlu bir şekilde görselleştirilmesine yönelik bir başka

yaklaşım ise çok boyutlu ölçeklemedir. Çok boyutlu ölçekleme, mesafe matrisi çok boyutlu veriler kullanılarak hesaplanan ve orijinal mesafe matrisine yaklaşan iki boyutlu uzayda bir konfigürasyon bulmayı amaçlamaktadır. Fakat çok boyutlu ölçekleme söz konusu olduğunda elde edilen bulguların yorumlanması zor olabilmekte, sürecin birkaç kez daha gerçekleştirilmesi zorunluluğu ortaya çıkabilmektedir. Üstelik, yeni nesneleri aynı alana yerleştirmenin basit bir yolu da bulunmamaktadır (Gurney, 1997; Haykin, 2009; Kohonen, 2014).

Kohonen'in Öz Örgütlemeli Harita Yöntemi (Kohonen, 2001) yukarıda bahsedilen sorunu çok boyutlu ölçeklemeye benzer şekilde ele almakta fakat birimler arasındaki mesafeleri yeniden hesaplamaya çalışmak yerine birimlerin yer aldığı uzayı yeniden oluşturmaya çalışmakta, başka bir ifadeyle birimlerin bu uzayda yerlerini gösteren veri noktalarına ilişkin komşulukları sabit tutmayı amaçlamaktadır. Eğer çok boyutlu uzayda iki birim çok benzer ise bu birimlerin iki boyutlu düzlemdeki konumları da çok benzer olacaktır. Birimleri "sürekli bir uzayda" eşleştirmek yerine, birimlerin düğüm ya da nöronlar kullanılarak haritalandırıldığı gösterim yöntemleri tercih edilmektedir Çok boyutlu ölçekleme en büyük farklılıklara odaklanırken, Kohonen'in Öz Örgütlemeli Harita Yöntemi büyük benzerliklere odaklanmaktadır. Bir başka deyişle, çok boyutlu ölçekleme ile gerçekleştirilen iki boyutlu bir gösterimde büyük bir mesafe ile gerçek bir mesafenin tahmini doğrudan yorumlanabilirken; Kohonen'in Öz Örgütlemeli Harita Yöntemi ile sadece aynı veya komşu birimlere eşlenen nesnelerin çok benzer olduğu söylenebilir. Kohonen'in Öz Örgütlemeli Harita Yöntemi birimlerin hangi kümede yer aldığının belirlenmesi sürecinde yapay sinir ağlarını temel almaktadır. Modeli meydana getiren düğümler rekabete dayanan (competitive learning) bir süreç ile eğitilmektedir.

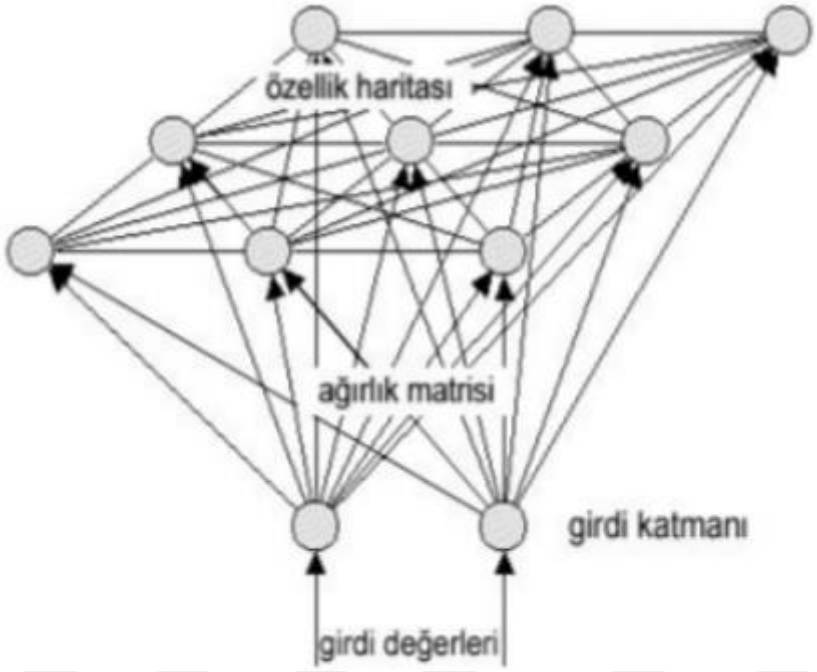
Araştırma kapsamında, modelin nasıl işlediğinin daha iyi anlaşılabilmesi için modele ilişkin öğrenme başlığı altında yapay sinir ağlarından da bahsedilmiştir. Ancak öncelikle, yöntemin özelliklerine ilişkin daha ayrıntılı bilgiler sunulmuştur.

Kohonen'in Öz Örgütlemeli Harita Yöntemi ile kurulan model, kısmen çok boyutlu ölçekleme gibi doğrusal olmayan yöntemlere benzemektedir. Bu yöntemler kapsamında çoğunlukla çok boyutlu bir veri seti, iki boyutlu bir Öklid düzlemi üzerine, düzlem üzerindeki çıkıntıların karşılıklı mesafeleri yaklaşık olarak eşit olacak şekilde eşleştirilir. Benzer ögeler birbirine yakın konumlandırılır; farklı ögeler ise ekranda sırasıyla birbirinden ayrılır. Ögeler daha sonra, soyut bir düzlem üzerinde temsil

edilirler. Bununla birlikte model, verilerin yerel ortalamaları olan modellere göre girdi verilerini temsil eder. Sadece bazı özel durumlarda girdi ögelerinin görüntüsüyle olan ilişkisi haritada birebir olabilmektedir. Özellikle endüstriyel ve bilimsel uygulamalarda haritalama bire birdir; bir başka deyişle birimin harita üzerinde yansıyan görüntüleri klasik vektör nicelemesinde (ölçülebilir özellikleri sayısal olarak dile getirme) niceleme (quantization) K-Ortalamalar ile karşılaştırılabilen, girdi veri dağılımının yerel ortalamalarıdır. Vektör nicelemesinde, yerel ortalamalar bir dizi kod çizelge vektörleri ile temsil edilir. Kohonen'in Öz Örgütlemeli Harita Yöntemi, yerel ortalamaların temsil edilmesi anlamında modeller olarak da adlandırılan ve sonlu bir küme olan kod çizelgesi vektörlerini (codebook vectors) kullanır. Bir giriş vektörü, tüm modellerle karşılaştırılarak harita dizisindeki belirli bir düğüme eşlenir ve "kazanan" olarak adlandırılan en iyi eşleşen model vektör nicelemesinde olduğu gibi tanımlanır. Kohonen'in Öz Örgütlemeli Harita Yöntemi ile K-Ortalamalar arasında eğitim süresi, potansiyel sonuçlar, temel alınan algoritma (Kohonen'in Öz Örgütlemeli Harita Yöntemi için yapay sinir ağı), istenilen küme sayısına ilişkin sonuçların elde edilip edilememesi, kümelerin merkezkaç kuvveti ve küme büyüklüğü mü yoksa geometri temelli mi oluşturulduğu vb. farklılıklar dışında, Kohonen'in Öz Örgütlemeli Harita yöntemi ile K-Ortalamalar kümelenmesi arasındaki en temel farklılık, Kohonen'in Öz Örgütlemeli Harita Yöntemi kaynak verilere benzeyen projeksiyon görüntüleri arasındaki topografik ilişkileri de yansıtmaktadır. Bu nedenle yöntem aynı zamanda, "Veriye ilişkin topografik ilişkilerin topografik bir harita üzerinde düzenli bir şekilde temsil edildiği bir veri sıkıştırma yöntemi" şeklinde de tanımlanabilir (Kohonen, 2001; Kohonen, 2014).

Kohonen'in Öz Örgütlemeli Harita Yöntemi, istatistiksel testlerin karşılaması gereken herhangi bir varsayıma dayanmamaktadır. Kümelerin başlangıç sayısına, değişkenlere ilişkin olasılık dağılımlarına ve değişkenlerin bağımsızlığına ilişkin herhangi bir varsayım gerektirmemesi; özellikle de çok boyutlu veri setleri (çok boyutluluk istatistiksel korelasyonları anlamsız hale getirmekte ve dolayısıyla istatistiksel yöntemler bu türden veri setlerini analiz etmede yetersiz ve güçsüz kalmaktadır) ile çalışılmak istendiğinde diğer yöntemlere göre çok daha iyi sonuçlar veren Kohonen'in Öz Örgütlemeli Harita Yöntemini diğer yöntemlerden daha kullanışlı hale getirmektedir (Dasu ve Johnson, 2003; Dunham, 2003; Penn, 2005).

Sinir sistemlerinden esinlenerek oluşturulan yapay sinir ağlarında sinir hücreleri olarak tanımlanan nöronlar ve bilginin nöronlar arasında taşınmasında rol oynayan düğümler kullanılmaktadır. İnsan beyni taklit edilerek oluşturulan yapay sinir ağlarında girdi değerleri ile özellik haritası Şekil 3'te gösterilmiştir.



Şekil 3. Yapay sinir ağlarının yapısı

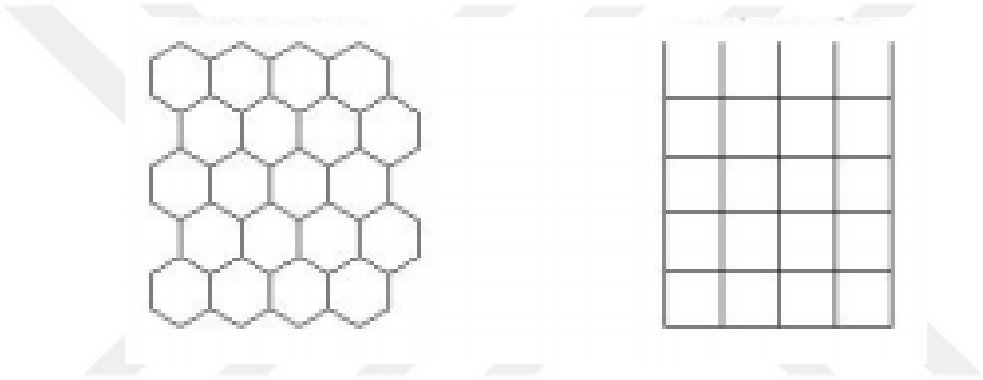
Yapay sinir ağlarından esinlenerek gerçekleştirilen model oluşturma süreci dört temel basamaktan oluşmaktadır. İlk basamak başlatma (initialization) basamağıdır. Bu basamakta, düğümlere ilişkin ağırlıklar küçük ve rastgele değerler ile başlatılır. İkinci basamak, rekabet (competition) basamağıdır. Bu basamakta ise, her bir girdi modeli için, nöronlar, rekabete temel oluşturan diskriminant fonksiyonunun ilgili değerlerini hesaplar. Diskriminant fonksiyonunun en küçük değeri olan özel nöron, kazanan nöron ünvanını alır. Üçüncü adım işbirliği (cooperation) adımıdır. Bu adımda kazanan nöron, uyarılmış nöronların komşuluk anlamında mekânsal olarak yerini belirler ve böylece komşu nöronlar arasında işbirliğinin temeli oluşur. Dördünce ve son adım ise adaptasyon (adaptation) adımıdır. Bu adımda uyarılmış nöronlar, örüntü deseni ağırlıklarının uygun şekilde ayarlanması yoluyla, giriş modeline göre diskriminant fonksiyonunun bireysel değerlerini azaltır. Böylece, kazanan nöronun, benzer bir girdi modeli uygulamasına verdiği tepki artmaktadır (Gurney, 1997; Haykin, 2009; Kohonen, 2014). Aşağıda, yapay sinir ağlarını temel alan modele ilişkin öğrenme sürecinden bahsedilmektedir.

Modele ilişkin öğrenme süreci. Kohonen'in Öz Örgütlemeli Harita Yöntemi, sinir ağı modellerinden, özellikle de çağrışımsal bellek ve uyarlamalı öğrenme modellerinden ortaya çıkmıştır (Kohonen, 1984). Yöntem, beynin işlevlerinin mekânsal organizasyonunu açıklama konusunda serebral korteks gözlemlerine dayalı sonuçlar elde etmeyi temel almaktadır. Kohonen'den önce, Malsburg'un (1973) mekânsal sıralı hat detektörleri ve Amari'nin sinirsel alan modeli (1980) Kohonen'in Öz Örgütlemeli Harita Yönteminin temelini oluşturmaktadır. Kohonen'in ağları, kendi kendini organize eden sinir ağlarını temel almaktadır. Kendi kendini organize etme yeteneği yeni olasılıkların oluşmasına zemin hazırlar. Aynı zamanda bu özellik, insan beyninde şekillenen en doğal öğrenme şeklidir. Yeni olasılıklar öğrenme sürecinde şekillenmektedir. Kohonen'in ağları, kendi kendini organize eden, rekabetçi tipte öğrenme yöntemini kullanan ağ grupları sunmaktadır.

Kohonen'in Öz Örgütlemeli Harita Yöntemi kapsamında ağ girişleri üzerinde sinyaller oluşturulur ve daha sonra giriş vektörüne en iyi şekilde karşılık gelen yani kazanan nöron belirlenir. Nöronların rekabetine ilişkin şema ve sinaptik hücrelerin modifikasyonları çeşitli formlarda olabilir. Yönteme ilişkin, rekabete dayalı ve kendi kendini organize eden bir algoritma ile farklılaşabilen birçok alt tip söz konusudur. Bu alt sistemlerden belki de en önemlisi kazan-kazan işlevini benimsemiş olan rekabetçi sinir ağı anlayışıdır. Bunun yanında, sinir ağı tarafından kontrol edilen ve öğrenmedeki nöronların yerel anlamdaki sinaptik esnekliğini değiştiren başka bir alt sistem de vardır. Öğrenme yer olarak, en aktif nöronların komşuluğu ile sınırlıdır. Kontrol ile ilgili olan alt sistemin esnekliği spesifik olmayan sinirsel etkileşimlere dayanabilir ancak bu durum daha çok bir kimyasal kontrol etkisidir. Kendi kendini organize etme sisteminin oluşması, nöral sinyal aktarımı ve esneklik kontrolünün ayrılması sayesinde mümkün olur. (Kaski, Kangas ve Kohonen, 1998; Kohonen, 2014). Bununla birlikte, Kohonen'in Öz Örgütlemeli Harita Yöntemi altta yatan herhangi bir nöral ya da başka bir bileşene atıfta bulunmaksızın saf ve soyut matematiksel bir formda da ifade edilebilir (Kaski, Kangas ve Kohonen, 1998; Oja ve diğ., 2003).

Genel anlamda K-Ortalamalar algoritması Kohonen'in Öz Örgütlemeli Harita Yöntemi ile gerçekleştirilen kümeleme analizinin özel bir durumu olarak düşünülebilir (K-Ortalamalar kapsamında komşuluk ilişkileri dikkate alınmazken Kohonen'in Öz Örgütlemeli Harita Yönteminde komşuluk ilişkileri dikkate alınmaktadır) (Ripley,

1996). Bu benzetme çerçevesinde, her birim bir “kümeye’’ karşılık gelir ve kümelerin sayısı, tipik olarak dikdörtgen veya altıgen bir şekilde düzenlenmiş olan nöronların yer aldığı ızgaranın boyutu ile tanımlanır ve model kapsamında, analiz işlemi gerçekleştirilmeden önce eğitime tabi tutulur. Köşeler sayılmadığı taktirde altıgen şeklinde nöronlardan oluşan ızgara formu altı, dikdörtgen şeklinde nöronlardan oluşan ızgara formunun ise dört komşuya sahiptir. Kohonen’in Öz Örgütlemeli Harita Yöntemi ile oluşturulan yapının topolojik bir yapı olması sebebiyle altıgen şeklinde olan nöronlar ile daha düzgün haritalar oluşturulduğu düşünülmektedir (Dasu ve Johnson, 2003; Kloptchenko, Eklund, Karlsson, Back, Vanharanta ve Visa, 2004; Kohonen, 2014). Şekil 4 kapsamında altıgen ve dikdörtgen nöronlardan oluşan ızgaralar gösterilmektedir.



Şekil 4. Altıgen ve dikdörtgen nöronlar

Kohonen’in Öz Örgütlemeli Harita Yöntemi kapsamındaki eğitim süreci, K-Ortalamalar algoritması ile gerçekleştirilen kümeleme analizine çok benzemektedir. Eğitim süreci, her birime bir kod vektörü atanarak başlar ve bu durum, eğitim sürecinin temel taşıdır. Her bir birime rastgele veri kümesi atanır. Eğitim süresince, nesneler haritaya tekrar tekrar rastgele bir sırayla sunulur. “Kazanan birim” yani mevcut eğitim nesnesine en benzer olanı, mevcut eğitim nesnesine daha da benzer hale gelecek şekilde güncellenmektedir. Güncelleme işlemi süresince ağırlıklandırılmış ortalama kullanılır ve yeni nesnenin ağırlığı analizin eğitim parametrelerinden biridir. Eğitim süreci için, öğrenme oranı alfa olarak adlandırılmaktadır ve genellikle varsayılan değer olan 0.05 şeklinde ayarlanmaktadır (Kohonen, 2001; Kohonen, 2014).

Kısacası, modele ilişkin algoritma uygulandığında, konfigürasyon (dikdörtgen veya altıgen) ve ağdaki nöronların sayısı önceden tanımlanmış olmalıdır. Bazı durumlarda, haritadaki maksimum nöron sayısının kullanılması önerilmektedir. Başlangıçtaki

öğrenme yarıçapı genelleme kapasitesini etkiler. Haritada yer alan düğüm sayısının öğrenme sürecindeki örnek sayısını aştığı durumda algoritmanın başarısı büyük ölçüde seçilen başlangıç yarıçapına bağlıdır. Harita binlerce nörondan meydana geliyorsa, eğitim süreci uzun zaman alır. Bu nedenle nöron sayısını belirlerken nöron sayısının kabul edilebilir miktarda olmasına dikkat edilmelidir.

Nöronlara ilişkin ağırlıkları başlatmak için üç yol vardır. Bunlardan birincisi, tüm ağırlıklara rastgele değerler verildiğinde rastgele değerler ile başlangıç; ikincisi, eğitim örneğinden rastgele seçilen örneklerin başlangıç değerleri olarak tanımlandığı örneklerle başlangıç; üçüncüsü ise ağırlıkların orijinal veri kümesinin iki temel özvektörü arasında geçen doğrusal alt uzay boyunca doğrusal bir düzende vektör değerleri ile başlatıldığı doğrusal başlangıçtır.

Modele ilişkin öğrenme süreci, nöron (vektör) düzeltme dizisi olarak adlandırılabilir. Öğrenmenin her adımında bir vektör, en benzer nöron katsayısı vektörü ile eşleştirilmek için veri kümesi içerisinden seçilir. Giriş vektörüne en çok benzeyen sadece bir kazanan nöron aşağıda verilen eşitlik yardımıyla seçilir. Bu durumda benzerlik, Öklid uzayında hesaplanan vektörler arasındaki mesafedir. C kazanan nöron, x vektör, w_c kazanan nöronun ağırlıklandırılmış değeri, w_i i. sıradaki vektörün ağırlıklandırılmış değeri olmak üzere aşağıdaki denklem elde edilir:

$$|x - w_c| = \min\{|x - w_i|\}$$

Kazanan nöron belirlendiğinde, sinir ağına ilişkin ağırlıklar düzeltilir. Kazanan nöronu ve bu nöronun ızgara üzerindeki komşularını tanımlayan vektör giriş vektörüne doğru taşınır.

Model eğitilirken verilerin tamamı aynı anda eğitime katılmaz. Veriler belli sayıda nöronlar halinde eğitimde yer alır. İlk nöron eğitilir ve bu ilk nöronun eğitiminden sonra modelin başarısı test edilir, başarıya göre geriye yayılım (“backpropagation”) ile ağırlıklar güncellenir. Daha sonra yeni eğitim kümesi ile model tekrar eğitilip ağırlıklar tekrar güncellenir. Bu işlem her bir eğitim adımında tekrarlanarak model için en uygun ağırlık değerleri hesaplanmaya çalışılır. Bu eğitim adımlarının her birine “epoch” (döngü) denilmektedir. X(t) vektörü, t defa eğitime katılan eğitim seti içerisinden rastgele seçilir. Uzaklık ölçütü olarak kullanılan H(d,t) gösterimi, kazanan nöron ve ızgara üzerinde yer alan komşu nöronlar arasındaki zaman ve mesafeye ilişkin

monoton azalan bir fonksiyondur. Bu fonksiyon, mesafe fonksiyonu ve zamanla öğrenme oranı fonksiyonu olmak üzere iki kısma ayrılır.

Alanyazında genellikle iki mesafe fonksiyonu kullanılır. Bunlardan birincisi basit sabittir. İkincisi ise gauss fonksiyonudur. Genelde en iyi sonuç, gauss mesafe fonksiyonu kullanılarak elde edilir. Gauss fonksiyonuna ilişkin formülde t, epoch sayısı; d, uzaklığa ilişkin gösterge; h(d,t) nöronlar arasındaki etkileşimi; σ(t) ise zamana göre azalan komşuluğa ilişkin yarıçapı göstermektedir. Aşağıda, gauss fonksiyonuna ilişkin formül yer almaktadır:

$$h(d,t) = -e^{-\frac{d^2}{2\sigma^2(t)}}$$

Gauss fonksiyonu zamanla azalan bir fonksiyondur. Genellikle bu değer, öğrenmenin başlangıç aşamasında yeterince büyük olan ve bir kazanan nöron bırakılana kadar yavaş yavaş azalan öğrenme yarıçapı olarak adlandırılır. En fazla kullanılan fonksiyon, zamanla doğrusal olarak azalan fonksiyondur.

Öğrenme oranının α(t) olduğunu varsayalım. Bu oran, zamanın azalan bir fonksiyonudur. Öğrenme oranına ilişkin form zaman ile ters orantılıdır:

$$\alpha(t) = \frac{A}{t + B}$$

Yukarıdaki fonksiyonda A ve B katsayıları, kümeleme sürecinin başlangıcında sabit olarak alınan noktanın yatay ve dikey eksendeki konumunu gösteren değerlerdir. Bu fonksiyon kullanıldığında, eğitim setindeki tüm vektörler eğitim sonunda öğrenme sonucuna yaklaşık olarak eşit katkıda bulunmuş olurlar.

Öğrenme iki ana aşamadan oluşmaktadır. Başlangıç aşamasında öğrenme oranı ve yarıçap değerleri, nöron vektörlerinin örneklemin dağılımına uyacak şekilde dağıtacak kadar büyük seçilir. Öğrenme oranı parametre değerleri ilk aşamadaki değerlerinden daha düşük olduğunda, ağırlıklara ince ayar yapılmaktadır. Doğrusal başlatma kullanıldığında, ağırlıklara ilişkin ince ayar süreci atlanabilir (Bishop 1995; Gurney, 1997; Haykin; 2009)

Modele ilişkin görselleştirme. Kohonen'in Öz Örgütlemeli Harita Yöntemi kapsamında, verinin farklı bölgelerindeki kümelenme yoğunluğunun gösteriminde tek bir grafik kullanılmaktadır. Düzenlenmiş bir haritadaki referans vektörlerinin yoğunluğu arttıkça o bölgede daha fazla birimin bir araya geldiği ve dolayısıyla

kümelemenin gerçekleştiği sonucuna ulaşılmaktadır (Kohonen, 2014; Ritter, 1991). Kümelenmiş alanlarda referans vektörleri birbirine yakın, kümeler arasındaki boş alanlarda ise daha seyrek olacaktır. Böylece, veri setinin küme yapısı, komşu birimlerin referans vektörleri arasındaki mesafeler gösterilerek görünür hale getirilebilir.

Kümelere ilişkin harita oluşturulurken, her bir referans vektör çifti arasındaki mesafe hesaplanır ve ölçeklendirilir, böylece mesafeler isteğe bağlı olarak belirli bir minimum ve maksimum değer arasına sığar. Harita ekranında, ölçeklenmiş her mesafe değeri, karşılık gelen harita birimlerinin ortasındaki noktanın grilik seviyesini veya rengini belirler. Harita birimlerine karşılık gelen noktalardaki gri seviye değerleri, en yakın mesafe değerlerinin bazılarının ortalamasına (altıgen bir ızgara üzerinde, örneğin alt sağ köşeye doğru olan altı mesafenin üçünün ortalamasına) göre ayarlanır. Bu değerler belirlendikten sonra harita oluşturulabilir veya değerler mekânsal olarak düzleştirilebilirler. Ortaya çıkan haritadaki kümelerin şekilleri, kümelerin gerçek şekilleri konusunda gerçekçi bir bilgi sunmayabilir. Kümeleme algoritmalarının çoğu, belirli şekillerin kümelerini tercih etmektedir (Jain ve Dubes, 1988).

İdeal küme sayısının belirlenmesi. Kohonen'in Öz Örgütlemeli Harita Yöntemi ile ideal küme sayısının belirlenebilmesi için R programı kapsamında dirsek (elbow) metodu ve siluet katsayısı kullanılabilmektedir. Kohonen'in Öz Örgütlemeli Harita Yöntemi için dirsek metodu küme içi kareler toplamının küme sayısına göre değişimine ilişkin görsel yorumlanırken kullanılmaktadır. Bu noktada dirsek yöntemi, küme sayısının bir göstergesi olarak açıklanan varyans miktarını dikkate almaktadır. Açıklanan varyans miktarı göz önünde bulundurularak kümeler birbiri ile çakışmayacak şekilde modellenmelidir. Küme içi kareler toplamının küme sayısına göre değişimine ilişkin görsel yorumlanırken, marjinal kazancın düştüğü yani grafiğin düz bir plato şeklinde olmaya başladığı noktanın iz düşümünün ideal küme sayısına işaret ettiği belirtilmektedir. Hem uyuşma (cohesion) hem de ayrışma (seperation) anlamında kombine bir ölçü olarak kabul edilen siluet katsayısı da ideal küme sayısının belirlenmesinde kullanılmaktadır. Siluet katsayısı, küme elemanlarının komşu kümelere ne kadar yakın olduğu konusunda bilgi vermektedir. Katsayı, [-1, +1] aralığında bir değer almaktadır. +1 civarındaki puanlar komşu kümelerden çok uzakta olunduğuna işaret etmekte iken, 0 civarında olan değerler kümeler arasındaki sınırlara yakın olunduğunu göstermektedir. Negatif değerler ise örneklerin yanlış

sınıflandırılmış olma ihtimalini göstermektedir (Jain ve Dubes, 1988; Kaufman ve Rousseeuw; 1990; Kleinberg, 2002).

K-Ortalamalar Kümeleme Analizi

Modelin tanımı. K-Ortalamalar algoritması, Yale Üniversitesi'nden J. A. Hartigan ve M. A. Wong (1979) tarafından geliştirilmiş olan bir bölümleme tekniğidir. Çok büyük veri setlerinden az sayıda küme elde etme konusunda en yararlı yöntemlerden biri olduğu düşünülmektedir. K-Ortalamalar Yöntemi genel olarak, veri setini meydana getiren nesnelerin, objelerin veya bireylerin bazı nitelik veya özelliklerine göre ''k'' sayıda gruba ayrılması şeklinde tanımlanmaktadır. Bu tanım kapsamında yer alan ''k'' pozitif bir tam sayıdır. K-Ortalamalar Yöntemi, genel anlamda en fazla kullanılan denetimsiz kümeleme algoritmasıdır. Dolayısıyla yöntem, en fazla kullanılan prototip tabanlı algoritmalardan bir tanesidir. Kolay uygulanabilmesi, çıktılarının kolay yorumlanması ve hızlı bir şekilde kümeleme analizinin gerçekleşmesi gibi özellikler K-Ortalamalar algoritmasının en popüler kümeleme algoritması olmasındaki başlıca faktörlerdir. Ayrıca normal dağılım varsayımını karşılamayan veri setleri söz konusu olduğunda diğer kümeleme yöntemleri gibi gayet dirençli bir kümeleme yöntemidir (Genolini ve Falissard, 2010; Kaufman ve Rousseeuw, 1990; Usami, 2014).

K-Ortalamalar algoritması kullanılarak gerçekleştirilen kümeleme analizi hiyerarşik kümeleme yaklaşımlarından daha büyük veri setleri üzerinde kullanılabilir. Bu duruma ek olarak, gözlemler bir kümeye kalıcı olarak bağlı değildir. Gözlemlerin kümelere olan aidiyeti belirlenmeye çalışılırken, algoritma gereği genel çözümler üretilir. Bununla birlikte tüm değişkenlerin sürekli olması gerekmektedir. K-Ortalamalar Yöntemi ile gerçekleştirilen bir kümeleme analizi sonucunda, bir küme içerisinde yer alan elemanlar arasındaki benzerlikler üst düzeyde iken, kümeler arası elemanlar arasındaki benzerlikler çok düşüktür.

Kümeleme süreci. K-Ortalamalar Yönteminin popüler olmasının en önemli sebebi, büyük veri setlerine uygulanabilmesidir. Kümeleme sürecinde araştırmacı, başlatma metodunu (initialization method) bir başka deyişle oluşturulacak olan küme sayısını kendisi belirlemektedir. K-Ortalamalar algoritması, veri setini bir dizi küme merkezi aracılığıyla ayırarak her bir gözlemi bir kümeye atar. Bu atama işlemi sonucunda yeni küme merkezleri belirlenir ve bu süreç devam ederek küresel

kümeler oluşturulur. K-Ortalamalar algoritmasının çalışma prensibinin adımları aşağıdaki gibidir:

1) K tane merkez (centroid) seçilir (rastgele seçilen K sayıda satır).

2) Her bir veri noktası en yakın merkeze atanır.

3) Bir kümede yer alan tüm veri noktalarının ortalaması alınarak merkezler yeniden hesaplanır.

4) Veri noktaları kendisine en yakın merkezlere atanır.

5) Gözlemler atanana kadar veya maksimum iterasyon sayısına ulaşılana kadar (R programında varsayılan iterasyon sayısı 10'dur) üçüncü ve dördüncü adımlar devam eder.

İdeal küme sayısının belirlenmesi. K-Ortalamalar Yöntemi söz konusu olduğunda, araştırmacı tarafından analiz öncesinde belirlenen ''k'' değeri analizin performansını etkileyebilmektedir. Bu sebeple, daha önceden belirlenmiş bir ''k'' değeri kullanmak yerine, farklı k değerleri için elde edilen sonuçların karşılaştırılması benimsenebilir. Veri setine ilişkin belirli özelliklerin yansıtılması hususunda makul ''k'' değerleri kullanmak çok önemlidir. Bu doğrultuda, seçilen ''k'' değerleri veri setindeki eleman sayısından önemli ölçüde küçük olmalıdır.

İstatistiksel yazılımların birçoğu K-Ortalamalar Yöntemi için kullanıcı tarafından belirlenecek olan bir küme sayısının daha önceden belirlenmesi prensibine göre çalışır. Yukarıda da bahsedildiği gibi, tatmin edici bir kümelenme sonucuna ulaşmak için genellikle araştırmacının farklı k değerleri ile algoritmayı çalıştırdığı bir dizi yineleme gerekmektedir. Kümeleme sonucunun geçerliği, geçerliğe ilişkin herhangi bir istatistiksel ölçü elde edilmezse sadece görsel anlamda değerlendirilebilir. Fakat geçerlik anlamındaki bu görsel değerlendirme ile araştırmacıların çok boyutlu veri setleri için kümeleme sonuçlarını değerlendirmek çok zordur.

İdeal k değerini belirlemek için istatistiksel bazı ölçütler söz konusudur. Bu ölçüler, genellikle olasılıkçı kümeleme yaklaşımlarının (şans faktörünün söz konusu olduğu kümeleme yaklaşımları) kombinasyonu şeklinde uygulanmaktadır ve veri setlerine ilişkin dağılımlar hakkındaki bazı varsayımlarla hesaplanmaktadır. Bir takım Gauss dağılımları kullanılarak oluşturulan ve aynı zamanda İki Aşamalı Kümeleme Analizinde de geçerliğe ilişkin delil olarak kullanılabilen Bayes Bilgi Kriteri (BBK) veya

Bayes Akaike Bilgi Kriteri (ABK) istatistikleri, her bir farklı küme çözümü için ayrı ayrı hesaplanan istatistiklerdir. Her iki istatistik, log-olabilirlik fonksiyonunu temel almaktadır. ABK/BBK oranının en düşük olduğu noktada ideal küme sayısına ulaşıldığı söylenebilir (Burnham ve Anderson, 2002; Hastie, Tibshirani ve Friedman, 2009)

Yukarıda bahsedilen yöntemlerin dışında, veri seti elemanlarının benzerliğini temel alan ve uyuşma (cohesion) başlığı altında incelenen gruplar içi kareler toplamı ile kümelerin ayrımına ilişkin bir ölçü olan ve ayrışma (seperation) başlığı altında incelenen kümeler arası kareler toplamı yöntemleri, ideal küme sayısını belirlemede kullanılan yaygın ve teknik olarak kuvvetli olan yöntemlerdendir. Gruplar içi kareler toplamı, her kümedeki varyans miktarına ilişkin bilgi vermektedir. Bu değer ne kadar düşük olursa, bölümleme (partioning) o derece iyi olmakta; gruplar içi kareler toplamına ilişkin değer azaldığında kümeleme performansı artmaktadır. Kısacası yöntem, bir kümenin içindeki noktaların birbirine olabildiğince yakın olması gerektiği prensibini temel almaktadır. İdeal küme sayısı (k) arttıkça bu değer monoton olarak azalacaktır. Bu nedenle, eğrinin düzleştiği (artış oranında önemli bir düşüşü gösteren "dirsek" yöntemi) aralıktan optimal bir k değeri seçilir. Dirsek noktasına karşılık gelen kümelerin sayısının seçilmesi, çok fazla küme olmaksızın makul bir performans sağlar. Bu durumda, bu yöntem ideal küme sayısına ulaşılmasına yardımcı olmaktadır. Kümeler arası kareler toplamı yöntemi ise prensip olarak gruplar içi kareler toplamı yöntemine zıt çalışmaktadır. İdeal olarak, kümeleme analizi sonucu oluşan kümeler birbirinden iyi bir şekilde ayrılmalıdır, yani kümeler arası kareler toplamına ilişkin değer ne kadar fazla olursa, kümeler birbirinden o derece iyi ayrılmış anlamına gelmektedir. Kümeler arası kareler toplamına ilişkin görsel yorumlanırken gruplar içi kareler toplamına ilişkin görselin yorumlanmasında kullanılan dirsek yöntemindan yararlanılmaktadır (Fovell ve Fovell, 1993; Hartigan ve Wong, 1979; Macqueen, 1967).

İki Aşamalı Kümeleme Analizi

Modelin tanımı. SPSS 11.5 ve sonraki sürümler, İki Aşamalı Kümeleme Analizi adı altında yeni bir yöntem sunmaktadır. İki Aşamalı Kümeleme Analizi, büyük veri setleri söz konusu olduğunda rahatlıkla kullanılabilen SPSS programına ilişkin bir kümeleme yöntemidir. Analiz, genellikle veri seti çok büyük olduğunda hiyerarşik kümeleme analizi ve K-Ortalamalar Yöntemine alternatif olarak kullanılmaktadır

(Garson, 2014; Norusis, 2010). Veri setinde yer alan değişkenler hem kategorik hem de sürekli değişkenler olduğunda kullanılabilir. Kategorik ve sürekli değişkenlerle ayrı ayrı ve birlikte kullanılabilmesi, otomatik bir şekilde en ideal küme sayısının belirlenmesi ve analiz sonucunda gözlenen kümelerle uyum sağlamayan gözlemlerin istendiğinde veri setinden ayıklanabilmesi İki Aşamalı Kümeleme Analizinin en önemli özellikleridir. Analiz süreci, ön kümeleme aşaması ve kümeleme aşaması olmak üzere iki adımdan meydana gelmektedir (Garson, 2014; Tkaczynski, Rundle-Thiele, Zhang, Ramakrishnon ve Livny, 1996).

Ön kümeleme aşamasında gözlemlerin küçük alt kümeler halinde ilk kümeleme işlemi gerçekleştirilmektedir ve bu alt kümeler daha sonra ayrı gözlemler olarak ele alınmaktadır. Gözlemin önceden oluşturulmuş kümelenme ile birleştirilip birleştirilmediği ya da yeni bir kümelenme oluşturulup oluşturulmayacağına karar verilir. Bu yeni gözlemlerin gruplandırılması işlemi uzaklık kriteri göz önünde bulundurularak hiyerarşik kümeleme yöntemiyle gerçekleştirilmektedir. İki Aşamalı Kümeleme Analizinde kullanılan algoritma ile, küme sayısı belirlenebilmekte veya daha önce atanabilecek küme sayısı bulunabilmektedir. İkinci adım, ön kümeleme sonucu elde edilen alt kümelerin analizin temelini oluşturduğu ve alt kümelerin gerekli sayıda kümeye ayrıldığı yönlendirme işlemidir. Alt kümelerin sayısı gözlem sayısından önemli ölçüde daha küçük olduğu için, geleneksel gruplama yöntemlerinin kullanımı daha kolaydır. Alt küme sayısı ne kadar fazla ise, yöntem o derece hassastır (Cameron ve Miller, 2015; Tkaczynski ve ark., 2010; Zhang ve ark. 1996).

İki Aşamalı Kümeleme Analizinde bir veya daha fazla değişken kategorik ise gözlemlerin bu ölçümün en yüksek değerlerine sahip kümede gruplandırılacağı şekilde log-olabilirlik uzaklık ölçüsü kullanılır. Tüm değişkenler sürekli ise, Öklid mesafesi kullanılır ve böylece gözlemler en küçük Öklid mesafesine sahip kümede gruplanır. Log-olabilirlik yöntemi kategorik ve sürekli değişkenlerle uyumlu olduğundan SPSS algoritması, mesafe ölçütü olarak kümeleri birleştirmek için log-olabilirlik mesafe ölçüsünde bir azalma kullanır. Log-olabilirlik mesafesi ölçümünü kullanan İki Aşamalı Kümeleme Analizi süreci, sürekli değişkenler için normal dağılımı; kategorik değişkenler için ise çoklu normal dağımı gerektirir. Fakat normallik varsayımı karşılanmasa bile analiz iyi sonuçlar vermektedir (Amprik testler, analizin hem bağımsızlık hem de normallik varsayımına karşı oldukça sağlam olduğunu göstermektedir). Analize ilişkin tek ve en önemli varsayım örneklemin büyük

olmasıdır (n>200). Yani İki Aşamalı Kümeleme Analizi homojen olmayan çok büyük veri setlerine uygulanabilmektedir (Garson, 2014; Cameron ve Miller, 2015; Norusis, 2010). Bir diğer mesafe ölçüsü ise Öklid uzaklığıdır. Mesafe ölçüsü olarak Öklid uzaklığı tüm değişkenlerin sürekliği olduğu durumlarda kullanılmaktadır. İki nokta arasındaki Öklid mesafesi açıkça tanımlanmıştır. İki küme arasındaki mesafe, onların merkezleri arasındaki Öklid mesafesi ile; kümelerin merkezi ise belirli bir kümelenme için tüm değişkenlerden oluşan vektör olarak tanımlanır. İki Aşamalı Kümeleme Analizine ilişkin süreç, ilk kümenin oluşturulması ile başlar. Bu adımda sıralı kümeleme yöntemi kullanır. Gözlemler analiz edilir ve verilen gözlemin yeni bir kümelenme oluşturup oluşturmayacağına karar verilir. Bu karar, mesafe kriterlerine dayanmaktadır (Cameron ve Miller, 2015; Garson, 2014; Rundle-Thiele, S.,Kubacki, K., Tkaczynski, A., Parkinson, J., 2015).

Sonuç olarak İki Aşamalı Kümeleme Analizi Yönteminin en önemli özellikleri, hem sürekli hem de kategorik verilerin birlikte analiz edilebilmesi, büyük veri setlerinin bu yöntem ile analiz edilebilmesi ve bu tür verilerin işlenmesi için gereken süre bakımından, bu yöntem ile diğer yöntemlere göre daha kısa sürede analizin gerçekleştirilebilmesidir. İki Aşamalı Kümeleme Analizi hibrit bir yöntemdir. Bu yöntemin temel avantajları, Ward'ın minimum varyans yöntemi ile K-Ortalamalar Yönteminin gerektirdiği küme sayısının hesaplanması ve karma ölçekli veri setleri için kullanılabilmesidir (Kuo, Ho ve Hu, 2002). Yöntemin kayıp değerleri olan ögeleri analiz için dikkate almaması İki Aşamalı Kümeleme Analizine ilişkin bir dezavantaj olarak söylenebilir.

İdeal küme sayısının belirlenmesi. İki Aşamalı Kümeleme Analizinde küme sayısının otomatik olarak belirlenmesi için, hiyerarşik kümeleme analizi ile uyumlu iki aşamalı prosedür geliştirilmiştir. İlk adımda, BBK veya ABK istatistikleri, farklı sayıda kümeyle her bir farklı küme çözümü için hesaplanır. İkinci adımda ilk tahmin, hiyerarşik kümelenmelerdeki her bir aşamada en yakın iki küme arasındaki en fazla mesafe artışının bulunmasıyla geliştirilir. SPSS, en ideal küme sayısını göstermenin yanında bir de küme sayısının geçerliği için siluet katsayısına ilişkin bilgi vermektedir (Garson, 2014; Kayri, 2007; Zhang, Ramakrishnon ve Livny, 1996).

Önem düzeyi. Önem düzeyi, veri setinde yer alan her bir değişkenin oluşturulan model üzerindeki etkisine göre istatistiksel anlamlılığını temsil etmektedir. Önem ölçütü aslında, yordayıcıların modele yaptığı katkıya dayalı olarak her bir

tahmin edicinin kümeleme analizinde ne derece katkı sağladığına ilişkin bir sıralamadır. Ölçüt, veri madencilerinin modele hiçbir şekilde katkıda bulunmayan sadece analiz sürecini uzatan değişkenlerin belirlenmesi konusunda bilgilendirilmesine yardımcı olur. Kümeleri oluşturan değişkenlerin göreli katkısı (önemi), her iki değişken türü (sürekli ve kategorik) için ayrı ayrı hesaplanmaktadır. Önem değerleri 0-1 arasında derecelendirilmektedir. 0 kümeleri belirlemede en önemsiz değişkeni ve 1 ise son derece önemli değişkeni ifade etmektedir. Önem düzeyi formülsel anlamda, sürekli değişkenler söz konusu ise t testine; kategorik değişkenler söz konusu ise ki-kare anlamlılık testine dayanmaktadır (Ceylan ve diğ., 2017; Garson, 2014; Kayri, 2007).

İlgili Araştırmalar

Çalışmanın bu bölümünde ilgili araştırmalara yer verilerek elde edilen sonuçlar bir bütün olarak değerlendirilmiştir.

Kohonen'in öz örgütlemeli harita yöntemi ile ilgili araştırmalar. Kiang (2001), Kohonen'in Öz Örgütlemeli Harita Yöntemi ile yığınsal hiyerarşik kümeleme yöntemi sonuçlarını karşılaştırmıştır. Çalışma kapsamında, Kohonen'in Öz Örgütlemeli Harita Yöntemi ile elde edilen sonuçlar yorumlanırken uzman bilgisinin sürece dahil edilmesinin avantajlı olduğu, Kohonen'in Öz Örgütlemeli Harita Yöntemi ile hem iki kategorili hem de sürekli değişkenlerin girdi olarak sorunsuz bir şekilde kullanılabildiği sonucuna ulaşılmıştır.

Oğuzlar (2005), Bursa Emniyet Müdürlüğünden alınan veriler ile suçlu profilinin belirlenmesi amacıyla Kohonen'in Öz Örgütlemeli Harita Yöntemini kullanmıştır. Çalışma kapsamında, parametrik olmayan ve yeni bir yaklaşım olan Kohonen'in Öz Örgütlemeli Harita Yöntemi ile suçlu profillerine ilişkin tanımlamalar C5.0 kural algoritması kullanılarak gerçekleştirilmiş ve 12 ayrı kümenin oluştuğu sonucuna ulaşılmıştır.

Taşkın ve Emel (2010) çalışmalarında Kohonen ağları ile parakendecilik sektöründe bir kümeleme uygulaması gerçekleştirmişlerdir. Bir işletmenin 10000 adet müşterisine ait veri tabanı kullanılmış ve Kohonen' Öz Örgütlemeli Harita Yöntemi tekniği ile kümeleme gerçekleştirilmiştir.

Özşahin ve Yüreğir (2012) Türkiye'de otomotiv sektöründe faaliyet gösteren firmaları, bilanço ve gelir tablolarından elde edilen finansal oranları kullanmak suretiyle

Kohonen'in Öz Örgütlemeli Harita Yöntemi ile kümelere ayırmıştır. Çalışmada, otomotiv sektörü içerisinde yer alan işletmelerin finansal başarısını ve ihracat durumunu hangi faktörlerin ne düzeyde etkilediğini belirlemek için ısı haritalarından yararlanılmıştır.

İnce, İmamoğlu ve Keskin (2013) çalışmalarında tüketici profilleme çalışması gerçekleştirmişlerdir. Çalışmalarında Kohonen'in Öz Örgütlemeli Harita ve K-Ortalamalar Yöntemini kullanmak suretiyle, tüketicilerin alışveriş motivasyonu ile birlikte karar verme stillerine dayalı tüketici profili çıkarılmıştır. Toplamda 1459 adet müşteriyle anket gerçekleştirilmiş ve tüketici profilleri oluşturulmuştur. Sonuçta Kohonen'in Öz Örgütlemeli Harita Yönteminin, K-Ortalamalar Yöntemine göre daha üstün sonuçlar ortaya çıkardığı raporlanmıştır.

Özçalıcı (2016) yaptığı çalışmada, BIST 50 Endeksinde listelenen hisse senetlerini kümelere ayırmak amacıyla Kohonen'in Öz Örgütlemeli Harita Yöntemini kullanmıştır. Çalışmada etkin portföy oluşturma problemi üzerinde durulmuş ve Kohonen'in Öz Örgütlemeli Harita yönteminin birbirlerine benzeyen hisse senetlerini aynı kümede, birbirlerine benzemeyen hisse senetlerini de farklı kümelerde toplayabildiği ifade edilmiştir. Çalışmada ayrıca diğer görsel ve istatistiksel yöntemler kullanmak suretiyle Kohonen'in Öz Örgütlemeli Harita Yönteminin çok daha başarılı bir kümeleme gerçekleştirildiği raporlanmaktadır.

Qiao ve Jiao (2018) çalışmalarında, PISA 2012 Amerika Birleşik Devletleri örnekleminden elde ettikleri 426 kişilik çalışma grubuna ilişkin verileri kullanmışlardır. 11 değişkenin girdi olarak kullanıldığı çalışma sonuçlarına göre, 0.84 kappa istatistiği ile K-Ortalamalar Yöntemi ile elde edilen ideal küme sayısının beş; 0.96 kappa istatistiği ile Kohonen'in Öz Örgütlemeli Harita Yöntemi ile elde edilen ideal küme sayısının dokuz olduğu sonucuna ulaşılmıştır. Ayrıca çalışma sonucuna göre, her iki yönteme ilişkin doğru sınıflandırma yüzdesinin de tatmin edici ve birbirine yakın olduğu belirlenmiştir.

K-ortalamalar kümeleme analizi ile ilgili araştırmalar. Çakmak (1999), kümeleme analizini genel olarak incelemiş daha sonra kümeleme sonuçlarını bir geçerlik problemi olarak ele almış ve geçerlik tekniklerinden bazılarını gözden geçirmiştir. Uygulama bölümünde aşamalı kümeleme yöntemleri yardımıyla oluşturulabilecek küme sayıları belirlenmiş ve eğitim yapıları birbirine benzeyen iller,

farklı küme sayıları için aşamalı olmayan kümeleme tekniklerinden K-Ortalamalar Yöntemiyle kümelendirilmiştir. Elde edilen kümelerin geçerliliğini test etmek amacıyla kümeleme sonuçlarına diskriminant analizi uygulanmış ve iller yeniden sınıflandırılmıştır. Sonuç olarak, oldukça yüksek doğru sınıflandırma oranları bulunmuş ve K-Ortalamalar Yöntemiyle elde edilen kümelerin anlamlı olduğu sonucuna varılmıştır.

Ersöz (2009), OECD ülkelerine ilişkin sağlık verilerini girdi değişkenleri olarak kullandığı çalışma kapsamında K-Ortalamalar, hiyerarşik kümeleme ve k-medoid (noktaların merkeze olan uzaklıklarının temel alındığı kümeleme yöntemi) yöntemine ilişkin sonuçları karşılaştırmıştır. Çalışma sonucunda, Türkiye'nin her üç kümeleme yöntemi sonucunda da Meksika ile aynı kümede yer aldığı gözlemlenmiştir.

Şekerkaya ve Cengiz (2010) ise çalışmalarında kadın tüketicilerin alışveriş merkezi tercihlerine göre kümelenmesi problemi üzerinde durmaktadırlar. 304 adet AVM müşterisine anket uygulamışlardır ve kümeleme aracı olarak K-Ortalamalar Yöntemini kullanmışlardır. Araştırma sonucuna göre kadınların AVM tercihlerine göre üç grupta kümelenmiş ve bu kümeler sahip oldukları nitelikler itibariyle potansiyeller, aktifler ve duyarsızlar olarak adlandırılmıştır.

Acar (2012), tarafından yapılan çalışmada PISA 2009 sonuçlarına göre Türkiye'nin OECD'ye üye ve aday ülkeler arasındaki yeri K-Ortalamalar Yöntemi ve ayırma analiziyle belirlenmeye çalışılmıştır. Matematik, Fen Bilimleri ve Okuma Yeterliği değişkenlerinin ele alındığı çalışmanın örneklemini 2009 yılında PISA uygulamasına katılan 65 ülkeden toplam 475.460 öğrenci oluşturmuştur. Kümeleme analizi sonuçlarına göre; 1. kümede dokuzu aday toplam 13 ülkenin, 2. kümede beşi OECD'ye aday toplam 30 ülkenin; 3.kümede beşi aday toplam 10 ülkenin ve 4. kümede hepsi aday toplam 12 ülkenin sınıflandığı görülmüştür. Çalışmada ayırma analizine göre doğru sınıflama yüzdesinin %96,9 oranında olduğu bulunmuştur.

Antonenko, Toy, Niederhauser (2012), öğrencilerin çevrimiçi bir öğrenme ortamında problem çözme aktivitesine katılırken öğrenme davranışının özelliklerini analiz etmek için hiyerarşik bir kümeleme yöntemi (Ward kümelenmesi) ve hiyerarşik olmayan bir kümeleme yöntemi (k-ortalamalar) kullanmıştır. Çalışma sonucunda, K-Ortalamalar Yönteminin büyük örneklemlerde rahatlıkla kullanılabileceği sonucuna ulaşılmıştır.

DeFreitas, Benard (2015), öncelikle eğitimde kümeleme analizi üzerine yapılan araştırmaları inceleyerek kullanılan algoritmaları belirlemiştir. Daha sonra, Öğrenme Yönetim Sistemi (ÖYS) günlük verileriyle kümelenme algoritmalarının göreceli performansını göstermek için bir vaka tabanlı deney sunmuştur. ÖYS içerisinde kümeleme analizi yapmak için ve hangi tekniğin en uygun olduğunu belirlemek için bölüm tabanlı (K-Ortalamalar), yoğunluk tabanlı (DBSCAN) ve hiyerarşik (BIRCH) yöntemleri karşılaştırmıştır. Bölüm tabanlı metotların en yüksek Siluet Katsayısı değerlerini ürettiği ve kümeler arasında daha iyi dağılım gösterdiğini sonucuna varılmıştır. Sonuç olarak BIRCH algoritmasının ayrıca oldukça iyi bir performans sergilemekte ve algoritma küme sayısının önsel tanımlamasını gerektirmemesi sebebiyle yeni veri kümelerinde küme gruplarını bulmak için iyi bir başlangıç noktası olarak işlev gördüğü sonucuna ulaşılmıştır.

Hamalainen, Kumpulainen, Mozgovoy (2015) kümeleme yöntemlerini karşılaştırmış ve sonuç olarak kümeleme konusunda “En iyi yöntem şudur” vb. bir kuralın olamayacağını vurgulamışlardır. Kümeleme yöntemlerinin birbirine göre üstünlüklerinin olduğu, prensip olarak, hibrit yöntemlerinin çoğu zaman en bilgilendirici ve çekici kümeleme modellerini ürettiği, ancak kullanımlarını sınırlayan pratik problemlerin olduğu belirtilmiştir. Katı (strict) kümeleme yöntemleri arasında, yoğunluğa dayalı yöntemler, spektral kümeleme, kernel k-ortalamalar ve hiyerarşik CHAMELEON algoritmasının en umut verici görünen algoritmalar olduğu; K-Ortalamalar Yönteminin EVM’de, tipik eğitimsel veriler için en uygun yöntemlerden biri olduğu ve bu sebepten ötürü çok popüler olduğu vurgulanmıştır. K-Ortalamalar Yönteminin diğer alanlarda, etkinliği ve belki de daha kolay tespit edilebilir kümeleri nedeniyle popülerliğinin daha anlaşılabilir bir durumda olduğu sonucuna varmışlardır.

Atalay ve Öztürk (2016), eğitim durumları itibariyle benzer özellikler gösteren illerin hangileri olduğunu incelemiştir. Araştırma kapsamında kullanılan veriler, Türkiye İstatistik Kurumu (TÜİK)’nun, adrese dayalı altı yaş üstü nüfus verileri olup, 2010 yılına aittir. Türkiye’deki illerin eğitim durumlarını gösteren değişkenler belirlenmiş ve bu değişkenlerle kümeleme analizi yapılmıştır. Analiz safhasında K-Ortalamalar Yöntemi kullanılmış ve uygun küme sayısının altı olduğu belirlenmiştir.

Aksu, Güzeller ve Eser (2017), PISA 2012 öğrenci anketi kapsamında bulunan öz yeterlik, ilgi ve tutum ortalama puanlarını göz önünde bulundurarak PISA 2012 katılımcı ülkeleri içerisinden 43 ülkenin nasıl kümelendiğini incelemiştir. Çalışma

kapsamında küme dağılımlarını inceleme anlamında hiyerarşik kümeleme yöntemi kullanılmıştır. Hiyerarşik kümeleme sonucu elde edilen kümelerin geçerliğine ilişkin kanıt sunmak için ise K-Ortalamalar ve diskriminant analizi yöntemi kullanılmıştır. Çalışmanın sonuçları incelendiğinde, öz yeterlik puanlarına göre sekiz, ilgi puanlarına göre yedi, tutum puanlarına göre ise altı farklı küme oluştuğu belirlenmiştir. Alt boyutlara ilişkin oluşan kümeler incelendiğinde, öz yeterlik puanları göz önünde bulundurulduğunda Japonya ve Şangay'ın tek başına birer küme, ilgi puanları göz önünde bulundurulduğunda Romanya'nın tek başına bir küme, tutum puanları göz önünde bulundurulduğunda ise Norveç ve Danimarka ile Japonya ve Kore'nin tek başına birer küme oluşturduğu belirlenmiştir.

Navarro ve Ger (2018), kümelemede model oluşturmada iç geçerlik ve kararlılık anlamında en başarılı algoritmaları belirlemeye çalışmıştır. İç geçerlik ve kararlılık ölçümlerine göre hangi algoritmaların daha iyi performans gösterdiğini belirlemek için yedi farklı algoritmanın performansı karşılaştırılmış, K-Ortalamalar ve PAM'in (partition around medoids) bölüm algoritmaları arasında en iyi performansı gösterdiği ve DIANA'nın (Divisive Analysis) ise hiyerarşik algoritmalar arasında en iyi performansı gösterdiği belirlenmiştir.

İki aşamalı kümeleme analizi ile ilgili araştırmalar. Kayri (2007), İki Aşamalı Kümeleme Analizini kullandığı çalışmada İki Aşamalı Kümeleme Analizinin avantajı ve dezavantajlarını belirlemeye çalışmıştır. Araştırma sonuçlarına göre İki Aşamalı Kümeleme Analizinin ideal küme sayısı konusunda bilgi verdiği ve analiz ile optimal alt popülasyon sayısının belirlenebildiği sonucuna ulaşılmıştır. Çalışma kapsamında kümeleme analizi teknikleri log-olabilirlik uzaklık ölçütüne göre tanımlanmış ve bu çalışma log-olabilirlik ölçütünü kullanarak grupların nasıl oluşturulabileceği konusunda bir örnek teşkil etmiştir. Ayrıca ABK ve BBK temel alınarak ideal küme sayısı konusunda karşılaştırma yapılmıştır. Bu çalışmanın sonunda, değişkenlerin benzerliğine göre BBK kullanılarak yedi küme belirlenmiştir. En ideal küme sayısını elde etmek için BBK'nın ABK karşısında kullanılabileceği sonucuna varılmıştır. Ayrıca İki Aşamalı Kümeleme Analizinde hem sürekli hem de kategorik değişkenlerin kullanılabileceği sonucuna varılmıştır.

Yıldırım ve Akın (2009), veri madenciliğinde önemli bir yere sahip olan kümeleme yöntemlerini karşılaştırmış ve İstanbul'da öğrenim gören 3468 ortaöğretim öğrencisinden elde edilen verileri kullanmıştır. Çalışma kapsamında, öğrencilerin

şiddet eğilimlerine göre gruplandırılması ve bu eğilime yol açan sebeplerin ortaya çıkarılması amaçlanmıştır. Merkeze dayalı bölümleyici ve hiyerarşik kümeleme yöntemlerinden; K-Ortalama, İki Aşamalı Kümeleme ve CLARA (Clustering Large Applications) ile elde edilen kümelerin karşılaştırıldığı çalışmada ve bu yöntemlerin üstün ve zayıf yönlerini incelemiştir.

Yılmaz (2012) çalışmasında üniversite öğrencilerinin eğlence ya da iletişim amacıyla, internet kullanımına göre profillerini belirlemek ve internetteki ilgisine bağlı olarak profillerinin farklı olup olmadığı amacıyla İki Aşamalı Kümeleme Analizini kullanmıştır. Çalışmanın örneklemini 358 üniversite öğrencisi oluşturmuştur. Çalışma kapsamında, üniversite öğrencilerinin eğlence ve iletişim amacıyla internet kullanımı anlamında iki kümeye ayrıştığı ve İnternet'e olan ilginin bu bölünme üzerinde büyük etkiye sahip olduğu sonucuna ulaşılmıştır. Aynı zamanda çalışma sonucuna göre ilk küme çoğunlukla erkek, interneti yoğun kullanan ve internete büyük önem veren öğrencilerden; ikinci kümenin ise interneti daha az kullanan ve internete orta düzeyde önem veren öğrencilerden oluştuğu sonucuna ulaşılmıştır.

Arı, Özköse ve Calp (2016) çalışmasında, Borsa İstanbul'da faaliyet gösteren 90 firma, finansal tablolarından elde edilen bilgileri kullanmak suretiyle kümelere ayrılmıştır. Çalışmalarında İki Aşamalı Kümeleme Yöntemini kullanmışlardır. Çalışma kapsamında uygulanan üç farklı analizden ilk uygulamada 12 faktör ve 90 birimden oluşan matris İki Aşamalı Kümeleme Analizine alınmış, sonuçta küme kalitesi orta derecede olan iki adet küme elde edilmiştir. İkinci uygulamada veri seti varyans analizine tabi tutularak faktörlerden birimler için istatistiksel olarak anlamlı farklılık arz etmeyen 5 faktör elenmiş, İki Aşamalı Kümeleme Analizi elde kalan 7 faktör üzerinden yapılmıştır. Burada küme kalitesi oldukça yüksek olan yine iki küme elde edilmiştir. Üçüncü uygulamada ise daha önceki uygulamalarda eleman sayısı çok yüksek olan küme ayrıştırmak istenmiş, sonuçta 3 kümeli ancak kalitesi biraz daha düşük bir sonuç elde edilmiştir.

Önen (2018), TIMSS-2015 uygulamasını göz önünde bulundurarak matematik başarısı üzerinde etkisi olduğu düşünülen öğrenci ve öğretmene ilişkin nitelikler ile öğretimsel nitelikler açısından dördüncü ve sekizinci sınıf öğrencilerini kümelere ayırmış ve her bir kümeye ilişkin bir öğrenci profili belirlenmiştir. Çalışma kapsamında İki Aşamalı Kümeleme Analizi kullanılmıştır. Gerçekleştirilen kümeleme analizi sonucunda dördüncü sınıf düzeyinde üç küme, sekizinci sınıf düzeyinde ise iki

kümenin ortaya çıktığı göze çarpmaktadır. Dördüncü sınıf düzeyindeki kümelerin oluşmasında matematik başarısı, öğrenci ve öğretmen niteliklerinin kümelerin oluşmasında en etkili olan özellikler olduğu belirlenmiştir. Kümeleme işleminde en az etkisi olan değişkenin öğretmen beyanı göz önünde bulundurularak belirlenen öğretimsel nitelikler olduğu görülmüştür. Sekizinci sınıf düzeyinde ortaya çıkan iki kümenin oluşmasındaki en önemli özelliklerin ise öğrenci niteliklerinin; matematik başarısı, öğretmen nitelikleri ile öğretimsel niteliklerin ise kümelemede düşük düzeyde etkili olduğu görülmüştür. Hem dördüncü hem de sekizinci sınıf düzeyi için matematik dersinde en başarılı olan öğrencilerin matematik dersi için kendine güven düzeyi çok düşük, matematik öğrenmeyi seven, matematik dersine ilişkin öğretimin ilgi çekici olduğunu düşünen, okul anlamındaki aidiyet hissi yüksek ve akran baskısına çok az maruz kalan öğrencileri olduğu göze çarpmaktadır. Hem dördüncü hem de sekizinci sınıf düzeyi için matematik dersi anlamında başarı düzeyi düşük öğrencilerin öğrenmekten hoşlanmayan, matematik dersine ilişkin öğretimin dikkat çekici olmadığımı düşünen, okul için aidiyet hissi düşük düzeyde ve akran baskısı ile karşı karşıya kalan öğrenciler olduğu görülmüştür.

Tekin (2018), üç farklı kümeleme analizi yöntemini kullanarak Borsa İstanbul'da işlem gören hisse senetlerinden etkin bir portföy oluşturulmasını amaçlamıştır. Aynı zamanda çalışmada hisse senetlerinden etkin bir portföy oluşturmada kümeleme analizi yöntemlerinin kullanılabilirliği sınanmıştır. Çalışma kapsamında hiyerarşik kümeleme yöntemlerinden Ward yöntemi, hiyerarşik olmayan kümeleme yöntemlerinden K-Ortalamalar ve İki Aşamalı Kümeleme Yöntemi kullanılarak toplam 69 adet hisse senedi kümelenmiştir. Kümeleme analizinde kullanılan finansal göstergeler şirketlerin finansal tablolarından ve hisse senedi fiyat hareketlerinden elde edilmiştir. Çalışma sonucunda her üç yönteme göre oluşan kümelerin genel itibariyle benzer şekillendiği sonucuna ulaşılmıştır.

İlgili araştırmalar bir bütün olarak değerlendirildiğinde K-Ortalamalar Yönteminin çalışmalarda sıklıkla kullanıldığı, İki Aşamalı Kümeleme Analizinin ve Kohenen'in Öz Örgütlemeli Harita Yönteminin ise eğitim bilimleri alanında yapılan çalışmalar kapsamında çok fazla kullanılmadığı sonucuna ulaşılmıştır. Bunun yanında ilgili araştırmalarda belirli sayıdaki çalışmada sadece iki farklı yöntemin sonuçları incelenip karşılaştırılırken daha fazla sayıda yöntemin sonuçlarının incelendiği ve karşılaştırıldığı bir araştırma bulgusuna rastlanamamıştır. Bunlara ek olarak

kümeleme analizlerinin çoğunlukla Matlab ve SPSS Clemente programlarında gerçekleştirildiği, R programı ile kümeleme analizi yapılmadığı belirlenmiştir. İlgili araştırmalar değerlendirildiğinde kümeleme analizinden elde edilen sonuçların ne düzeyde geçerli olduğuna ilişkin delil sunmak amacıyla farklı ölçütlere göre farklı kümeleme yöntemlerini inceleyen çalışmaların olmadığı belirlenmiştir. Bu nedenle ilgili alan yazına katkı sağlamak amacıyla farklı kümeleme yöntemlerinden elde edilen sonuçların özellikle PISA gibi geniş ölçekli bir sınavdan elde edilen büyük veriler yardımıyla incelenmesi gerektiği sonucuna ulaşılmıştır. Bu sayede alanda çalışma yapacak araştırmacıların farklı kümeleme yöntemlerinin birbirlerine göre üstün ve zayıf yönlerini görmeleri ve araştırmalarına bu yönde şekil vermeleri önemli görülmektedir. Bu sayede araştırmacıların kümeleme analizinde kullanbilecekleri yöntemlerin çeşitliliği konusunda bilgi sahibi olacakları düşünülmektedir.

Bölüm 3
Yöntem

Bu bölümde araştırmanın yöntemi, çalışma grubu, veri toplama araçları, verilerin toplanması ve analizine yer verilmektedir.

Araştırmanın Türü

Bu çalışmada, 2015 PISA'ya katılım gösteren OECD ülkeleri farklı kümeleme yöntemleri ile modellenmiş ve oluşturulan modeller incelenmiştir Araştırma kapsamında kullanılan değişkenler PISA 2015 öğrenci anketinde yer alan fen öğretimine ilişkin maddeler ve öğrencilerin olası başarı puanı ortalamalarıdır. Araştırma, farklı kümeleme modellerinin incelenmesi amacıyla yapılmıştır. Çalışma bu yönüyle tarama modelinde betimsel araştırmalar altında değerlendirilebilir.

Çalışma Grubu

Araştırmanın amaçları doğrultusunda çalışma verisini OECD tarafından düzenlenen PISA sınavına katılan öğrencilerin hazır verileri oluşturmaktadır. Sınava 72 ülkeden toplam 540.000'e yakın öğrenci katılmıştır. Bu ülkelerden 35 tanesi OECD üyesi olup, bu 35 ülkenin öğrenci sayısı toplamda 253.140'tır. OECD ülkelerinden biri olan Slovenya'nın kayıp veri oranının kabul edilebilir düzeyin üzerinde olması sebebiyle toplam 34 ülke ile kümeleme analizi gerçekleştirilmiştir. Çalışma kapsamında veri kaynağı olarak kullanılan ülkelerin isimleri Tablo 1'de gösterilmiştir.

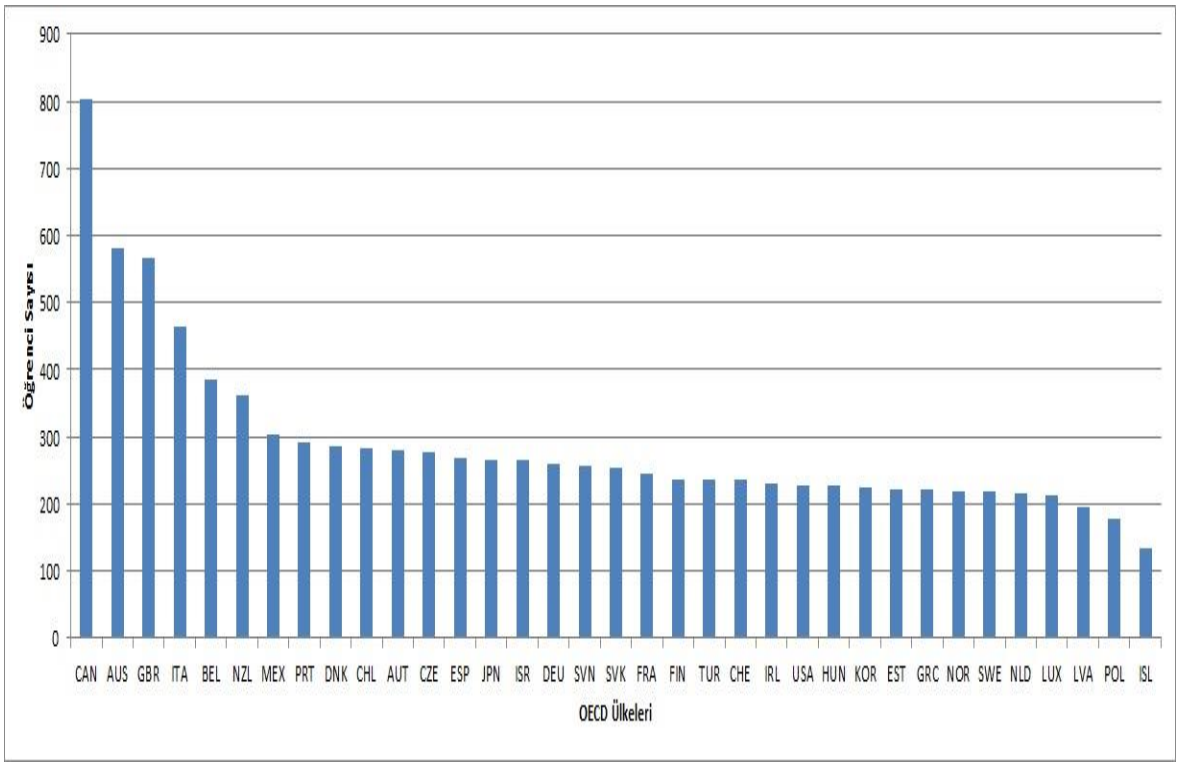
Tablo 1

Ülke Kodları

Sıra	Kod	Ülkeler	Sıra	Kod	Ülkeler
1	JPN	Japonya	18	ISR	İsrail
2	KOR	Kore	19	HUN	Macaristan
3	EST	Estonya	20	DNK	Danimarka
4	BEL	Belçika	21	CHE	İsviçre
5	CAN	Kanada	22	GRC	Yunanistan
6	AUT	Avusturya	23	GBR	İngiltere
7	CZE	Çek Cumhuriyeti	24	CHL	Şili
8	NOR	Norveç	25	SWE	İsveç
9	LUX	Lüksemburg	26	POL	Polonya
10	NLD	Hollanda	27	AUS	Avustralya
11	DEU	Almanya	28	USA	Amerika

12	SVK	Slovakya	29	LVA	Litvanya
13	ITA	İtalya	30	PRT	Portekiz
14	ESP	İspanya	31	NZL	Yeni Zelanda
15	ISL	İzlanda	32	IRL	İrlanda
16	FRA	Fransa	33	TUR	Türkiye
17	FIN	Finlandiya	34	MEX	Meksika

Çalışma kapsamında kümeleme analizine dâhil edilen 34 ülkenin öğrenci sayılarına ilişkin sütun grafiği Şekil 5'te gösterilmiştir.



Şekil 5. Öğrenci sayılarının ülkelere göre dağılımı

Ülkelerin katılımcı sayılarına ilişkin Şekil 5 incelendiğinde, en fazla katılımcının Kanada'dan, en az katılımcının ise İzlanda'dan olduğu göze çarpmaktadır. Katılımcı sayısı Kanada'dan sonra gözle görülür bir düşüş göstermekte ve Kanada'yı Avustralya ve İngiltere izlemektedir. En az katılımcının bulunduğu ülke olarak ise, İzlanda'yı sırasıyla Polonya ve Litvanya izlemektedir.

Araştırma kapsamında, verilerin analizine geçilmeden önce veriler, analize hazır hale getirilme amacıyla ''Veri Ön İşleme'' sürecine tabi tutulmuştur. Bu süreçte ilk olarak, kayıp veri analizi gerçekleştirilmiştir.

Kayıp veri analizi anlamında, çoklu değer atama yöntemi kullanılmıştır. Çoklu değer atama yöntemi, veri setinin yapısı gereği SPSS 20 programı tarafından lojistik

regresyon tabanlı olarak gerçekleştirilmiştir. Çoklu değer atama yöntemi, farklı yapılarda veri setlerine uygulanabilen bir veri atama yöntemidir. Yöntemin çalışma prensibi, tamamlanmamış veri setinin daha önceden belirlenmiş olan çoğaltma sayısı kadar çoğaltılarak, her bir kopyada kayıp değerler yerine olası değerlerin atanması şeklindedir. Çoklu değer atama yöntemi, yanlılığa ve uç değerlere karşı dirençli, geçerliği yüksek bir yöntemdir. Yöntem tesadüfi yapıda kayıp veriye sahip olan veri setlerine uygulanmaktadır (Bodner, 2008; Graham, 2009; Little ve Rubin, 1987). PISA anketlerine ilişkin veri setindeki kayıp verilerin rastlantısal bir yapıya sahip olması sebebi ile araştırma kapsamında kullanılan veri setine kayıp veri atama yöntemlerinden çoklu değer atama yöntemi uygulanmıştır (Adams, Lietz ve Berezner, 2013; Kaplan ve Su, 2016). Kayıp veri ataması sonucunda, araştırma kapsamında kullanılan değişkenler göz önünde bulundurulduğunda kayıp veri miktarı çok fazla olan Slovenya'ya ilişkin veriler veri setinden çıkartılmıştır.

Kayıp veri atama işleminden sonra veri setine sistematik örnekleme uygulanmıştır. Veri setine sistematik örnekleme uygulanmasının sebebi, PISA sınavında 253.140 öğrenciden elde edilen verilere R programı kapsamında kümeleme analizi yapmak için eldeki veri setinin çok büyük olması sebebiyle bilgisayarın hesaplama süresinin gecikmesi ve bu gecikmenin giderilebilmesi için yüksek işlemcili bilgisayarlara ihtiyaç duyulmasıdır. Program tarafından kümeleme analizi kapsamında en ideal sürede sonuç veren maksimum birey sayısının 10.000 ile sınırlı olması nedeniyle orantı sabiti k=25 (253.140/10.128) olarak belirlenmiştir. Excel programında macro oluşturularak evrende yer alan her 25 öğrenciden 1'i örnekleme alınmıştır. Bu işlemlerimn ardından çalışma kapsamında analize dâhil edilen öğrenci sayısı 10.128 olarak belirlenmiştir. Sistematik örnekleme sonucunda veri setinin 9870 öğrenciye ilişkin verilerden meydana geldiği görülmektedir. Sistematik örnekleme seçkisiz olmayan ve evren listesinden belli aralıklarla seçilen kişilerin yer aldığı örnekleme yöntemidir (Monette, Sullivan ve Jong, 1990).

Çalışmanın veri setini, PISA 2015 öğrenci anketinde yer alan fen öğretimine ilişkin maddelere verilen cevaplar oluşturmaktadır. Fen öğretimine ilişkin bu maddeler, öğretmen yönetimindeki fen bilgisi öğretimi, fen bilgisi öğretmenlerinden alınan geri bildirim, fen bilgisi derslerine ilişkin uyarlanabilir öğretim ve sorgulama temelli fen bilgisi öğretimi olmak üzere dört farklı kavramsal boyutta değerlendirilmektedir. PISA

öğrenci anketinde yer alan ve çalışma kapsamında veri toplama aracı olarak kullanılan alt testlerde yer alan maddelere ilişkin bilgiler Tablo 2'de gösterilmiştir.

Tablo 2

Maddelere İlişkin Bilgiler

Kavramsal Boyut	Kavramsal Alt Boyut	Madde Sayısı	Maddeler	İfadeler
Fen bilgisi öğretimi	Öğretmen yönetimindeki fen bilgisi öğretimi (Faktör 1)	4	ST103Q01NA	Öğretmen bilimsel fikirleri açıklar
			ST103Q03NA	Bütün sınıf öğretmenle birlikte konuyu tartışır
			ST103Q08NA	Öğretmen sorularımızı bizimle tartışır
			ST103Q11NA	Öğretmen fikrini söyler
	Fen bilgisi öğretmenlerinden alınan geri bildirim (Faktör 2)	5	ST104Q01NA	Öğretmen performansıma ilişkin bilgi verir
			ST104Q02NA	Öğretmen güçlü yanlarım hakkında bana dönüt verir
			ST104Q03NA	Öğretmen beni hangi alanlarda hala gelişebileceğim konusunda bilgilendirir
			ST104Q04NA	Öğretmen performansımı nasıl yükseltebileceğimi söyler
			ST104Q05NA	Öğretmen öğrenme amaçlarıma nasıl ulaşacağım konusunda bana tavsiyede bulunur
	Fen bilgisi derslerine ilişkin uyarlanabilir	3	ST107Q01NA	Öğretmen derse ilişkin hazırlığını sınıfın ihtiyaçlarına ve bilgi düzeyine göre yapar

öğretim (Faktör 3)	ST107Q02 NA	Öğretmen, bir öğrencinin bir konuyu veya görevi anlamada zorluk çekmesi durumunda kişisel yardım sağlar
	ST107Q03 NA	Öğretmen, çoğu öğrencinin anlamada zorluk çekeceğini düşündüğü bir dersin yapısında değişikliğe gider
Sorgulama temelli fen bilgisi öğretimi (Faktör 4)	ST098Q01 TA	Öğrencilere fikirlerini açıklama fırsatı verilir
	ST098Q02 TA	Öğrenciler, laboratuvarda deneyler yaparak zaman geçirirler
	ST098Q03 NA	Öğrencilerin fene ilişkin sorularla ilgili tartışması gerekir
	ST098Q05 TA	Öğrencilerden yaptıkları bir deneyden sonuç çıkarmaları istenir
	ST098Q06 TA	Öğretmen bilimsel bir düşüncenin birkaç farklı olgu için nasıl uygulanabileceğini açıklar
	ST098Q07 TA	Öğrencilerin kendi deneylerini tasarlamalarına izin verilir
	ST098Q08 NA	Araştırmalara ilişkin tartışma gerçekleştirilir
	ST098Q09 TA	Öğretmen, fene ilişkin kavramların yaşamla olan ilişkisini açık bir şekilde belirtir
	ST098Q10 NA	Öğrencilerden fikirlerini test etmek için araştırma yapmaları istenir

Not: Maddelere ilişkin derecelendirmeler, öğrencilerin fen bilgisi derslerinde ne sıklıkla ilgili durumlarla karşı karşıya kaldıklarına ilişkindir (1= Asla, 2= Bazı derslerde, 3= Çoğu derste, 4= Hemen hemen her derste)

Verilerin Analizi

Araştırma kapsamında kullanılan fen eğitimine ilişkin alt boyutlarda yer alan madde sayısı toplamının 21 olması ve bu 21 maddeye ilişkin puanlar girdi olarak kullanıldığı takdirde bu puanlara ilişkin yapılacak olan yorumların model incelemesinin önüne geçeceği ve yorumlama kolaylığı göz önünde bulundurularak çalışmada faktör puanları kullanılmıştır. Ayrıca, faktör puanlarının ilişkili değişkenlerin ağırlıklandırılmış kombinasyonları olmasının bu puan türlerini gerçek değerler karşısında daha güvenilir ve daha kaliteli yapması da faktör puanlarının araştırma kapsamında tercih edilmesinin sebeplerindendir (Fiedler ve Mcdonald, 1993). Faktör puanları, gizil değişkenlere ilişkin bilgi sahibi olmak ve örneklemi oluşturan bireylere ilişkin puanların gizil boyuttaki göreceli durumunu belirlemek gibi çok çeşitli amaçlar için kullanılabilir. Kavramsal olarak faktör puanı, gizil faktörün doğrudan ölçülebildiği varsayıldığında o kişinin gözlemlenebilecek puanıdır ve değişkenlere ilişkin yeni değerlerdir. Faktör puanları ağırlıklandırılmış, ortalaması 0 standart sapması 1 şeklinde ölçeklendirilmiş puanlardır. Teoride her ne kadar faktör puanlarının ortalamasının 0'a eşit olduğu belirtilse de, pratikte 0 noktasından kaymalar görülebilmektedir (Mulaik, 2009; Thompson, 2004). Faktör puanları, bilgisayar programları kullanılarak oluşturulabilmektedir. Araştırmada kullanılan faktör puanları SPSS 20 programı kullanılarak oluşturulmuştur. Faktör puanlarının kestirimine ilişkin değişik yöntemler mevcuttur. Araştırma kapsamında kullanılan faktör puanları, en küçük kareler yöntemini temel alan, çok kolay bir şekilde ve istatistiksel geçerliği diğer yöntemlerle kestirilen faktör puanları ile karşılaştırıldığında daha yüksek olan regresyon yöntemi ile kestirilmiştir (Grice, 2001; Mulaik, 2009).

Faktör puanlarına ilişkin yorum yapılırken, orijinal puan toplamı veya orijinal puanların ortalaması kullanılarak yapılan yorumlara benzer yorumlar yapılmaktadır. Bir başka deyişle, sıfır faktör puanı kişinin ilgili özniteliklerinin önem derecesinin örneklem için ortalamaya yakın olduğu, pozitif faktör puanı ortalamanın üzerinde olduğu, negatif faktör puanı ise kişinin ilgili özniteliklerinin önem derecesinin ortalamanın altında olduğu anlamına gelmektedir (DiStefano, Zhu ve Mindrilla, 2009; Grice, 2001; Wells, 1999).

Çalışma kapsamında girdi değişkeni olarak fen bilgisi öğretimi ile ilgili 21 maddeye ilişkin dört alt boyutu temsilen faktör puanları ortalaması ile birlikte10 farklı olası fen

başarı puanının ortalaması alınmış ve elde edilen değişken de diğer dört değişkenle birlikte analizlerde kullanılmıştır.

Araştırmada, Kohonen'in Öz Örgütlemeli Harita Yöntemi, K-Ortalamalar Yöntemi ve İki Aşamalı Kümeleme Analizi kullanılmıştır. Kohonen'in Öz Örgütlemeli Harita Yöntemi ile kümeleme analizi ve K-Ortalamalar Kümeleme Analizi R programı ile İki Aşamalı Kümeleme Analizi ise SPSS 20 ile gerçekleştirilmiştir. R programı kapsamında gerçekleştirilen analizlerde 18 farklı fonksiyon paketi kullanılmıştır.

Kümeleme analizi, bir gözlemler setinin yapısal karakteristiklerini ölçmeye çalışan objektif bir metottur. Kümeleme analizinde verilere ilişkin normal dağılım varsayımı olmakla birlikte normallik varsayımı prensipte kalmakta, çok fazla dikkate alınmamaktadır. Ayrıca kümeleme analizinde kovaryans matrisine ilişkin herhangi bir varsayım bulunmamakta, diğer çok değişkenli analizlerde aranan doğrusallık, eş varyansa sahip olma gibi varsayımlar da aranmamaktadır. Kısacası kümeleme analizi, varsayımların ihlallerine dirençli (robust) bir analiz türüdür (Garson, 2014; Hair, Black, Babin ve Anderson, 2009; Tatlıdil, 1992).

Kümeleme analizi varsayımların ihlali söz konusu olduğunda her ne kadar dirençli bir analiz türü olsa da, araştırma kapsamında kullanılan veri seti ile ilgili daha ayrıntılı bilgi sahibi olmak ve herhangi bir veri ön işleme hatasına sebebiyet vermemek adına, veri setine ilişkin normallik ve doğrusallık varsayımları incelenmiştir.

Veriyi ön işleme anlamında en önemli basamaklardan bir tanesi ölçeklendirmedir. Ölçeklendirme, normalleştirme ve standartlaştırma olmak üzere iki farklı başlık altında incelenmektedir. Normalleştirme, sınırlı aralıkta değerlere ihtiyaç olduğunda faydalıdır. Standartlaştırma ise, belirli mesafe ölçütlerine dayanan özellikler arasındaki benzerlikleri karşılaştırmak için önem arz edebileceğinden kümeleme analizinden önce normalleştirmeye göre daha sık kullanılmaktadır ve standartlaştırma ile ilgili veriler ortalaması 0 standart sapması 1 olan puanlara dönüştürülür.

Veri madenciliğinde gürültülü/kirli veri olarak tanımlanan uç değerlerin analiz sonuçlarını olumsuz yönde etkilememesi amacıyla çalışma kapsamında ele alınan değişkenlerin aynı ölçek düzeyine indirgenmesi amaçlanmıştır. Bu sebeple olası fen başarı puanları ortalamasına ilişkin yüksek değerler standardize edilmiş ''z'' değerlerine dönüştürülmüş ve analizlerde olası fen başarı puanları ortalaması yerine standardize edilmiş z değerleri kullanılmıştır.

Kohonen'in Öz Örgütlemeli Harita Yöntemi denetimsiz öğrenme gerçekleştiren bir sinir ağı yapısına sahip olduğundan ve kümeleme analizi ile veri setinin temel yapısının belirlenmeye çalışılması amaçlandığından, araştırmada kullanılan veri seti eğitim ve test veri seti olmak üzere ikiye ayrılmamış, hem eğitim hem de test sürecinde veri setinin tamamı kullanılmıştır.

Araştırmanın birinci alt problemi için Kohonen'in Öz Örgütlemeli Harita Yöntemi kapsamında eğitim sürecine ilişkin bilgi elde etme anlamında en yakın birime olan ortalama uzaklık-iterasyon sayısı grafiği; modele ilişkin oluşturulan iki boyutlu haritanın niteliğini belirlemede sayı grafiği; nöronların komşuluk mesafesine ilişkin bilgi elde etmede komşuluk mesafesi grafiği; kurulan modeldeki birim ve değişkenlerin dağılımındaki kalıplar hakkında bilgi sahibi olmada kod vektörleri haritası; değişkenlerin kümelemedeki önemine ilişkin bilgi elde etmede ısı grafikleri; ideal küme sayısını belirlemede ise küme içi kareler toplamının değişimi, farklı küme sayıları için siluet grafikleri ve kalibrasyon grafiğinden yararlanılmıştır.

Araştırmanın ikinci alt problemi için K-Ortalamalar Yöntemi kapsamında ideal küme sayısını belirlemede grup içi kareler toplamı, kümeler arası hata değerleri, ABK ve BBK değerleri, Gap istatistiği ve kümelere ilişkin dağılım grafiğinden yararlanılmıştır.

Araştırmanın üçüncü alt problemi için İki Aşamalı Kümeleme Analizi kapsamında ideal küme sayısını belirlemede farklı küme sayıları için siluet değerlerinden; Schwarz'ın Bayesçi Ölçütü'nden, BBK değişiminden, BBK oranından ve uzaklık ölçüsü oranından; girdi değişkenlerinin kümelerin oluşmasındaki önem düzeyi için ise değişkenlere ilişkin önem düzeyi grafiğinden yararlanılmıştır.

Bölüm 4

Bulgular ve Yorumlar

Bu bölümde, belirlenen alt problemlerin sırasına göre bulgular ve ilişkili tablolara yer verilmiştir. Belirlenen alt problemler sırayla başlıklar halinde belirtilmiş ve bulgular yorumlanmıştır. Araştırma kapsamında kullanılan üç farklı kümeleme analizine ilişkin bulgulara geçilmeden önce kümeleme analizinde girdi olarak kullanılan faktör puanları ile ilgili ön bilgi sahibi olma anlamında betimsel istatistikler, normallik ve korelasyon değerleri incelenmiştir. Çalışma kapsamında öğrencilerin PISA sınavından elde ettikleri sonuçlara göre kümeleme analizinde kullanılan değişkenlerin isimleri ile merkezi eğilim ve değişkenlik ölçüleri Tablo 3'te gösterilmiştir.

Tablo 3

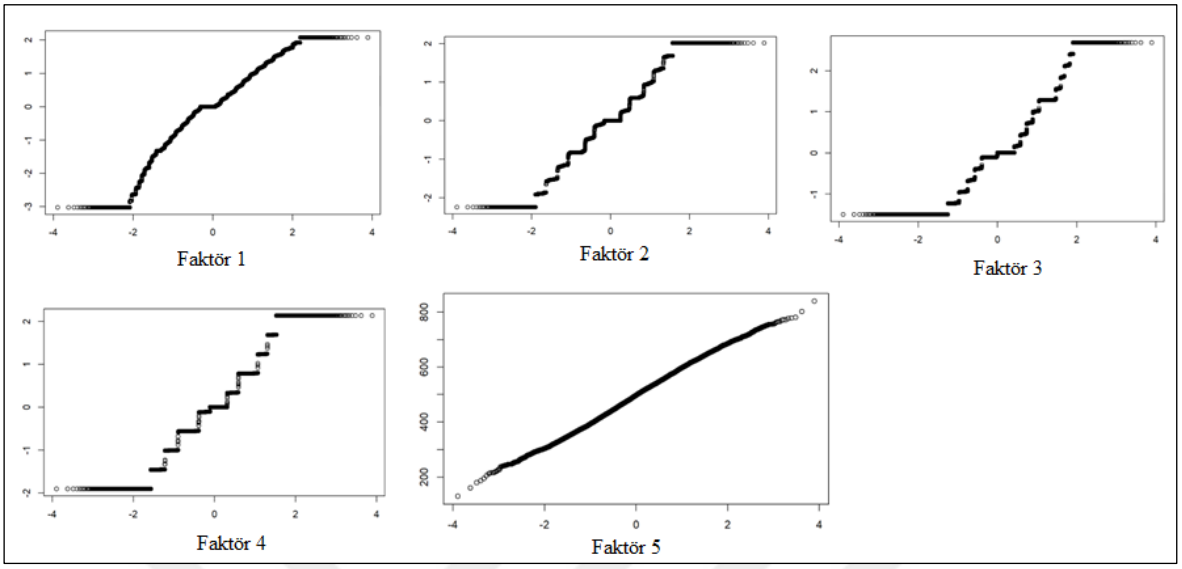
Değişkenlere İlişkin Betimsel İstatistikler

	Faktör 1	Faktör 2	Faktör 3	Faktör 4	PVSCIENCE
Min.	-3.025730	-2.240240	-1.497220	-1.90580	130
1. Çeyrek	-0.494815	0.790430	-0.661040	-0.55828	426.8
Medyan	-0.003210	0.001050	-0.023180	-0.00086	497.5
Ortalama	0.003238	0.009943	-0.007968	0.01963	495.8
Maksimum	2.071930	2.008300	2.681290	2.13676	839.7
Standart Sapma	0.9945847	0.9960432	0.9881865	0.9990277	97.79843

Not: Faktör 1 (Öğretmen yönetimindeki fen bilgisi öğretimi), Faktör 2 (Fen Bilgisi öğretmenlerinden alınan geri bildirim), Faktör 3 (Fen Bilgisi derslerine ilişkin uyarlanabilir öğretim), Faktör 4 (Sorgulama temelli fen bilgisi öğretimi), Faktör 5/PVSCIENCE (Fene ilişkin olası başarı puanı ortalaması)

Değişkenlerin betimsel istatistiklerine ilişkin Tablo 3 incelendiğinde, kayıp veriye rastlanmadığı, PVSCIENCE değişkeninin 130 ile 839.7 arasında değerler aldığı görülmektedir. Yine aynı tablo incelendiğinde, faktör1, faktör2, faktör3 ve faktör4'ün ortalamalarının 0 standart sapmalarının ise 1'e çok yakın olduğu; PVSCIENCE değişkeninin standart sapma değerinin ise 97.79843 olduğu gözlemlenmektedir. Daha önce de bahsedildiği gibi, analizlerde olası fen başarı puanları ortalaması yerine standardize edilmiş z değerleri kullanılmıştır. Veri setinde yer alan

değişkenlerin dağılımına ilişkin bilgi sahibi olma anlamında oluşturulan Q-Q grafikleri Şekil'de gösterilmiştir. Q-Q grafikleri Şekil 6'da gösterilmiştir.



Şekil 6. Faktörler için elde edilen q-q grafikleri

Araştırma kapsamındaki olası başarı puanları ortalamasına ilişkin Q-Q Plot grafiği incelendiğinde, ilgili değişkenin normal dağılıma çok yakın olduğu söylenebilir. Normallik varsayımı için kullanılan istatistiksel yöntemlerden biri olan Kolmogorov Smirnov normallik testi sonuçları Tablo 4'te gösterilmiştir.

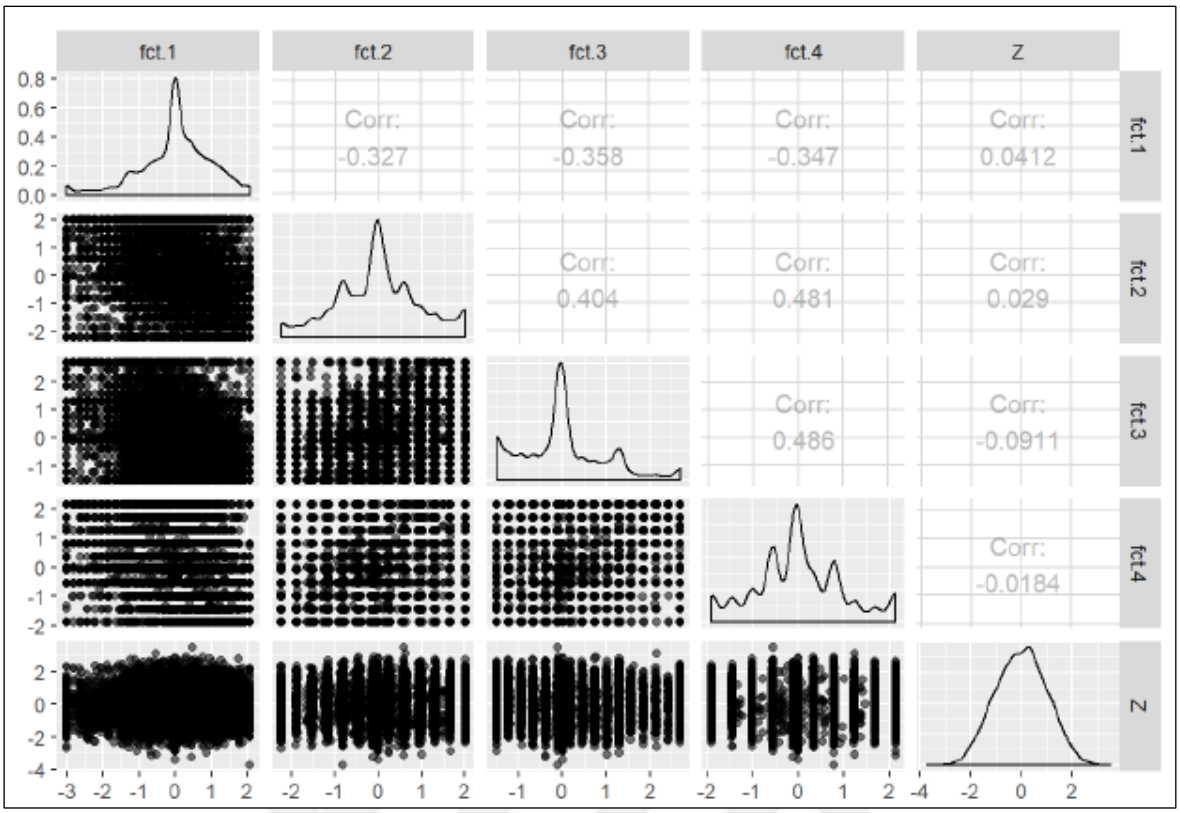
Tablo 4

Kolmogorov Smirnov Normallik Testi Sonucu

	Kolmogorov Smirnov Değeri	p
Faktör 1	0.11	.080
Faktör 2	0.09	.076
Faktör 3	0.16	.064
Faktör 4	0.12	.058
PVSCIENCE	0.01	.052

Kolmogorov Smirnov test sonucu ve Q-Q Plot grafiği ile gözlemlenen normal dağılım sonucunu destekler niteliktedir (p>0.05).

Varsayımlar kapsamında ikinci olarak, doğrusallık varsayımı incelenmiştir. Doğrusallık varsayımı, analize dahil edilen değişkenlere ilişkin korelasyonların çok yüksek (0.8 ve üzeri) düzeyde olmaması ile ilişkilidir. Kısacası bu varsayımla, değişkenler arası ilişkilerin çok yüksek olması istenilen bir durum değildir. Şekil 7'de ilişkiler ve birlike dağılım matrisine ilişkin grafik gösterilmiştir.



Şekil 7. İlişkiler ve birlikte dağılım matrisi

Şekil 7 incelendiğinde, değişkenler arasında güçlü ilişkiler olmadığı, en yüksek ikili korelasyon değerinin 0.486 (faktör 3 ve faktör 4 arasında) olduğu göze çarpmaktadır. Daha önce de belirtildiği gibi, kümeleme analizinde varsayımların karşılanması gibi bir zorunluluğun olmaması ve standardizasyon işlemi ile varsayımların kısmen karşılanması göz önünde bulundurulduğunda, araştırma kapsamındaki analizlere devam edilmiştir.

Aynı zamanda araştırma kapsamında kayıp veri oranının çok yüksek olduğu Slovenya'ya ilişkin katılımcılar çalışma grubuna dahil edilmemiş ve analizler 34 ülkeye ilişkin katılımcı grubu ile gerçekleştirilmiştir.

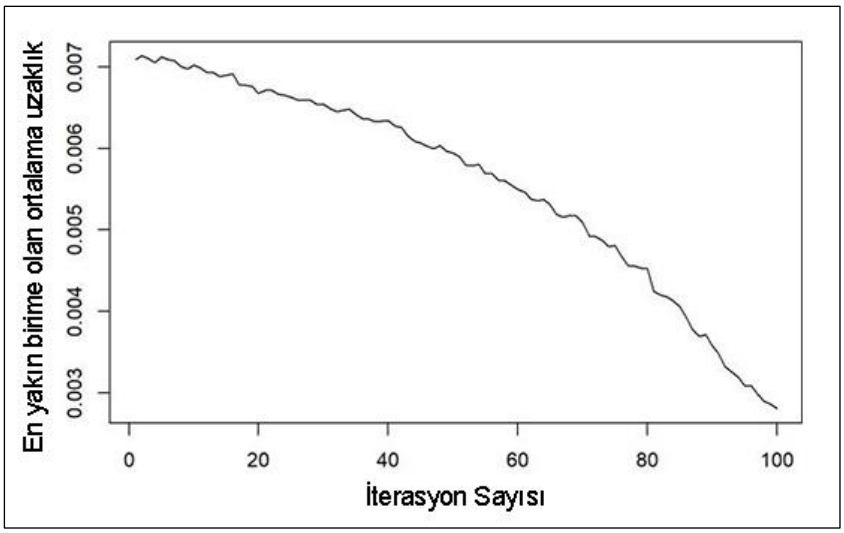
Slovenya'ya ilişkin kayıp veri oranının çok yüksek olması ve bu sebeple kayıp veri ataması sonucu atanan verilerin çok sık bir şekilde tekrar etmesi sonucunda değişkenler bazında korelasyonların hesaplanmadığı görülmektedir. Bu sebeple, Slovenya'ya ilişkin katılımcılar araştırma kapsamının dışında tutulmuş ve analizler 34 ülkeye ilişkin katılımcı grubu ile gerçekleştirilmiştir.

Birinci Alt Probleme İlişkin Bulgular

Çalışmanın birinci alt probleminde Kohonen'in Öz Örgütlemeli Harita Yöntemi ile elde edilen küme sayıları ve kümelere ilişkin istatistikler rapor edilmiştir (Kohonen, 2001). Kohonen'in Öz Örgütlemeli Harita yöntemi ile Kümeleme Analizinin gerçekleştirilmesi için çalışmada R programı kullanılmıştır. Kohonen'in Öz Örgütlemeli Harita Yönteminde çok boyutlu iki nesne benzer ise iki boyutlu düzlemdeki pozisyonları birbirine yakın olmalıdır. Bu yöntemde nesneleri "sürekli bir uzayda" eşlemek yerine, nesnelerin haritalandırıldığı nöronlardan meydana gelen iki boyutlu bir grafik kullanır. Çok boyutlu ölçekleme en büyük farklılıklara odaklanırken, Kohonen'in Öz Örgütlemeli Harita Yöntemi büyük benzerliklere odaklanmaktadır. Bir başka deyişle, çok boyutlu ölçekleme ile gerçekleştirilen iki boyutlu bir çizimde büyük bir mesafe gerçek bir mesafenin tahmini olarak doğrudan yorumlanabilirken; Kohonen'in Öz Örgütlemeli Harita Yöntemi ile sadece aynı veya komşu birimlere eşlenen nesnelerin çok benzer olduğu söylenebilir. Özetlemek gerekirse genel anlamda Kohonen'in yöntemi, temel bileşen analizi (Hotelling, 1933), çok boyutlu ölçekleme (Kruskal ve Wish, 1978) gibi klasik yöntemleri sinir ağı temelli bir şekilde devam ettirir ve kümelenme sonuçlarını kolayca anlaşılabilir bir geometrik ekran olarak görselleştirir.

Mevcut analizde nöron sayısı, başlangıçtaki grid olarak kullanılacak veri setinin 9,870 gözlemine uyum sağlama anlamında program tarafından makul sayı olarak belirlenen 30 × 30 = 900 nörondan oluşacak şekilde seçilmiştir. Grid için belirlenen şekil altıgendir. Yineleme sayısı ve öğrenme oranı varsayılan değerler olan 100 ve 0.05 seviyelerinde tutulmuştur (öğrenme oranı her iterasyonda doğrusal olarak 0.01 oranında azalmaktadır. Normalde öğrenme oranı, analizin başında belirlenir ve iterasyon sayısı değiştikçe değişim göstermez. Fakat Kohonen'in Öz Örgütlemeli Harita Yöntemi ile kümeleme söz konusu olduğunda, yakınsamanın sağlanması için öğrenme oranını azaltmak gerekmektedir. Bir başka deyişle, öğrenme oranı değişmezse, öğrenme sürecinin sonu gelmeyebilir.

Kohonen'in Öz Örgütlemeli Harita Yöntemi ile kümeleme analizinin ilk aşaması eğitim sürecidir. Eğitim sürecine ilişkin iterasyonlar arttıkça, her nöronun ağırlıklarından dolayı o nöron tarafından temsil edilen örneklere ilişkin mesafe azalır. Çalışma kapsamında iterasyon sayısına bağlı olarak elde edilen kümeler için küme içi uzaklıkların nasıl bir değişim gösterdiği Şekil 8'de gösterilmiştir.

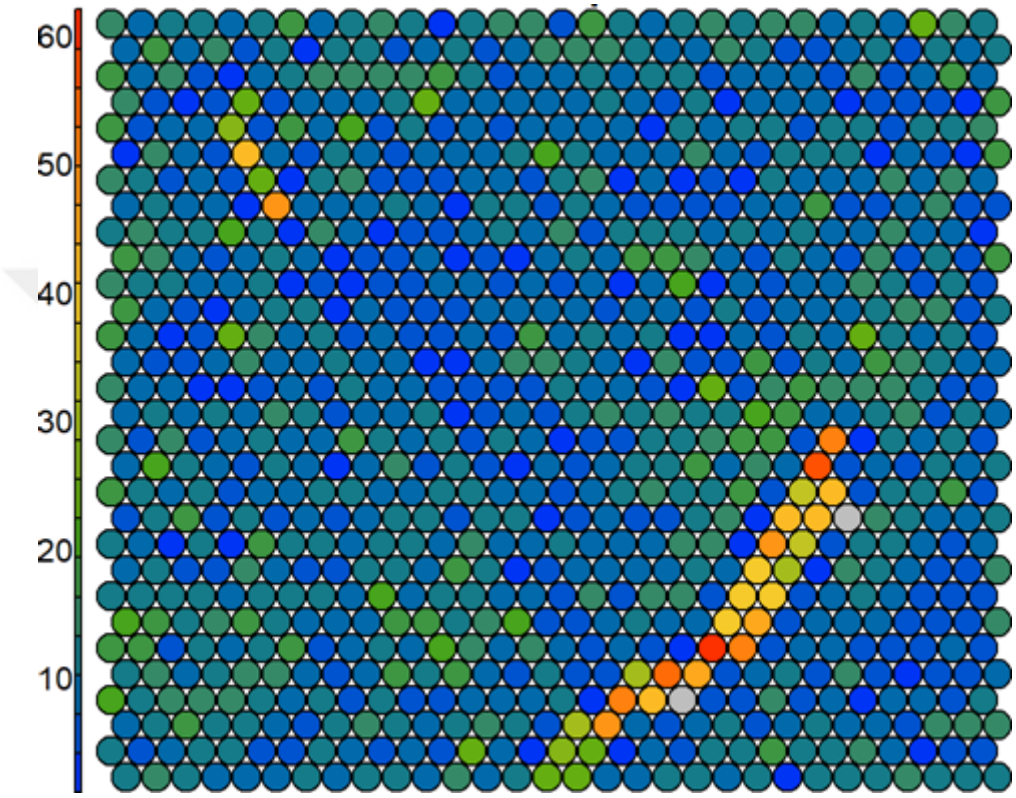


Şekil 8. Veri setinin eğitim süreci

Şekil 8 incelendiğinde Kohonen'in Öz Örgütlemeli Harita Yönteminin eğitim sürecine ilişkin iterasyon sayısı arttıkça, her bir nöronun ağırlığından o nöron tarafından temsil edilen örneklere olan uzaklığın azaldığı göze çarpmaktadır. Aynı zamanda eğitim sürecine ilişkin grafik incelendiğinde, gözlemler ile onlara en yakın birim arasındaki ortalama mesafenin stabilize olmamasına rağmen son iterasyonlarda daha hızlı bir düşme eğiliminin olduğu görülmekte, dolayısıyla eğitim sürecinin etkili olduğu göze çarpmaktadır. Bu aşamada istenilen, yamaç birikinti grafiğinde olduğu gibi çizgi grafiğinin düz bir platoya ulaşmasıdır. Aynı zamanda, iterasyon sayısının arttırılmasının hızlı bir düşme eğilimine girmesi sebebiyle daha iyi bir uyuma işaret etmeyeceği düşünülmüş, iterasyon sayısının arttırılması ile oluşabilecek olan aşırı uygunluk (overfitting) tehlikesine karşın iterasyon sayısını varsayılan miktar olan 100'de tutma kararı alınmıştır. Elde edilen bu sonuca göre iterasyon sayısının kümeleme analizi için yeterli olduğu belirlenmiştir.

Kohonen'in Öz Örgütlemeli Harita Yöntemi ile, her bir nörona kaç tane örnek eşlendiğinin sayısını görselleştirmeye yardımcı olan Sayı Grafiği (Counts Plot) elde edilmektedir. Bu görsel, kümelemeye ilişkin asıl haritanın kalitesinin bir ölçüsüdür. Şekil 9'da yer alan grafikte her bir nöronda yer alan birim sayıları renkler aracılığıyla gösterilmektedir. Grafiğin sol tarafında dikey olarak gösterilen renk paleti incelendiğinde maviden kırmızıya doğru metrik bir derecelendirme yapılmaktadır. Mavi renkler nöronlarda yer alan birim sayısının düşük olduğunu; kırmızı renkler ise nöronlarda yer alan birim sayısının fazla olduğunu göstermektedir. Çok sayıda mavi renkli nöronun olması haritanın boyutunun yani nöron sayısının veri seti için fazla

olduğuna ve dolayısıyla azaltılması gerektiğini gösterirken; çok sayıda kırmızı renkli nöronun olması ise harita için kullanılması gereken nöron sayısının arttırılması anlamına gelmektedir. Ayrıca, genel anlamda renk paletinde yer alan nöron başına 5-10 nöron hedeflenmesinin de homojenliğe katkı sağlayacağı düşünülmektedir (Kohonen, 2001; Wehrens ve Buydens, 2007). Çalışma kapsamında her bir nöronda yer alan birim sayılarına ilişkin Sayı Grafiği Şekil 9'da gösterilmiştir.

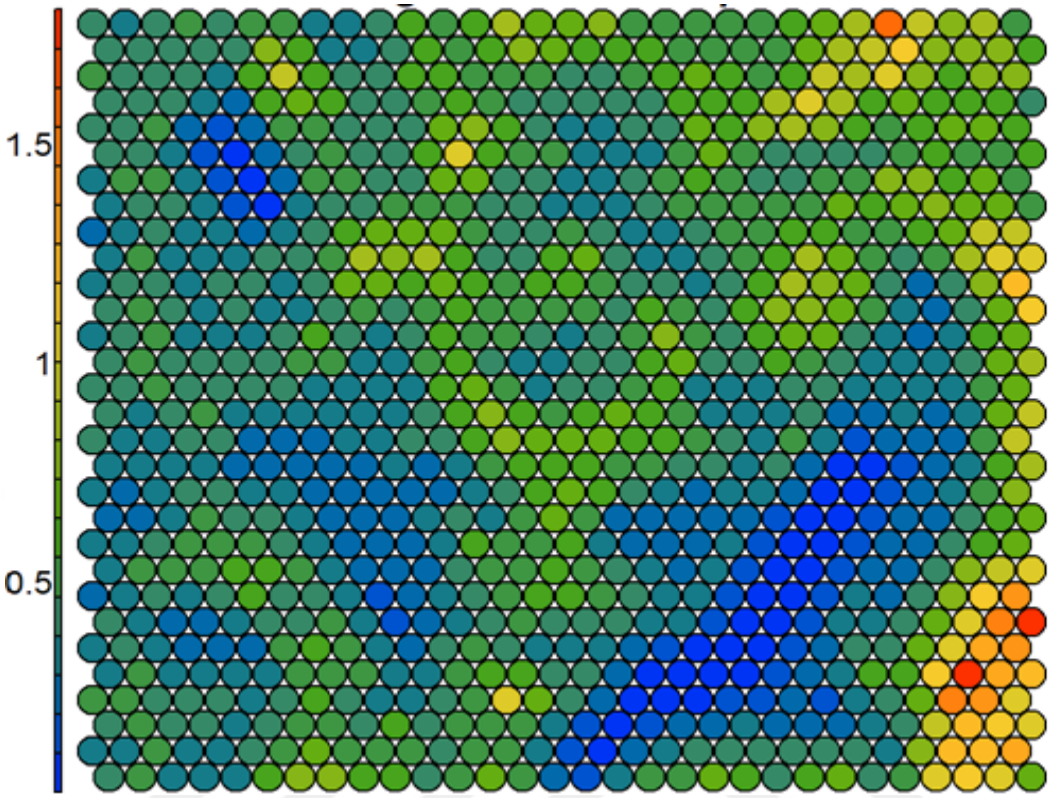


Şekil 9. Nöronlarda yer alan birimlere ilişkin sayı grafiği

Analize ilişkin Şekil 9'da gösterilen grafik incelendiğinde, nöron başına düşen gözlem sayısının nispeten homojen olduğu söylenebilir. Ayrıca genel anlamda, renk paletinde yer alan her bir renk için haritada en az 5-10 nöron yer aldığı görülmektedir. Sonuç olarak, Şekil 9'da gösterilen 900 nörondan meydana gelen harita ve veri setinde yer alan birim sayısı göz önünde bulundurulduğunda nöron sayısının yeterli olduğu kanısına varılmıştır.

Analiz sonucunda elde edilen grafiklerden bir diğeri ise, nöronların komşuluk mesafesine ilişkin bilgi veren ''Komşuluk mesafesi grafiğidir''. Bu grafik, "U-Matrisi" olarak da bilinir. Bu grafik, haritada yer alan her bir nöronun birbirine olan mesafesi hakkında bilgi vermekte, benzer nöronları bir araya getirirken sınırların kullanılmasını

önermektedir. Kohenen'in Öz Örgütlemeli Harita Yöntemiyle elde edilen kümeler ve kümeler arası mesafelere ilişkin Komşuluk Mesafesi Grafiği Şekil 10'da gösterilmiştir.

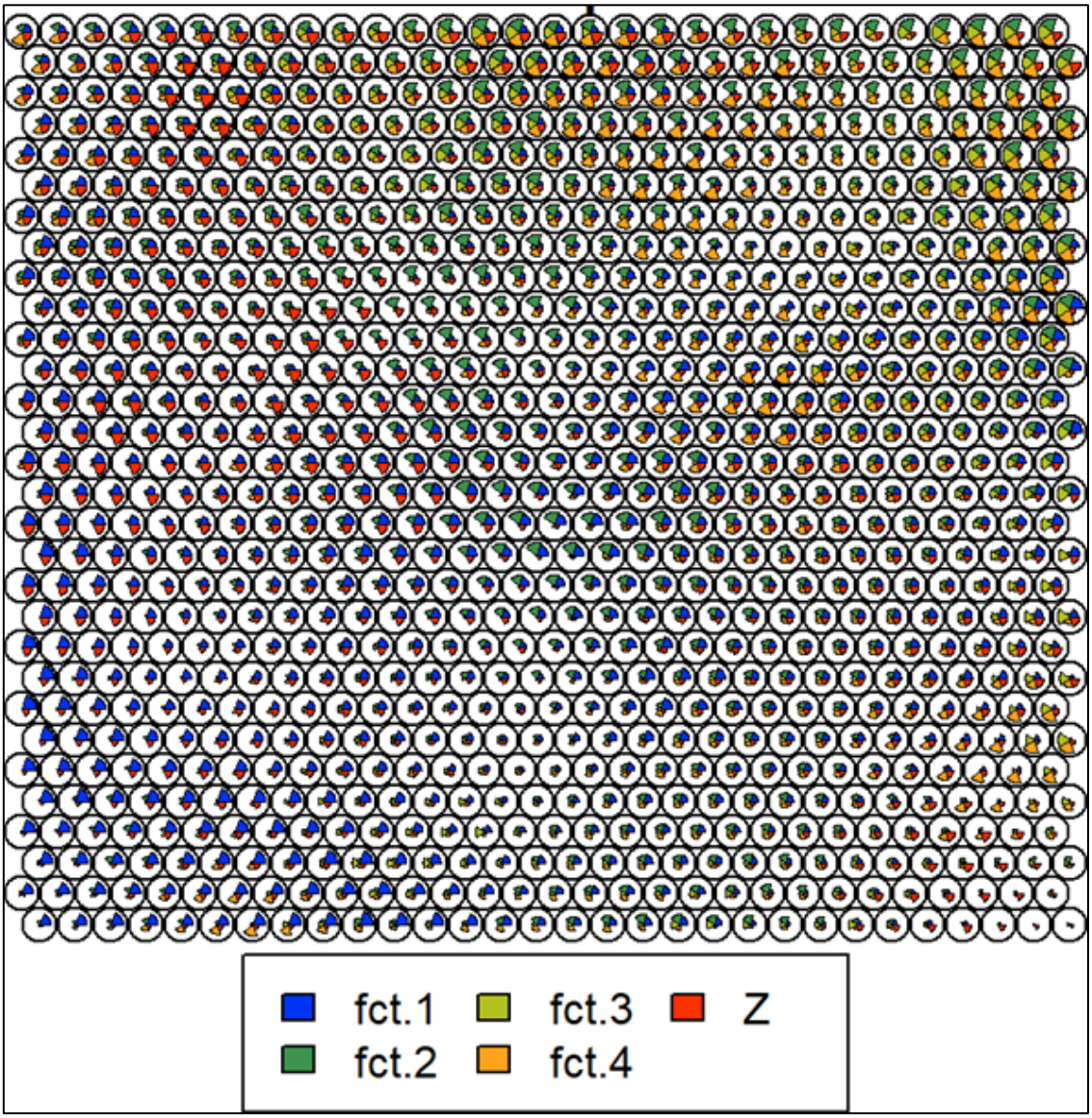


Şekil 10. Komşuluk mesafesi grafiği

Şekil 10 incelendiğinde, renk paleti maviden kırmızıya doğru ilerledikçe nöronların komşuluk mesafelerinin arttığı göze çarpmaktadır. Söz konusu olan mesafe Öklid mesafesidir. Komşuluk mesafelerinin renk paleti ile gösterildiği grafikte mavi renkler nöronlar arası mesafenin düşük olduğunu dolayısıyla nöron gruplarının benzer olduğuna işaret etmektedir. Grafikte yer alan kırmızı renkler ise nöronlar arası mesafenin fazla olduğunu dolayısıyla nöron gruplarının farklı olduğuna işaret etmektedir. Bu durum göz önünde bulundurulduğunda Şekil 10, Kohonen'in Öz Örgütlemeli Harita Yöntemi kapsamında oluşturulan kümeleri tanımlamak için kullanılabilir. Şekil 10'da gösterilen grafik incelendiğinde, komşularına çok yakın olan nöronların koyu mavi renkler olduğu, komşularına orta mesafede yakın olan nöronların yeşil renkler olduğu, komşularına uzak mesafede olan nöronların ise kırmızı renkler olduğu belirlenmiştir. Bu durum, görsel anlamda kümelerin birbirinden iyi bir şekilde ayrıştırılabileceğine ilişkin bir ipucudur.

Nöronlara ilişkin ağırlık vektörleri bir diğer adıyla "Kodlar", öz örgütlemeli bir harita oluşturmak için kullanılan değişkenlerin normalleştirilmiş değerlerinden oluşur. Her bir

nöronun ağırlık vektörü, o nöronla eşlenen örneklerin temsilcisidir. Haritadaki ağırlık vektörleri görselleştirerek, birimler ve değişkenlerin dağılımındaki kalıplar hakkında bilgi sahibi olunmaktadır. Kısacası, kod vektörlerinin dağılımına ilişkin bir harita, ilgili haritanın farklı alanlarının tanımlanmasında analize alınan değişkenlerin rolü hakkında bilgi vermektedir. Analiz sonucunda elde edilen Ağırlık Vektörlerine ilişkin harita Şekil 11'de gösterilmiştir.



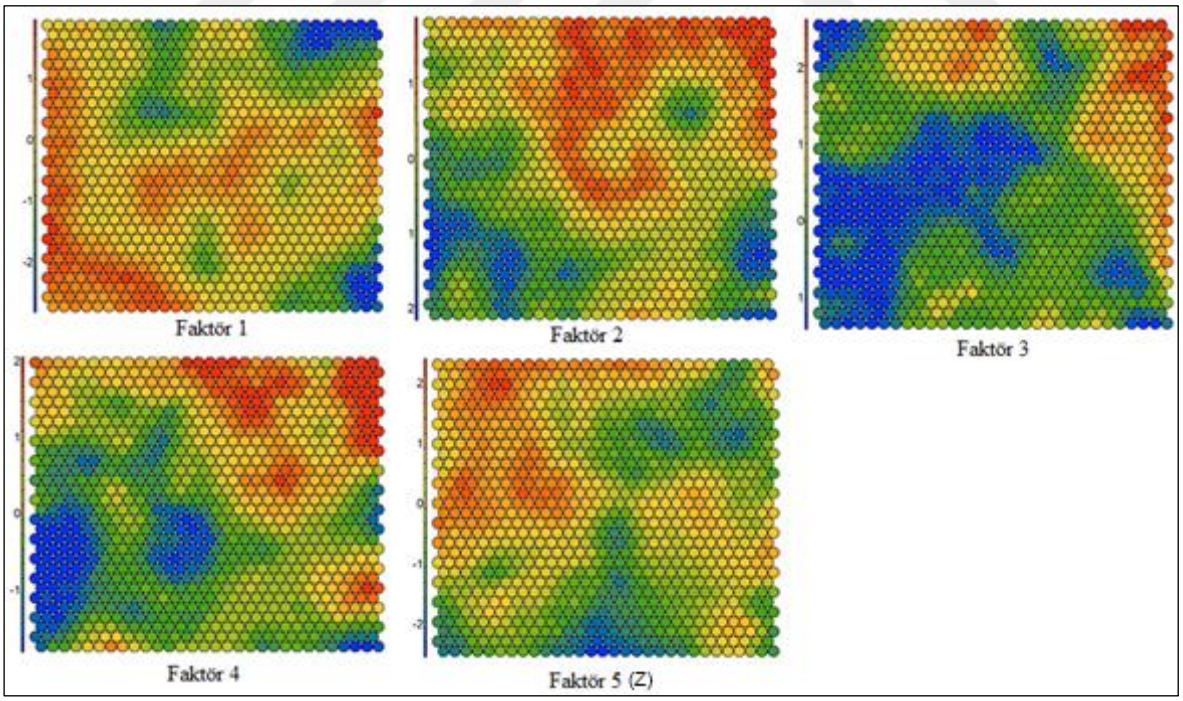
Şekil 11. Kod vektörlerinin dağılımına ilişkin harita

Şekil 11'de her bir nöronun analize alınan değişkenler tarafından nasıl bir ağırlığa sahip olduğu gösterilmektedir. Haritada yer alan değişkenlere ilişkin tanımlanan renklerin yoğunluğu ilgili değişkenin göreceli etkisi hakkında bilgi vermektedir. Kod vektörlerinin dağılımına ilişkin harita incelendiğinde, Faktör 1 (fct.1) değişkenine ilişkin ağırlığın, haritayı meydana getiren nöronların büyük bir çoğunluğunda yer

aldığı ve diğer değişkenlere göre ağırlığının daha fazla olduğu göze çarpmaktadır. Fen okuryazarlığına ilişkin olası başarı puan ortalaması olarak tanımlanan z değişkeninin ise ikincil düzeyde haritayı meydana getiren nöronların büyük bir çoğunluğunda yer aldığı ve diğer değişkenlere göre ağırlığının daha fazla olduğu göze çarpmaktadır.

Nöron sayısı ve değişken sayısı arttıkça kod vektörlerinin dağılımına ilişkin haritayı okumak zorlaşmaktadır. Analiz kapsamında oluşturulan kod vektörlerinin dağılımına ilişkin haritanın 900 nörondan meydana gelmesi, bu haritanın yorumlanmasını zorlaştırmaktadır. Bu noktada, analize alınan bütün değişkenlerin ağırlıklarını tek bir haritada belirlemeye çalışmak yerine, her bir değişken için yüksek ve düşük değerli alanlar arasındaki kontrastı vurgulamaya çalışan bir grafik oluşturulabilir. Oluşturulacak tek değişkenli bu grafiklerin yorumlanması, kod vektörlerinin yorumlanmasına nazaran daha kolaydır.

Şekil 12'de analiz kapsamında kullanılan beş değişkenin analiz sonucunda meydana gelen kümelerin oluşumunda nasıl bir öneme sahip olduğuna ilişkin Isı Grafikleri yer almaktadır.

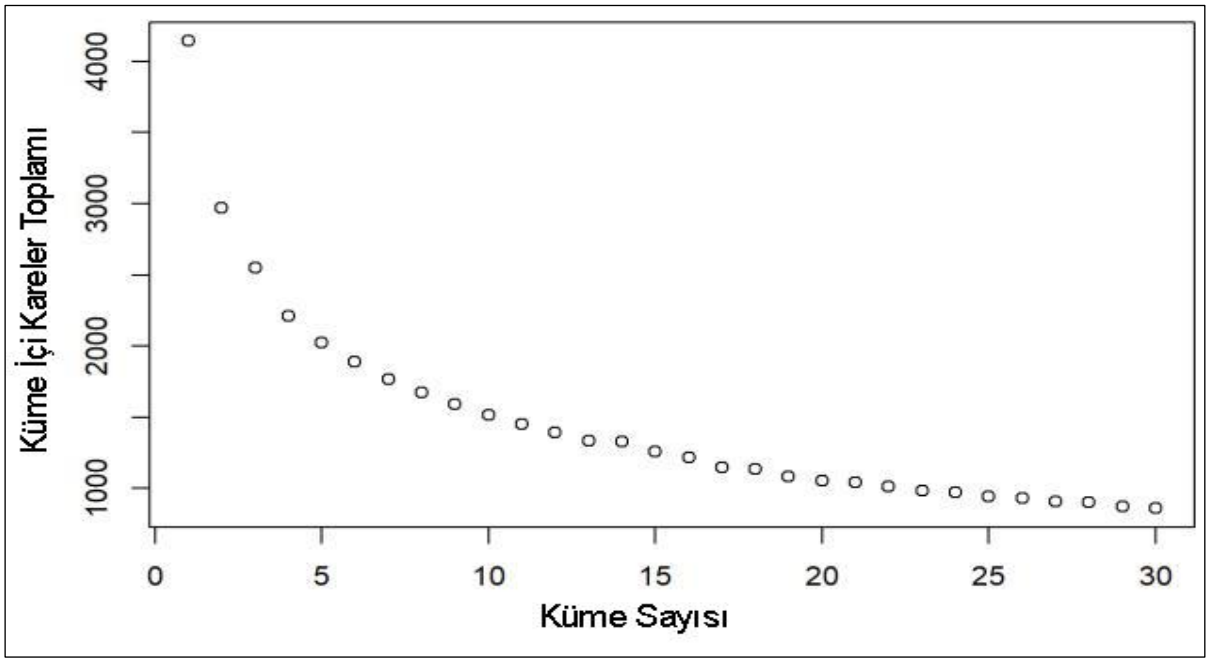


Şekil 12. Faktörlere ilişkin ısı grafikleri

Şekil 12'de gösterilen ısı grafikleri incelendiğinde dikey eksende yer alan birim sayılarının normalleştirilmiş değerler olduğu göze çarpmaktadır. Daha önce renk

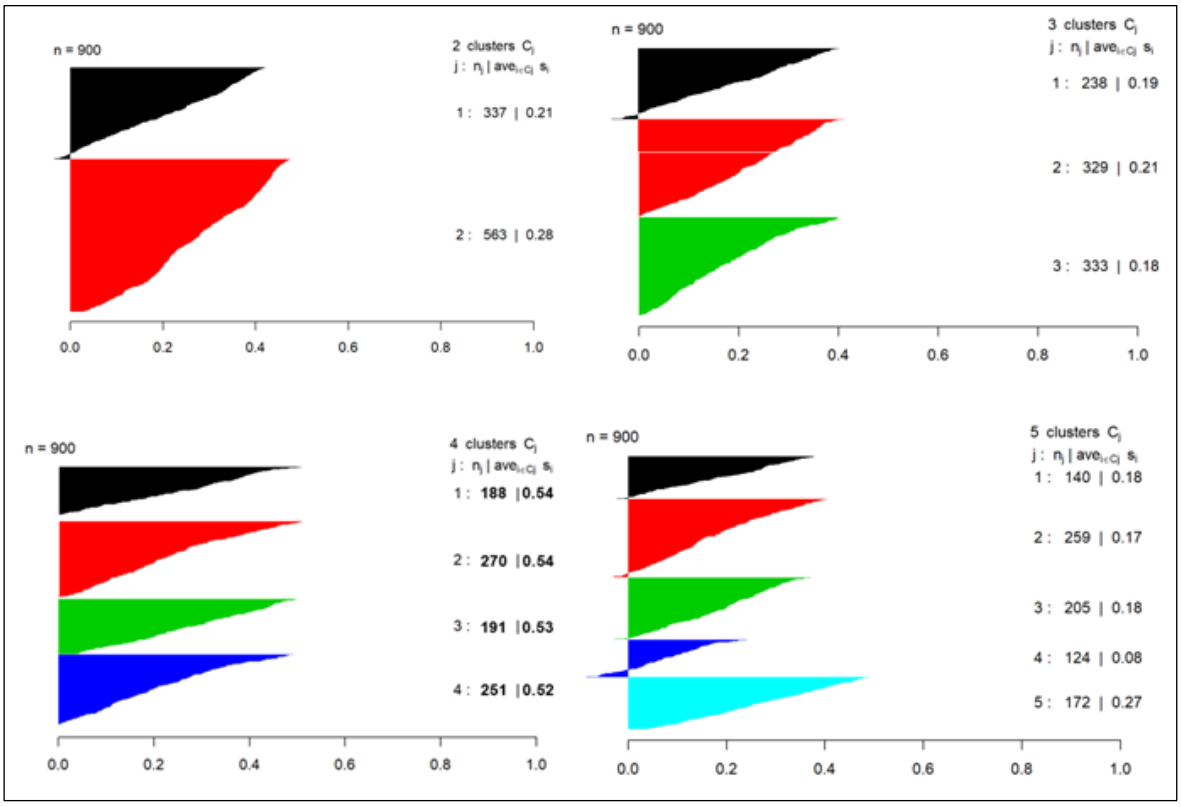
paletinde maviden kırmızıya doğru geçişin sayısal olarak 0'dan 60'a kadar olduğu göz önüne alındığında bu defa normalleştirilmiş değerlerin -2 ile +2 arasında değişkenlik gösterdiği belirlenmiştir. Isı grafikleri incelendiğinde, faktör1'in kümelenme anlamında en fazla etkiye sahip olduğu alanda faktör 3 en az etkiye; faktör 3'ün en az etkiye sahip olduğu alanda faktör 4'ün faktör 3 kadar olmasa da göreceli olarak en az etkiye sahip olduğu belirlenmiştir. Bunun yanında faktör 2'nin kümelenme anlamında en çok etkiye sahip olduğu alanda faktör 4'ün faktör 2 kadar olmasa da kümelenme anlamında etkili olduğu göze çarpmaktadır. Elde edilen bu sonuçlara göre Faktör 1 ile Faktör 3 arasında negatif yönde bir ilişki bulunurken Faktör 4 ile Faktör 2 arasında pozitif yönde bir ilişki olduğu söylenebilir. Bunun durumun sebebi Faktör 1 ile Faktör 3 için elde edilen grafikler farklı renklerde gösterilirken; faktör 4 ve faktör 2 için elde edilen grafiklerin aynı renkte gösterilmesidir.

Her bir değişkenin kümeleme analizinde nasıl bir etkiye sahip olduğuna ilişkin ısı grafiklerinden elde edilen bilgilerin ardından çalışmada elde edilen verilerin kaç küme altında bir araya geldiklerinin belirlenmesi aşamasına geçilmiştir. Bunun için öncelikle veri setinde yer alan birimler için elde edilecek küme sayılarına göre küme içi kareler toplamının nasıl bir değişkenlik gösterdiğine ilişkin Şekil 13'te gösterilen grafiğin incelenmesi gerekmektedir.



Şekil 13. Küme sayılarına göre küme içi kareler toplamının değişimi

Şekil 13'te gösterilen grafikte ideal küme sayısının belirlenebilmesi için grafiğin belli bir noktadan sonra düz bir plato şeklinde olması gerekmektedir. Bu grafik üzerinden elde edilen ideal küme sayısının belirlenmesi görsel olarak zor olduğundan sırasıyla küme sayısının k=2, k=3, k=4 ve k=5 olması durumunda elde edilen Siluet grafiklerinin incelenmesi gerekmektedir. Farklı küme sayıları için elde edilen Siluet grafikleri Şekil 14'te gösterilmiştir.



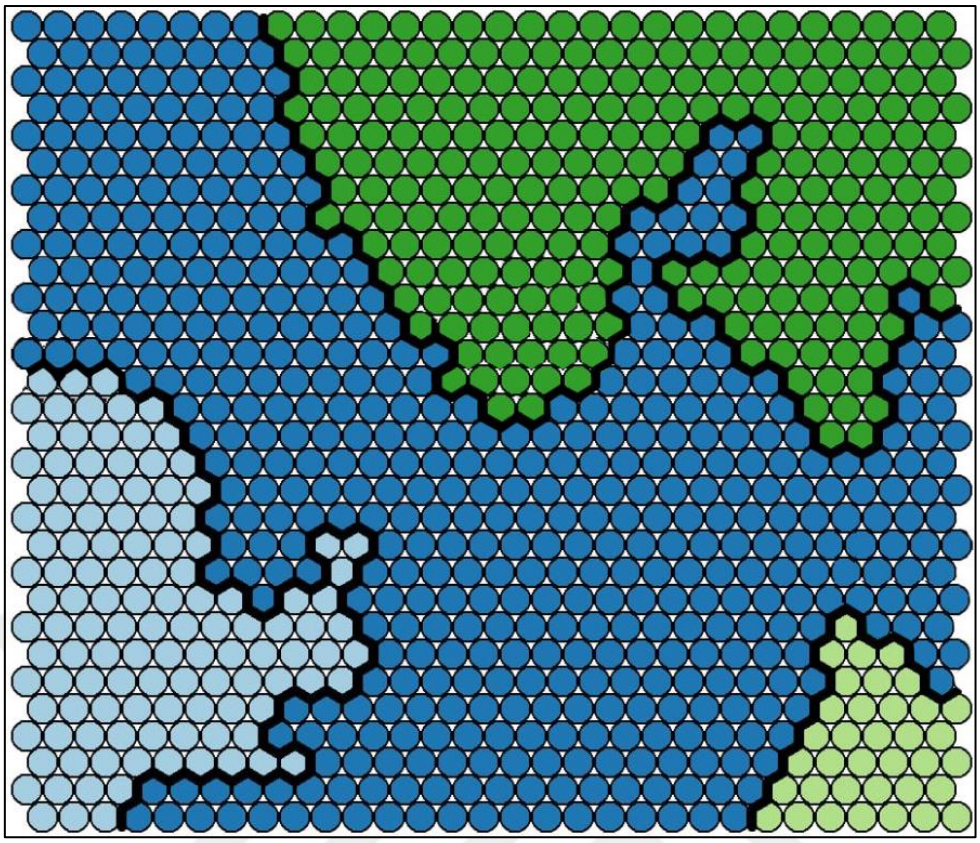
Şekil 14. Kümelerin geçerliğine ilişkin siluet grafikleri

Kümeleme analizi söz konusu olduğunda, elde edilen kümeler arasındaki ayırma mesafesini incelemek için, yani elde edilen küme sayısına karar vermek için siluet analizi kullanılabilir. Siluet analizine ilişkin grafik, bir kümedeki her noktanın komşu kümelerdeki noktalara ne kadar yakın olduğu ve kümelerin sayısı gibi parametreleri görsel olarak değerlendirme ile ilgilidir. Siluet değeri -1 ile +1 arasında bir değer alır.

Siluet katsayısı hesaplanırken hem kümeler arası hem de küme içi uzaklıklar dikkate alınmaktadır. R programı ile siluet analizi gerçekleştirildiğinde elde edilen grafik, hem kümeler için ayrı ayrı hesaplanmış siluet değerlerini hem de ortalama siluet değerini içermektedir.

Siluet katsayıları, dört farklı seviye altında değerlendirilir. Eğer elde edilen siluet katsayısı 0.25'e eşit veya bu değerden küçükse önem arz eden bir kümeye rastlanmadığı, 0.26-0.50 arasında ise elde edilen yapının zayıf olduğu ve farklı algoritmaların denenmesi gerektiği, 0.51-0.70 arasında ise makul bir yapının, 0.71-1.00 arasında ise güçlü bir yapının söz konusu olduğuna ilişkin yorum yapılabilir (Kaufman ve Rousseeuw, 1990).

Araştırma kapsamında elde edilen siluet grafikleri incelendiğinde, küme sayısı iki olarak belirlendiğinde birinci kümede 337, ikinci kümede 563 nöronun yer aldığı ve küme sayısına ilişkin ortalama silüet değerinin 0.25 olduğu; küme sayısı üç olarak belirlendiğinde birinci kümede 238, ikinci kümede 329, üçüncü kümede 333 nöronun yer aldığı ve küme sayısına ilişkin ortalama siluet değerinin 0.19 olduğu; küme sayısı dört olarak belirlendiğinde birinci kümede 188, ikinci kümede 270, üçüncü kümede 191, dördüncü kümede 251 nöronun yer aldığı ve küme sayısına ilişkin ortalama siluet değerinin 0.53 olduğu; küme sayısı beş olarak belirlendiğinde birinci kümenin 140, ikinci kümenin 259, üçüncü kümenin 205, dördüncü kümenin 124, beşinci kümenin 172 nörondan meydana geldiği ve küme sayısına ilişkin ortalama siluet katsayısının 0.18 olduğu tespit edilmiştir. İdeal küme sayısı dört olarak belirlendiğinde her bir küme için ayrı ayrı elde edilen sihouette değerlerinin 0.51'in üzerinde olduğu, yani dört kümenin de kabul edilebilir bir yapıya sahip olduğu ve analiz sonucu elde edilen ortalama siluet değerinin 0.53 olduğu belirlenmiştir. Şekil 14'te ideal küme sayısı iki, üç ve beş olarak belirlendiğinde yanlış kümelere atanan nöronların olabileceği (negatif yöndeki renk uzantıları) görülmektedir. Sonuç olarak elde edilen ortalama siluet katsayısı ile kümeler içi siluet katsayılarının 0.50'nin üzerinde olması ve yanlış kümelere atanan nöronların olmaması göz önünde bulundurulduğunda, belirlenen ideal küme sayısının 4 (dört) ve seçilen kümeleme yönteminin uygun olduğu söylenebilir. Bunun yanında birbirine benzeyen ve yan yana olan nöronların bir araya gelmesiyle oluşan ve ideal küme sayısına ilişkin bir diğer ölçüt olan kalibrasyon grafiği Şekil 15'te gösterilmiştir.



Şekil 15. Kümelere ilişkin kalibrasyon grafiği

Kalibrasyon grafiği oluşturma sürecinin başlangıç aşamasında her öğrenci (gözlem/birim) ayrı bir küme olarak kabul edilmektedir ve dolayısıyla kümelerin sayısı gözlem sayısına eşit olmaktadır. Algoritmanın bir sonraki adımında bu kümelerden benzerlik gösterenler tek bir küme oluncaya ya da istenen özellikleri sağlayana kadar birleştirilmektedir. Birleştirme sürecinde, kümelerin hem uzaklıkları hem de haritadaki konumları dikkate alınmaktadır. Nöronların bir araya gelmesiyle oluşan kalibrasyon grafiğinde harita yüzeyinde bulunan kümeler bitişiktir. Fakat analizde kullanılan girdi değişkenlerine ilişkin dağılımlara bağlı olarak harita yüzeyindeki bitişikliklerde farklılıklar olabilmektedir. Böyle bir durumda kümelerin homojenliğine ilişkin olumsuz durumlar oluşabilir. Bitişik kümeler oluşturabilmek için, analiz sonucunda elde edilen harita üzerinde birbirine benzer ve birbirine yakın nöronları birleştirme amacıyla hiyerarşik kümeleme yöntemi kullanılmaktadır (Kohonen, 2001). Bu kapsamda, alanyazında en fazla kullanılan yöntem Yığınsal Hiyerarşik Kümeleme Yöntemidir (Hierarchical Agglomerative Clustering Method). Kohonen'in Öz Örgütlemeli Harita Yöntemi ve yığınsal hiyerarşik kümeleme yönteminin birlikte kullanıldığı bu adım "Kalibrasyon'' aşaması olarak adlandırılmaktadır. Kalibrasyon aşaması ile Kohonen'in Öz Örgütlemeli Harita Yöntemi kullanılarak oluşturulan nöron sınıfları dağılımının

sıkıştırılmış bir temsiline ilişkin görsel ve homojen alanlardaki nöron sınıflarının güvenirliğine ilişkin bilgi elde edilmektedir. Kümeleme sonuçları yığınsal hiyerarşik kümeleme yöntemine ilişkin çizim fonksiyonu kullanılarak görselleştirilmekte ve sınırlar (kalın ve siyah çizgiler ile gösterilmektedir) istatistiksel anlamda en doğru şekilde belirlenmektedir. Şekil 11 incelendiğinde açık mavi, açık yeşil, koyu mavi, koyu yeşil olmak üzere dört farklı renkte kümelenme olduğu için ideal küme sayısının 4 olduğu belirlenmiştir. Bunlardan mavi renkle gösterilen kümede yer alan birim sayısının en fazla olduğu göze çarpmaktadır. Bunun yanında açık yeşil renkle gösterilen kümenin eleman sayısının ise en az olduğu görülmektedir. Kümelerde yer alan birim (öğrenci) sayılarına ilişkin elde edilen analiz sonuçları Tablo 5'te gösterilmiştir.

Tablo 5

Öğrenci Sayısının Ülkelere ve Kümelere İlişkin Dağılımı

Ülkeler	Küme 1	Küme 2	Küme 3	Küme 4
Japonya	**83**	7	136	39
Kore	**64**	6	129	*25*
Estonya	54	**16**	*120*	*33*
Belçika	**87**	**16**	**239**	44
Kanada	**165**	**14**	**396**	*28*
Avusturya	52	10	172	46
Çek Cumhuriyeti	51	**21**	165	39
Norveç	40	8	*113*	57
Lüksemburg	38	10	*120*	44
Hollanda	38	*3*	148	*26*
Almanya	45	*4*	187	*24*
Slovakya	43	5	157	49
İtalya	**78**	11	**279**	**96**
İspanya	45	5	146	73
İzlanda	21	*4*	*70*	40
Fransa	36	11	159	38
Finlandiya	32	2	151	51
İsrail	35	6	169	54
Macaristan	29	*2*	156	40
Danimarka	34	**17**	163	72
İsviçre	27	11	157	40
Yunanistan	25	*3*	119	74
İngiltere	63	*10*	321	**172**
Şili	29	11	**165**	78
İsveç	27	11	157	40
Polonya	*16*	*4*	*95*	64
Avustralya	52	10	172	46
Amerika	*20*	11	121	76
Litvanya	*17*	5	*109*	63
Portekiz	24	7	177	**85**
Yeni Zelanda	29	11	**214**	**108**
İrlanda	*18*	10	*120*	81
Türkiye	*17*	6	137	76
Meksika	21	13	150	**119**
Toplam	1455	301	5589	2040

Öğrenci sayısının ülkelere ve kümelere ilişkin dağılımını gösteren Tablo 5 incelendiğinde, birinci kümeye en fazla öğrenci veren ilk beş ülkenin sırasıyla Kanada (165), Japonya (83), Belçika (87), İtalya (78) ve Kore (64); en az öğrenci veren ilk beş ülkenin ise Polonya (16), Litvanya (17), Türkiye (17), İrlanda (18) ve Amerika (20) olduğu görülmektedir. İkinci kümeye en fazla öğrenci veren ilk beş ülkenin sırasıyla Çek Cumhuriyeti (21), Danimarka (17), Estonya (16), Belçika (16), Kanada (14); en az öğrenci veren ilk yedi ülkenin ise Macaristan (2), İngiltere (2), Hollanda (3), Yunanistan (3), Almanya (4), İzlanda (4) ve Polonya (4) olduğu görülmektedir. Üçüncü kümeye en fazla öğrenci veren ilk beş ülkenin sırasıyla Kanada (396), Şili (321), İtalya (279), Belçika (239), Yeni Zelanda (214); en az öğrenci veren ilk yedi ülkenin ise İzlanda (70), Polonya (95), Litvanya (109), Norveç (113), İrlanda (120), Estonya (120) ve Lüksemburg (120) olduğu görülmektedir. Dördüncü kümeye en fazla öğrenci veren ilk beş ülkenin sırasıyla İngiltere (172), Meksika (119), Yeni Zelanda (108), İtalya (96), Portekiz (85); en az öğrenci veren ilk beş ülkenin ise Almanya (24), Kore (25), Hollanda (26), Kanada (28) ve Estonya (33) olduğu görülmektedir. Tablo 5 incelendiğinde kümeleme analizi sonucunda ideal küme sayısı olarak belirlenen 4 farklı kümeye atanan toplam öğrenci sayısının 9385 olduğu belirlenmiştir. Tüm veri setindeki öğrenci sayısının 9870 olması sebebiyle Kohonen'in Öz Örgütlemeli Harita Yöntemi ile doğru sınıflama oranının %95,08 (9385/9870) olduğu tespit edilmiştir. Kümelerde yer alan öğrenci sayılarına bağlı olarak ülke bazında öğrencilerin kümelere dağılımlarına ilişkin yüzde değerleri Tablo 6'da gösterilmiştir.

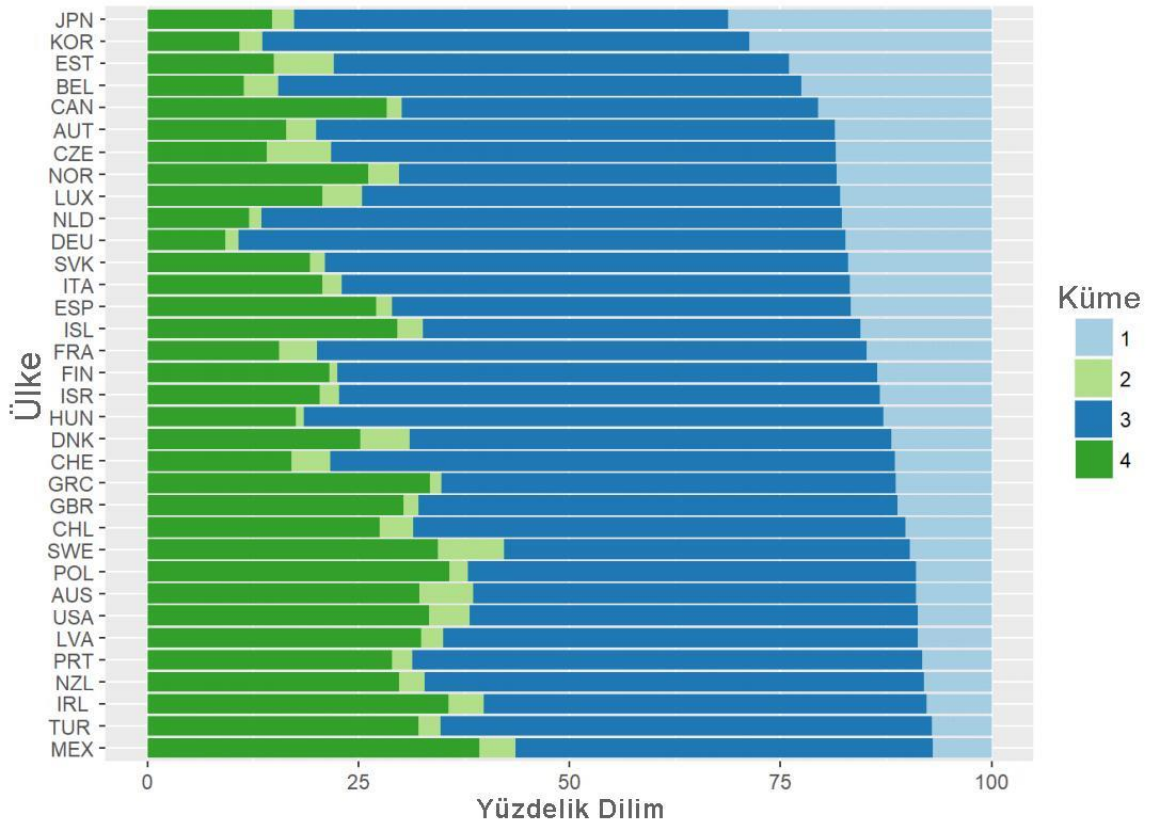
Tablo 6

Ülkelere ve Kümelere İlişkin Öğrenci Yüzdelerinin Dağılımı

Ülkeler	Küme 1	Küme 2	Küme 3	Küme 4
Japonya	**31.25**	2.67	*51.33*	14.73
Kore	**28.67**	2.64	57.73	*10.94*
Estonya	**24.01**	**7.08**	53.93	14.96
Belçika	**22.53**	4.14	61.91	*11.39*
Kanada	**20.52**	1.74	*49.34*	28.38
Avusturya	18.57	3.57	61.42	16.42
Çek Cumhuriyeti	18.47	**7.60**	59.78	*14.13*
Norveç	18.34	3.66	*51.83*	26.14
Lüksemburg	17.92	4.71	56.60	20.75
Hollanda	17.67	*1.39*	**68.83**	*12.09*
Almanya	17.30	*1.53*	**71.92**	*9.23*
Slovakya	17.04	1.79	61.88	19.28
İtalya	16.81	2.37	60.12	20.68
İspanya	16.72	1.85	54.27	27.13
İzlanda	15.55	2.96	51.85	29.62
Fransa	14.75	4.50	**65.16**	15.57
Finlandiya	13.55	*0.84*	63.98	21.61
İsrail	13.25	2.27	64.01	20.45
Macaristan	12.77	*0.88*	**68.72**	17.62
Danimarka	11.88	**5.94**	56.99	25.17
İsviçre	11.48	4.68	**66.80**	17.02
Yunanistan	11.31	*1.35*	53.84	**33.48**
İngiltere	11.13	1.76	56.71	30.38
Şili	10.24	3.88	58.30	27.56
İsveç	9.63	**7.79**	*48.16*	**34.40**
Polonya	8.93	2.23	53.07	**35.74**
Avustralya	8.89	**6.35**	52.54	32.20
Amerika	8.77	4.82	53.07	33.30
Litvanya	8.76	2.57	56.18	32.47
Portekiz	*8.19*	2.38	60.40	29.01
Yeni Zelanda	*8.01*	3.03	59.11	29.83
İrlanda	*7.72*	4.23	52.42	**35.61**
Türkiye	*7.04*	2.57	58.24	32.13
Meksika	*6.93*	4.29	*49.50*	**39.27**

Ülkelere ve kümelere ilişkin öğrenci yüzdelerinin dağılımını gösteren Tablo 6 incelendiğinde, yüzde olarak birinci kümeye en fazla öğrenci veren ilk beş ülkenin sırasıyla Japonya (31,25), Kore (28,67), Estonya (24,01), Belçika (22,53), Kanada (20,52); en az öğrenci veren ilk beş ülkenin ise sırasıyla Meksika (6,93), Türkiye (7,04), İrlanda (7,72), Yeni Zelanda (8,01) ve Portekiz (8,19) olduğu görülmektedir. Yüzde olarak ikinci kümeye en fazla öğrenci veren ilk beş ülkenin sırasıyla İsveç

(7,79), Çek Cumhuriyeti (7,60), Estonya (7,08), Avustralya (6,35) ve Danimarka (5,94); en az öğrenci veren ilk beş ülkenin ise Finlandiya (0,84), Macaristan (0,88), Yunanistan (1,35), Hollanda (1,39), Almanya (1,53) olduğu görülmektedir. Yüzde olarak üçüncü kümeye en fazla öğrenci veren ilk beş ülkenin sırasıyla Almanya (71,92), Hollanda (68,83), Macaristan (68,72), İsviçre (66,80), Fransa (65,16); en az öğrenci veren ilk beş ülkenin ise sırasıyla İsveç (48,16), Kanada (49,34), Meksika (49,50), Japonya (51,33) ve Norveç (51,83) olduğu görülmektedir. Yüzde olarak dördüncü kümeye en fazla öğrenci veren ilk beş ülkenin sırasıyla Meksika (39,27), Polonya (35,74), İrlanda (35,61), İsveç (34,40), Yunanistan (33,48); en az öğrenci veren ilk beş ülkenin ise sırasıyla Almanya (9,23), Kore (10,94), Belçika (11,39), Hollanda (12,09), Çek Cumhuriyeti (14,13) olduğu görülmektedir Bununla birlikte R programı kapsamında her bir ülkenin elde edilen 4 kümede yer alan öğrenci dağılımlarına ilişkin sütun grafikleri elde edilebilmektedir. Ülkelerin kümelerde yer alan öğrenci dağılımları Şekil 16'da gösterilmiştir.



Şekil 16. Öğrenci sayısının ülkelere ve kümelere ilişkin dağılımı

Öğrenci sayısının ülkelere ve kümelere ilişkin dağılımını gösteren Şekil 16 incelendiğinde, tüm ülkeler için en fazla koyu mavi ile gösterilen 3. kümede

öğrencilerin olduğu görülmektedir. Bunun ardından ülkelerin kümelere göre dağılımları sırasıyla koyu yeşil ile gösterilen 4. kümenin ikinci sırada, açık mavi ile gösterilen 1. kümenin üçüncü sırada ve açık yeşil ile gösterilen 2. kümenin ise son sırada olduğu görülmektedir. Başka bir ifadeyle ülkelerdeki öğrencilerin kümelerde yer alma oranları büyükten küçüğe doğru sırasıyla 3. küme, 4. küme, 1. küme ve 2.küme şeklindedir. Küme bazında inceleme yapıldığında kümede yer alan toplam öğrenci bakımından en büyük küme olan ve koyu mavi ile gösterilen 3. kümede Kanada, İngiltere, İtalya, Belçika ve Yeni Zelanda'nın bu kümeye en fazla öğrenci veren ülkeler olduğu görülmektedir. Bununla birlikte öğrenci sayısı bakımından ikinci sırada yer alan ve koyu yeşil ile gösterilen 4.kümede Meksika, Polonya, İrlanda, İsveç, Yunanistan ve Amerika'nın bu kümeye en fazla öğrenci veren ülkeler olduğu görülmektedir.

Kümelerin oluşmasında değişken olarak kullanılan faktör puanlarının her bir küme için nasıl bir etkiye sahip olduğunu belirlemek amacıyla küme profilleri elde edilmiştir. Tablo 7'de, Kohonen'in Öz Örgütlemeli Harita Yöntemi kapsamında dört farklı küme için faktör puanlarının ne düzeyde etkiye sahip olduğu gösterilmiştir.

Tablo 7

Kümeler için Alt Boyutlara ve Olası Başarı Puanına İlişkin Faktör Puanı Ortalamaları

		Öğretmen Yönetimindeki Fen Bilgisi Öğretimi	Fen Bilgisi Öğretmenlerinden Alınan Geri Bildirim	Fen Bilgisi Derslerine İlişkin Uyarlanabilir Öğretim	Sorgulama Temelli Fen Bilgisi Öğretimi	Olası Fen Başarı Puan Ortalaması
Küme	1	0.050	0.010	-0.070	-0.070	0,07897
	2	-0.010	0.030	-0.030	0.000	0,01284
	3	0.010	0.010	-0.020	0.020	0,08827
	4	-0.030	0.020	0.070	0.070	-0,08457
	Ort.	0.005	0.010	-0.008	0.002	0,02388

Tablo 7'de yer alan olası başarı puanı ortalamasına ilişkin z puanları incelendiğinde, çoğunluğunu olası başarı puanı en fazla olan öğrencilerin oluşturduğu kümenin

üçüncü küme olduğu görülmektedir. Üçüncü kümeyi sırasıyla birinci, ikinci ve dördüncü küme takip etmektedir.

Tablo 7 incelendiğinde, üçüncü küme için faktör 4'ün faktör puan ortalaması değerinin (0.02) 0.002 ortalama değerinden çok daha büyük olması, sorgulama temelli fen bilgisi öğretiminin üçüncü kümede en fazla etkiye sahip olduğunun bir göstergesi olarak yorumlanabilir. Bu durumun tam tersi olarak üçüncü küme için faktör 3'ün aldığı faktör puan ortalaması değerinin (-0.02) -0.008 ortalama değerinden büyük olması, fen bilgisi derslerine ilişkin uyarlanabilir öğretimin üçüncü kümede en az etkiye sahip olduğunun bir göstergesi olarak yorumlanabilir. Kısacası, üçüncü kümedeki öğrencilerin sorgulama temelli fen bilgisi öğretimini benimsedikleri, fen bilgisi derslerine ilişkin uyarlanabilir öğretimi ise çok az benimsedikleri tespit edilmiştir.

Tablo 7 incelendiğinde, birinci küme için faktör 1'in aldığı faktör puan ortalaması değerinin (0.05) 0.005 ortalama değerinden çok daha büyük olması öğretmen yönetimindeki fen bilgisi öğretiminin birinci kümede en fazla etkiye sahip olduğunun bir göstergesi olarak yorumlanabilir. Bu durumun tam tersi olarak, birinci küme için faktör 3'ün aldığı faktör puan ortalaması değerinin (-0.07) -0.008 ortalama değerinden çok daha büyük olması, fen bilgisi derslerine ilişkin uyarlanabilir öğretimin üçüncü kümede en az etkiye sahip olduğunun bir göstergesi olarak yorumlanabilir Kısacası, birinci kümedeki öğrencilerin öğretmen yönetimindeki fen bilgisi öğretimini benimsedikleri, fen bilgisi derslerine ilişkin uyarlanabilir öğretimi ise çok az benimsedikleri tespit edilmiştir.

Tablo 7 incelendiğinde, ikinci küme için faktör 2'nin faktör puan ortalaması değerinin (0.03) 0.01 ortalama değerinden çok daha büyük olması, fen bilgisi öğretmenlerinden alınan geri bildirimin ikinci kümede en fazla etkiye sahip olduğunun bir göstergesi olarak yorumlanabilir. Bu durumun tam tersi olarak ikinci küme için faktör 3'ün aldığı faktör puan ortalaması değerinin (-0.03) -0.008 ortalama değerinden büyük olması, fen bilgisi derslerine ilişkin uyarlanabilir öğretimin ikinci kümede en az etkiye sahip olduğunun bir göstergesi olarak yorumlanabilir. Kısacası, ikinci kümedeki öğrencilerin fen bilgisi öğretmenlerinden alınan geri bildirimi dikkate aldıkları, 'fen bilgisi derslerine ilişkin uyarlanabilir öğretimi ise çok az benimsedikleri tespit edilmiştir.

Tablo 7 incelendiğinde, dördüncü küme için faktör 3'ün faktör puan ortalaması değerinin (0.07) -0.008 ortalama değerinden çok daha büyük olması, fen bilgisi derslerine ilişkin uyarlanabilir öğretimin dördüncü kümede en fazla etkiye sahip olduğunun bir göstergesi olarak yorumlanabilir. Bu durumun tam tersi olarak, dördüncü küme için faktör 1'in aldığı faktör puan ortalaması değerinin (-0.03) 0.005 ortalama değerinin çok altında olması, öğretmen yönetimindeki fen bilgisi öğretiminin dördüncü kümede en az etkiye sahip olduğunun bir göstergesi olarak yorumlanabilir. Kısacası, dördüncü kümedeki öğrencilerin fen bilgisi derslerine ilişkin uyarlanabilir öğretimi benimsedikleri; öğretmen yönetimindeki fen bilgisi öğretimi ise çok az benimsedikleri tespit edilmiştir.

Elde edilen sonuçlar bir bütün olarak değerlendirildiğinde, üçüncü kümede yer alan öğrencilerin genel anlamda sorgulama temelli fen bilgisi öğretimini benimsediği ve diğer kümelerde yer alan öğrencilere göre olası fen başarı puanı daha yüksek öğrenciler oldukları tespit edilmiştir.

Kohonen'in Öz Örgütlemeli Harita Yöntemi esas alınarak yapılan çalışmalardan biri olan Taşkın ve Emel (2010) tarafından gerçekleştirilen araştırmada parakendecilik sektöründe bir kümeleme uygulaması gerçekleştirmişlerdir. Çalışmada bir işletmenin 10000 adet müşterisine ait veri tabanı kullanılmış ve Kohonen'in Öz Örgütlemeli Harita Yöntemi tekniği ile kümeleme gerçekleştirilmiştir. Clementine v8 programında gerçekleştirilen çalışmada sekiz farklı küme elde edilirken kümede etkili olan değişkenlerin önem dereceleri de rapor edilmiştir. Benzer bir çalışmada Oğuzlar (2005) Bursa Emniyet Müdürlüğünden alınan veriler yardımıyla suçlu profilinin belirlenmesi amacıyla Kohonen'in öz örgütlemeli harita yöntemini kullanmıştır. Çalışmada Clemetine 7.0 programı kullanılmış ve suçlu profillerine ilişkin tanımlamalar C5.0 kural algoritması kullanılarak toplam 12 ayrı kümenin elde edildiği sonucuna ulaşılmıştır. Özşahin ve Yüreğir (2012) tarafından gerçeleştirilen çalışmada Türkiye'de otomotiv sektöründe faaliyet gösteren 6 işletmenin bilanço ve gelir tablolarından elde edilen finansal oranları kullanılarak kümelenme yöntemi gerçekleştirilmeye çalışılmıştır. Çalışmada araştırmacılar tarafından Delphi 7.0 programlama dili kullanılarak bir yazılım geliştirilmiştir. Özçalıcı (2016) yaptığı çalışmasında BIST50 Endeksinde yer alan 50 adet hisse senedinin günlük standartlaştırılmış getiri ve risk değerlerini kullanarak hisse senetlerin nasıl kümelendiğini incelemiştir. Çalışmda Matlab programı kullanılarak Kohonen'in Öz

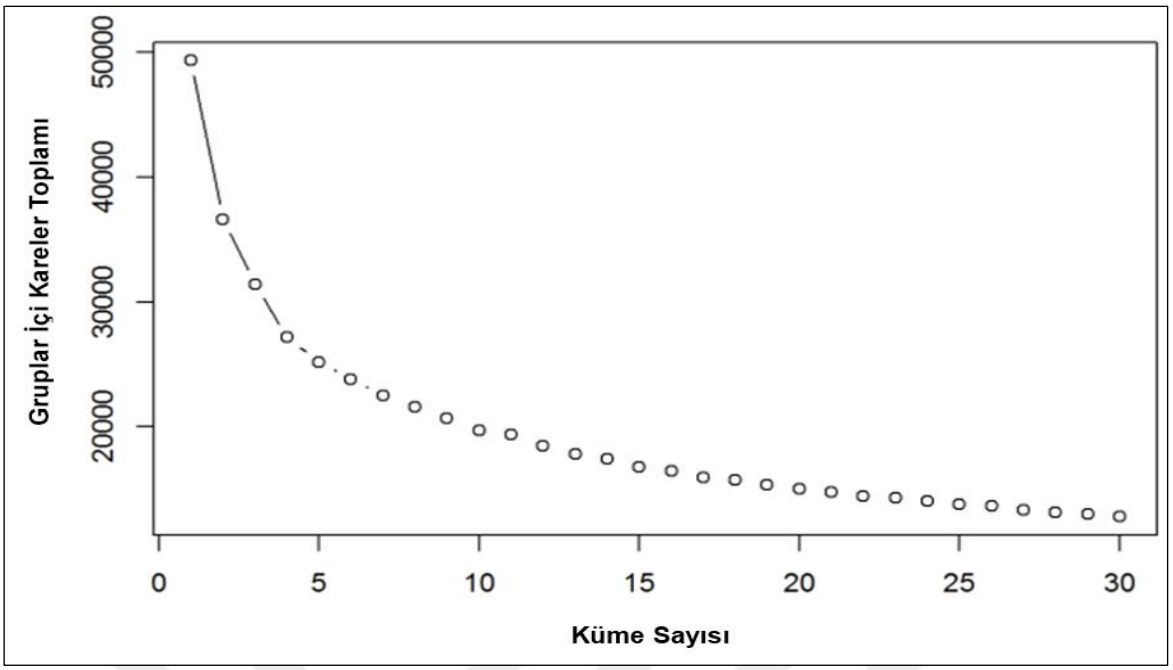
Örgütlemeli Harita Yöntemi ile iki farklı küme elde edilmiştir. İnce, İmamoğlu ve Kesin (2013) tarafından yapılan çalışmada tüketicilerin alışveriş motivasyon ve değerleri ile birlikte karar verme stillerine dayalı profilini çıkarmak için Kohonen'in öz örgütlemeli harita yöntemi kullanılmıştır. SOMToolbox yazılımı kullanılarak gerçekleştirilen çalışmada Kohonen'in Öz Örgütlemeli Harita Yönteminin K-Ortalamalar Yöntemine göre daha iyi olduğu sonucuna ulaşılmıştır. Birinci alt problem sonucunda üçüncü kümede etkili olan değişkenin sorgulama temelli fen bilgisi öğretimi ve fen başarı puanı olduğu belirlenmiştir. Bu sonuç, sorgulama temelli fen bilgisi öğretiminin öğrenci başarısını olumlu yönde etkilediğine ilişkin bulgular barındıran araştırma sonuçlarıyla benzerlik göstermektedir (Furtak ve diğ., 2012; Minner ve Levy, 2010; Schroeder ve diğ., 2007).

Kohonen'in Öz Örgütlemeli Harita Yöntemiyle ilgili çalışmalar bir bütün olarak incelendiğinde analizlerin farklı programlar ve yazılımlar tarafından gerçekleştirilirken eğitim bilimleri alanında açık kaynak kodlu R programlama dili kullanılarak yapılmış herhangi bir çalışmaya ait görgül araştırma bulgusuna rastlanamamıştır. Çalışmanın bu yönüyle alanyazında yapılan diğer çalışmalardan farklılık gösterdiği belirlenmiştir. Bunun yanında çalışmanın tek bir yazılım ya da program kullanmak yerine birden fazla kümeleme yöntemini aynı anda kullanarak elde edilen sonuçların incelenmesi bakımından diğer çalışmalardan farklılık gösterdiği belirlenmiştir.

İkinci Alt Probleme İlişkin Bulgular ve Yorumlar

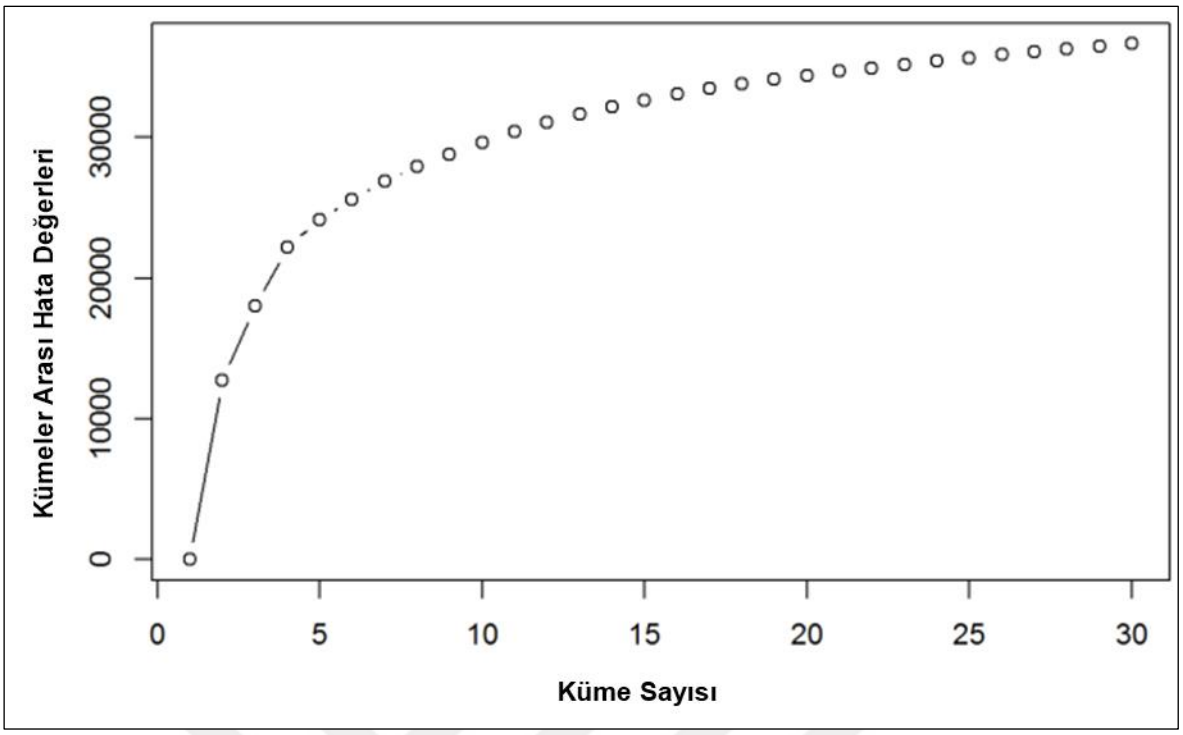
Çalışmanın ikinci alt probleminde K-Ortalamalar Yöntemi ile elde edilen küme sayıları ve kümelere ilişkin istatistikler rapor edilmiştir

K-ortalamalar algoritmasının en ideal küme sayısını bulması kesin olmadığından, algoritmanın çalıştırılması anlamında farklı parametreler ve rastgele örnekler kullanılarak analiz birçok kez gerçekleştirilmiştir. R programında k-ortalamalar algoritması kapsamındaki "nstart" parametresi, denemenin rasgele başlatılmasının belirlenmesine olanak sağlar, ayrıca her rasgele başlatma için algoritmaya izin verilen maksimum yineleme sayısını ayarlama anlamında iter.max parametresi de kullanılmaktadır. Bu nedenle farklı küme sayıları için gruplar içindeki kareler toplamının nasıl bir değişim gösterdiğinin incelenmesi gerekmektedir. Şekil 17'de küme sayısının değişiminin grup içi kareler toplamı ile değişimi gösterilmektedir.



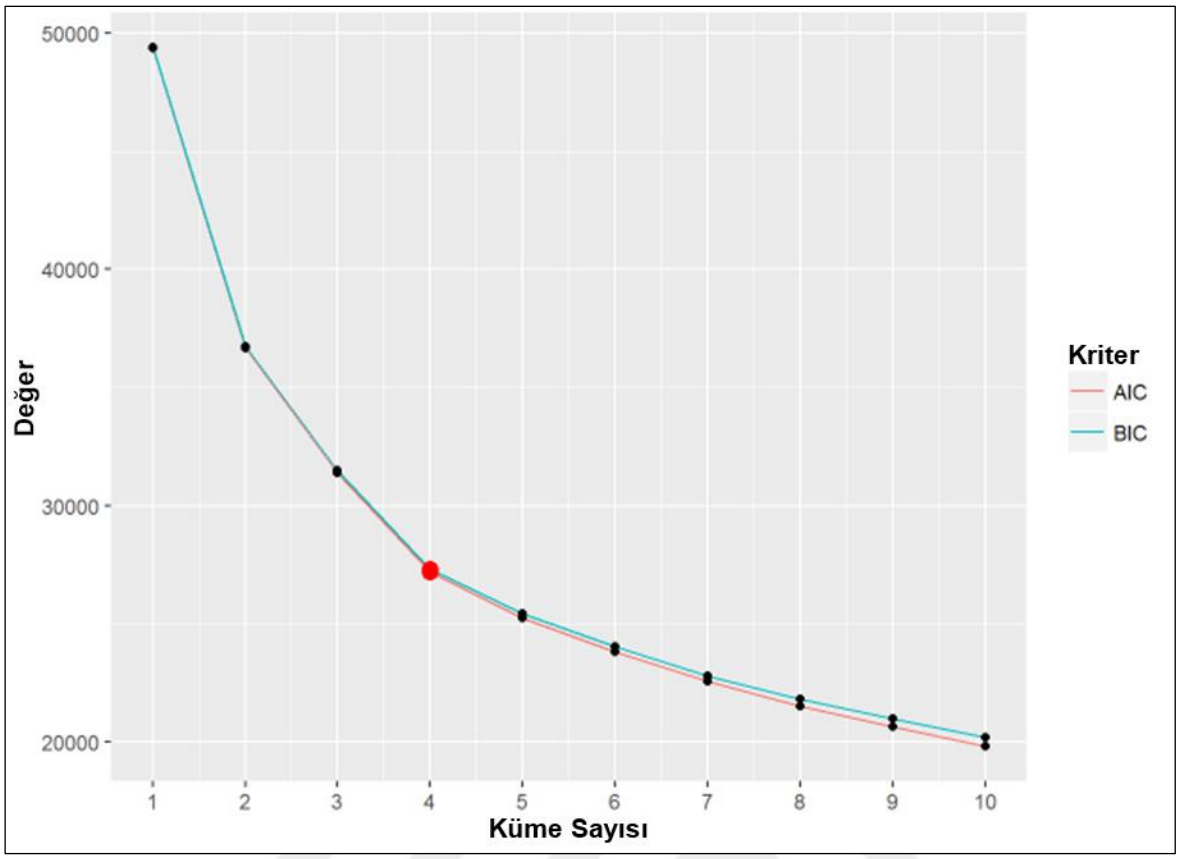
Şekil 17. Küme sayısı ile grup için kareler toplamının değişimi

Şekil 17 incelendiğinde kareler toplamının küme sayısı arttıkça azaldığı görülmektedir. Grup içi kareler toplamı hata değeri olarak da kabul edilmekte ve grafikte yer alan her iki nokta arası mesafe oluşan küme sayısını göstermektedir. Hata değerleri küme sayısı arttıkça azaldığından ideal küme sayısını belirlemede hata değerlerinin çok fazla değişme göstermediği nokta referans olarak alınabilmektedir. Buna göre küme sayısı dört oluncaya kadar hata değerlerinde hızlı bir azalma olmakla birlikte küme sayısı beş veya daha fazla olduğunda hata değerlerindeki değişimin çok fazla olmadığı görülmektedir. Bu sonuca göre ideal küme sayısının dört olabileceği düşünülmektedir. Bu sonucun doğrulanabilmesi amacıyla kümeler arasındaki kareler toplamının da incelenmesi gerekmektedir. Şekil 18'de küme sayılarına göre kümeler arası hata değerlerinin nasıl bir değişkenlik gösterdiği görülmektedir.



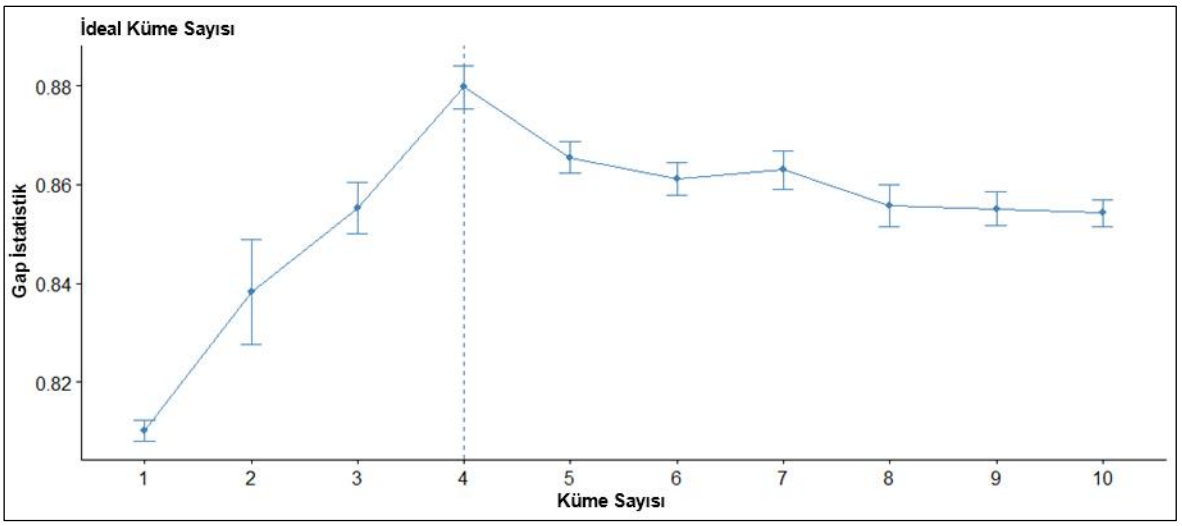
Şekil 18. Küme sayılarına göre kümeler arası hata değerlerinin değişimi

Farklı küme sayıları için elde edilen kümeler arasındaki hata değerlerinin değişimini gösteren Şekil 18 incelendiğinde, küme sayısı arttıkça kümeler arası hata değerlerinin de arttığı görülmektedir. İdeal küme sayısı belirlemek için daha önceki grafiklerde olduğu gibi hata değerleri arasındaki farkın çok fazla olmadığı noktanın referans olarak alınabileceği belirtilmektedir. Bir kümenin diğer kümelerden ne kadar iyi ve kesin bir şekilde ayrıştığının ölçütü olarak kabul edilen bu grafikte küme sayısı beş ve daha fazla olduğunda kümelerin çok iyi ayrışmadığı görülmektedir. Elde edilen bu sonuca göre ideal küme sayısının 4 (dört) olacağı düşünülmektedir. Bunlara ek olarak ideal küme sayısını belirlemek için oluşan kümelere ilişkin ABK ile BBK'nın farklı küme sayıları için nasıl bir değişkenlik gösterdiğinin incelenmesi gerekmektedir. Şekil 19'da her iki kriter için elde edilen çizgi grafiği gösterilmiştir.



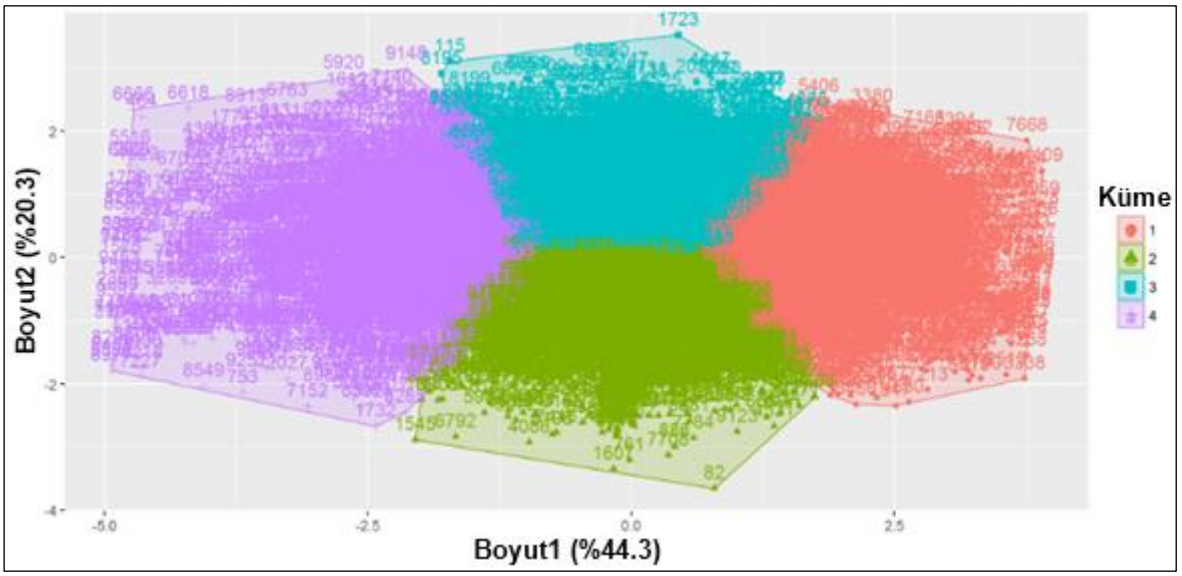
Şekil 19. Küme sayılarına göre ölçüt değerlerinin değişimi

İdeal küme sayısını belirlemek için ele alınan ABK ve BBK ölçütlerinin küme sayısına göre değişimini gösteren Şekil 19 incelendiğinde, her iki ölçütün de küme sayısı arttıkça azaldığı göze çarpmaktadır. Burada istenilen, çizgi grafiğinin düz bir plato şeklinde olmasıdır. Buna göre grafikteki azalma miktarının küme sayısı 4 (dört) ve daha fazla old uğunda nispeten daha az olduğu söylenebilir. Bunların yanında farklı küme sayıları için kümeler arası mesafeleri esas alarak elde edilen ve ideal küme sayısını belirleme yöntemlerinden biri olan grafik Şekil 20'de gösterilmiştir.



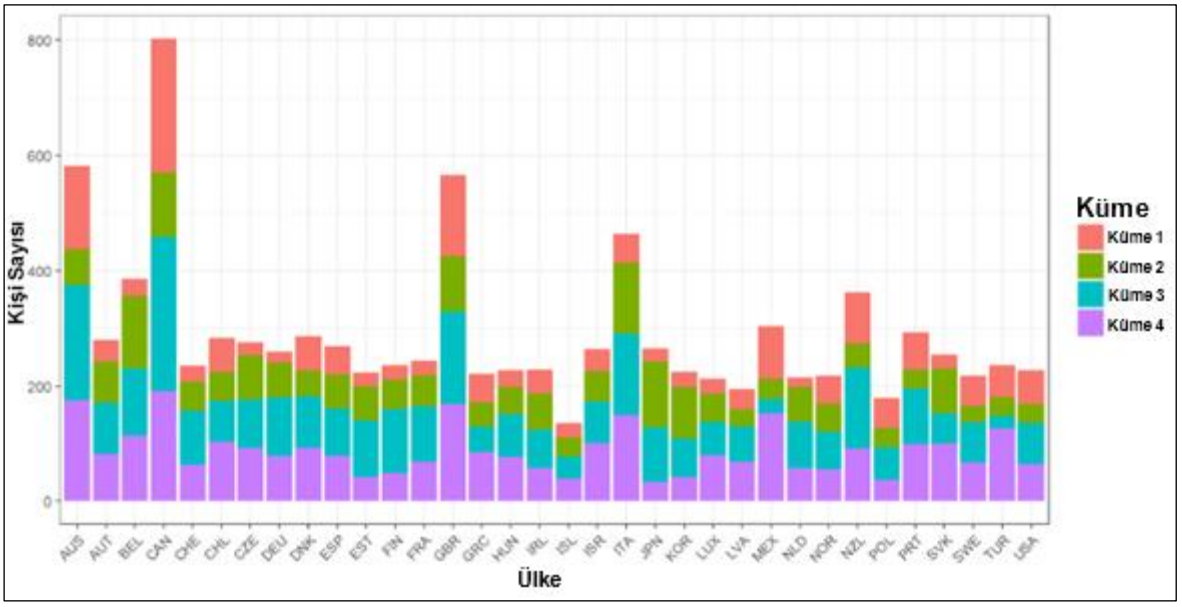
Şekil 20. Kümeler arası mesafelere ilişkin grafik

Farklı küme sayıları için küme içi değişkenlik katsayılarının değişimine ilişkin alt ve üst değerlerin gösterildiği Şekil 20 incelendiğinde küme sayı k=4 olduktan sonra grup içi değişkenliğin giderek azaldığı görülmektedir. Gap istatistik değerinin en büyük olduğu noktanın ideal küme sayısını göstereceği bilinmektedir. Buna göre ideal küme sayısının 4 olacağı düşünülmektedir. Her ne kadar ideal sayısının belirlenmesi sezgisel olarak kabul edilse de çalışmada son olarak verilen kümelere ilişkin renk dağılımları grafiği ideal küme sayısını belirlemede en etkili yöntem olarak görülmektedir. Şekil 21'de çalışma kapsamında elde edilen kümelere ilişkin dağılım grafiği gösterilmektedir.



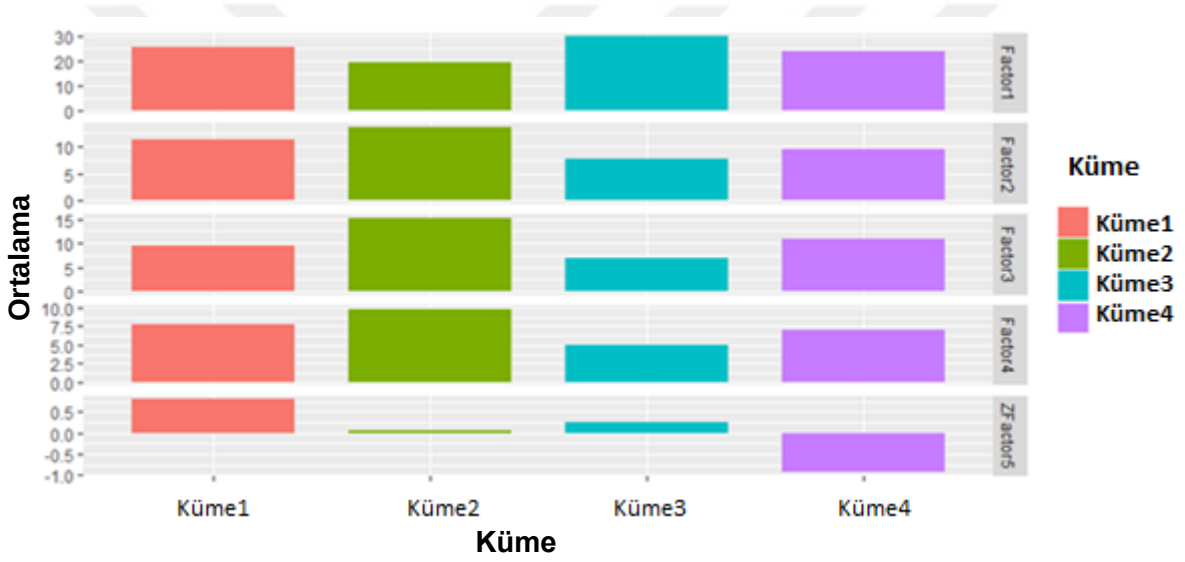
Şekil 21. Kümelere ilişkin dağılım grafiği

Şekil 21'de x ve y eksenlerinde yer alan değişkenler kümeleme analizinde en etkili olan başka bir ifadeyle açıkladığı varyans miktarı en fazla olan ilk iki değişkene Boyut1 (Dim1) ve Boyut2 (Dim2) ilişkin elde edilen kümeleri göstermektedir. Buna göre öğrencileri kümelere ayırmada birinci değişkenin toplam varyansın %44,30'unu ve ikinci değişkenin toplam varyansın %20,30'unu açıkladığı görülmektedir. Kümelere ayırmada ele alınan ilk iki değişkenin açıkladığı toplam varyans miktarı %64,60 ve bu oranın yeterli olduğu düşünülmektedir. Buna göre grafikte mor, yeşil, mavi ve turuncu renklerle gösterilen toplam 4 küme oluştuğu belirlenmiştir. Bu renklerin geçiş noktalarında birbiri üzerinde farklı renklerin olması üst üste binme sorunu (overlapping) olarak adlandırılmaktadır. Şekil 21'de verilen grafikte birbiri üzerine binen renklerin çok fazla olmaması kümelerin iyi bir şekilde ayrıştığını göstermektedir. Daha önce Kohenen'in Öz Örgütlemeli Harita Yöntemiyle kümelerde yer alan öğrenci sayılarının toplamı 9375 olarak belirlenmiş ve veri setindeki toplam öğrenci sayısının 9870 olduğu göz önüne alındığında kümelerde yer alan bireyleri doğru sınıflama oranının %95,08 (9375/9870) olduğu tespit edilmiştir. Şekil 21'de üst üste binen renklerin hangi kümeye ait olduğu tam olarak belirlenemeyen toplam 485 öğrenciye ait olduğu düşünülmektedir. Elde edilen sonuçlar bir bütün olarak incelendiğinde veri setindeki öğrencilerin toplam 4 kümeye ayrıştığı sonucuna ulaşılmıştır. Bunlara ek olarak her bir ülke için elde edilen 4 farklı kümede ilgili ülkeden o kümede kaç öğrencinin yer aldığına ilişkin sütun grafiği Şekil 22'de gösterilmiştir.



Şekil 22. Ülkelerin kümelerde yer alan öğrenci sayılarına ilişkin sütun grafiği

Toplam 34 ülkenin K-Ortalamalar Kümeleme Analizi ile belirlenen dört farklı kümede yer alan öğrenci sayıları Şekil 22'de görülmektedir. Örneklemde en fazla öğrenci sayısına sahip Kanada için öğrencilerin en fazla üçüncü ve en az ikinci kümede yer aldıkları görülmektedir. Yine benzer şekilde Avustralya örnekleminde öğrencilerin en fazla üçüncü ve en az ikinci kümede yer aldıkları görülmektedir. Bunun yanında örneklemde en az öğrenci sayısına sahip İsrail için kümelere göre öğrenci sayıları incelendiğinde öğrencilerin en fazla üçüncü ve en az birinci kümede yer aldıkları görülmektedir. Kümelere ayrışmada ele alınan faktör puanlarının her bir küme için nasıl bir etkiye sahip olduğunu belirlemek amacıyla küme profilleri elde edilmiştir. Şekil 23'te belirlenen dört farklı küme için faktör puanlarının ne düzeyde etkiye sahip olduğu gösterilmiştir.



Şekil 23. Kümelere ilişkin profil puanları dağılımı

Şekil 23 incelendiğinde birinci kümede en fazla etkiye öğretmen yönetimindeki fen bilgisi öğretimi (faktör 1) etkili olmuştur. Bununla birlikte Fen başarı puanı (faktör 5) en yüksek öğrencilerin birinci kümede yer aldığı belirlenmiştir. İkinci kümede en fazla etkiye Öğretmen yönetimindeki fen bilgisi öğretimi (faktör 1) etkili olmuştur. Bununla birlikte Fen Bilgisi öğretmenlerinden alınan geri bildirim (faktör2), Fen bilgisi derslerine ilişkin uyarlanabilir öğretim (Factor3) ve Sorgulama temelli fen bilgisi öğretimi (Factor4) değişkenlerinin en fazla bu kümede etkili oldukları görülmektedir. Üçüncü kümede en fazla etkiye Öğretmen yönetimindeki fen bilgisi öğretimi (faktör 1) sahip olmuştur. Ancak ikinci kümeyle karşılaştırıldığında bu kümedeki öğrencilerin fen başarı puanlarının daha yüksek olduğu görülmektedir. Dördüncü kümede en fazla etkiye Öğretmen yönetimindeki fen bilgisi öğretimi (faktör 1) sahip olmuştur. Bununla

birlikte fen başarı puanı en düşük olan öğrencilerin bu kümede yer aldığı görülmektedir.

Elde edilen sonuçlar bir bütün olarak değerlendirildiğinde birinci kümede yer alan öğrencilerin genel anlamda öğretmen yönetimindeki fen bilgisi öğretimini benimsediği ve diğer kümelerde yer alan öğrencilere göre fen başarı puanı daha yüksek öğrenciler oldukları tespit edilmiştir.

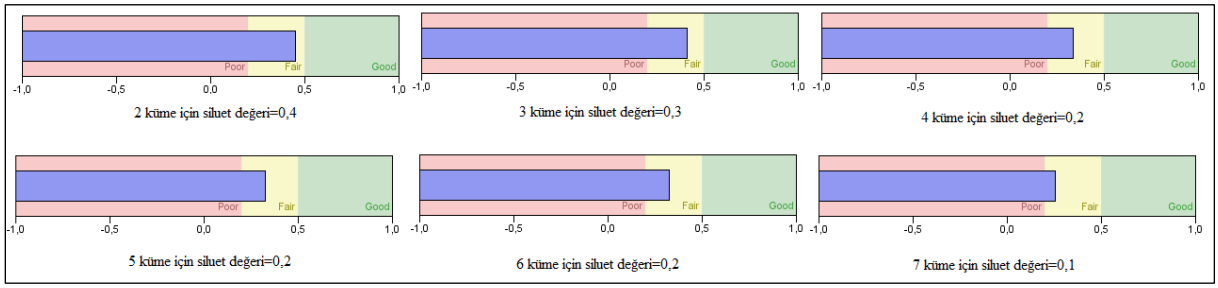
Alanyazında K-Ortalamalar Yöntemi kullanılarak yapılan çalışmalardan biri olan Çakmak (1999) tarafından gerçekleştirilen araştırmada kümeleme analizinin sonuçlarının diskriminant analizi yöntemiyle ne kadar uyumlu olduğu incelenmiştir. STATISTICA programında gerçekleştirilen çalışmada K-Ortalamalar Yöntemiyle 76 ilin iki kümeye ayrılması durumunda %100, altı kümeye ayrılması durumda %98.7 ve üç kümeye ayrılması durumda ise %97.4 doğru sınıflandırma sonucuna ulaşılmıştır. Ersöz (2009) tarafından yapılan çalışmada OECD ülkelerine ilişkin sağlık verileri girdi değişkenleri olarak ele alınarak K-Ortalamalar, hiyerarşik kümeleme ve k-medoid (noktaların merkeze olan uzaklıklarının temel alındığı kümeleme yöntemi) yöntemine ilişkin sonuçlar karşılaştırmıştır. SPSS 15.0 ve NCSS istatistik paket programları kullanıldığı çalışmada farklı yönteler kullanılarak elde edilen kümelerdeki ülkelerin farklılık gösterdiği belirlenmiştir. Şekerkaya ve Cengiz (2010) tarafından gerçekleştirilen çalışmada kadın tüketicilerin alışveriş merkezi tercihlerine göre kümelenmesi amaçlanmıştır. K-Ortalamalar Yönteminin kullanıldığı araştırma sonucuna göre kadınların AVM tercihlerine göre üç grupta kümelendiği ve bu kümelerin sahip oldukları nitelikler itibariyle potansiyeller, aktifler ve duyarsızlar olarak adlandırıldığı belirlenmiştir. Acar (2012) tarafından yapılan çalışmada PISA 2009 sonuçlarına göre Türkiye'nin OECD'ye üye ve aday ülkeler arasındaki yeri K-Ortalamalar Kümeleme ve ayırma analiziyle belirlenmeye çalışılmıştır. Kümeleme analizi sonuçlarına göre; 1. kümede dokuzu aday toplam 13 ülkenin, 2. kümede beşi OECD'ye aday toplam 30 ülkenin; 3.kümede beşi aday toplam 10 ülkenin ve 4. kümede hepsi aday toplam 12 ülkenin sınıflandığı görülmüştür. Çalışmada ayırma analizine göre doğru sınıflama yüzdesinin %96,9 oranında olduğu belirlenmiştir. Antonenko, Toy ve Niederhauser (2012) tarafından yapılan çalışmada öğrencilerin çevrimiçi bir öğrenme ortamında problem çözme aktivitesine katılırken öğrenme davranışının özelliklerini analiz etmek için K-Ortalamalar Yöntemi kullanmıştır. Verilerin analizinde SPSS programının kullanıldığı çalışmada K-Ortalamalar

Yönteminden elde edilen sonuçlar karşılaştırılmıştır. Analiz sonucunda K-Ortalamalar Yönteminin büyük örneklemlerde rahatlıkla kullanılabileceği sonucuna ulaşılmıştır. Cengiz ve Öztürk (2012) tarafından yapılan çalışmada eğitim yönleri itibariyle benzer özellikler gösteren illerin K-Ortalamalar Yöntemi kullanılarak nasıl kümelendiği incelenmiştir. SPSS paket programı kullanılarak gerçekleştirilen çalışmda siluet değerlerine göre ideal küme sayısının altı olduğu belirlenmiştir. K-Ortalamalar Yöntemiyle ilgili çalışmalar bir bütün olarak incelendiğinde analizlerin genellikle SPSS paket programı tarafından gerçekleştirildiği görülmektedir. Bunun yanında alanyazında açık kaynak kodlu R programlama dili kullanılarak yapılmış herhangi bir çalışmaya ait görgül araştırma bulgusuna rastlanamamıştır. Çalışmanın bu yönüyle alanyazında yapılan diğer çalışmalardan farklılık gösterdiği belirlenmiştir. Bunun yanında Aksu, Güzeller ve Eser (2017) tarafından yapılan çalışmada PISA 2012 öğrenci anketinde yer alan duyuşsal özelliklere göre ülkelerin öz yeterlik puanlarına göre 8, ilgi puanlarına göre 7 ve tutum puanlarına göre 6 farklı küme oluştuğu belirlenmiştir. PISA 2015 verileri kullanılarak yapılan çalışmada ise sistematik örnekleme yöntemiyle elde edilen 9837 öğrencinin K-Ortalamalar Yöntemiyle fen öğretimini etkileyen faktörler ile fen başarısına göre dört kümeye ayrıldığı ve bu yanıyla çalışmanın farklılık gösterdiği belirlenmiştir. İkinci alt problem sonucunda birinci kümede en etkili olan değişkenlerin öğretmen yönetimindeki fen bilgisi öğretimi ve fen başarı puanları olmasına karşın elde edilen bu sonuç Delen ve Bellibaş (2015) tarafından yapılan çalışmada öğretmen yönetimindeki fen bilgisi öğretimi ile fen başarısı arasında istatistiksel düzeyde manidar bir ilişki olmadığına dair bulgular ile farklılık göstermektedir. PISA 2012 verilerinin kullanıldığı çalışmada hiyerarşik doğrusal modelleme ile fen başarısı üzerinde biçimlendirici değerlendirme, öğretmen desteği, cinsiyet, sosyoekonomik durum ve okulun yeri değişkenlerinin etkili olduğu belirlenmiştir. Ancak, Hattie (2009) tarafından yapılan çalışmada başarı üzerinde en fazla etkiye sahip değişkenlerin eğitimin kalitesi ve öğretmen merkezli ders anlatımı olduğunun belirlenmesi çalışmada elde edilen bulgular ile benzerlik göstermektedir.

Üçüncü Alt Probleme İlişkin Bulgular ve Yorumlar

Çalışmanın üçüncü alt probleminde İki Aşamalı Kümeleme Analizi kulanılarak toplam 9870 öğrenciden elde edilen verilerin kaç kümeye ayrıldığı belirlenmiştir. Çalışmada ilk olarak farklı küme sayıları için elde edilen siluet değerlerinin nasıl bir değişim

gösterdiği incelenmiştir. Şekil 24'te k=2, k=3,…, k=7 için elde edilen siluet değerleri gösterilmiştir.



Şekil 24. Farklı küme sayıları için elde edilen siluet değerleri

Şekil 24 incelendiğinde, İki Aşamalı Kümeleme Analizi için küme sayısı sırasıyla 2, 3, 4, 5, 6 ve 7 olması durumunda siluet değerlerinin zayıf olduğu ve farklı algoritmaların denenmesi gerektiği görülmektedir. En yüksek siluet değeri küme sayısı 2 olduğunda elde edilirken küme sayısı 3 ve 4 olduğunda siluet değerlerinin çok fazla değişmediği ancak küme sayısı 5, 6 ve 7 olduğunda siluet değerlerinde azalma olduğu (<0,25) görülmektedir. Bu sonuca göre ideal küme sayısının 2, 3 veya 4 olabileceği belirlenmiştir. Elde edilen bu sonucun istatistiksel olarak desteklenmesi amacıyla analiz çıktısı olarak elde edilen otomatik kümeleme işlemine ilişkin sonuçlar Tablo 8'de gösterilmiştir.

Tablo 8

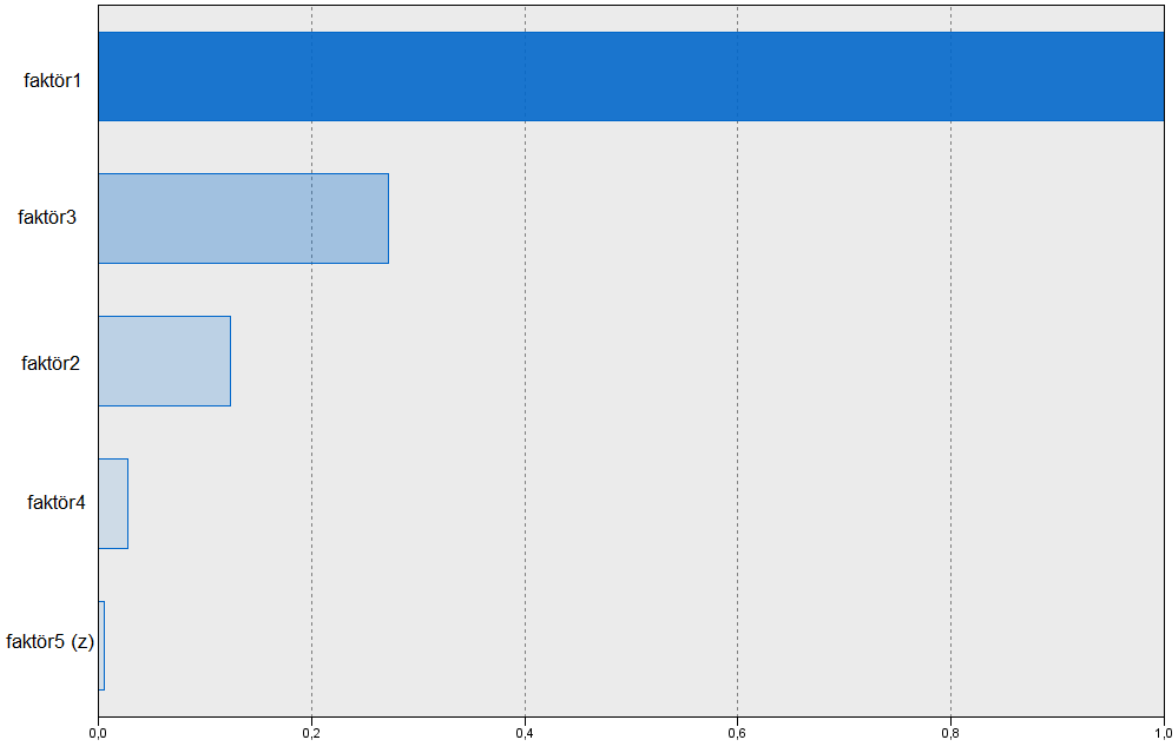
Otomatik Kümeleme Sonuçları

Küme Sayısı	Bayes Bilgi Kriteri (BBK)	BBK Değişimi	BBK Değişim Oranı	Uzaklık Ölçüsünün Oranı
1	34296.286			
2	**28868.831**	**-5427.455**	**1.000**	**2.090**
3	26319.408	-2549.422	.470	1.202
4	24213.022	-2106.387	.388	1.595
5	22927.142	-1285.879	.237	1.429
6	22054.826	-872.317	.161	1.255
7	21378.379	-676.446	.125	1.127
8	20788.732	-589.648	.109	1.074
9	20245.913	-542.819	.100	1.037

10	19725.474	-520.439	.096	1.230
11	19319.553	-405.921	.075	1.126
12	18969.200	-350.353	.065	1.119
13	18665.954	-303.246	.056	1.031
14	18374.478	-291.476	.054	1.044
15	18099.188	-275.289	.051	1.046

BBK Tablo 8 kapsamında yer alan ve ideal küme sayısını belirlemek için kullanılabilecek ölçütlerden bir tanesidir. Bu değerin en küçük olduğu noktada ideal küme sayısı görülebilmektedir. Ancak bu değer küme sayısına duyarlı olduğu için kullanılabilecek daha kararlı bir ölçüt BBK değişim oranıdır. Bu oranın diğerlerinden daha büyük olduğu nokta ideal küme sayısını göstermektedir. Buna göre ideal küme sayısının 2 olacağı düşünülmektedir. Elde edilen sonuçlar bir bütün olarak incelendiğinde küme sayısı 2 olması durumunda siluet değerinin de en yüksek değere ulaşması sebebiyle veri setinde yer alan birimlerin toplam 2 kümeye ayrıştığı sonucuna ulaşılmıştır.

Bunun yanında İki Aşamalı Kümeleme Analizinde değişkenlerin genel olarak kümelere ayrışmadaki önem düzeyine ilişkin çıktı elde edilmektedir. İki Aşamalı Kümeleme Analizinde girdi değişkenlerinin kümelerin oluşmasındaki önem düzeyine (predictor importance) ilişkin sonuçlar Şekil 25'te gösterilmiştir.



Şekil 25. Değişkenlere ilişkin önem düzeyi

Şekil 25 incelendiğinde kümelerin oluşmasında en etkili değişkenin öğretmen yönetimindeki fen bilgisi öğretimi (faktör1) olduğu belirlenmiştir. Sonrasında sırasıyla Fen Bilgisi derslerine ilişkin uyarlanabilir öğretimi (faktör3), fen bilgisi öğretmenlerinden alınan geri bildirim (faktör2), sorgulama temelli fen bilgisi öğretimi (faktör4) ve fene ilişkin olası başarı puanı ortalaması (faktör5) puanlarının kümelemede etkili diğer değişkenler olduğu belirlenmiştir.

Alanyazında İki Aşamalı Kümeleme Analizi kullanılarak yapılan çalışmalardan Kayri (2017) tarafından gerçekleştirilen araştırmada İki Aşamalı Kümeleme Analizinin avantajları ve dezavantajları incelenmiştir. SPSS programında gerçekleştirilen analiz sonucunda, İki Aşamalı Kümeleme Analizi ile hem sürekli hem de kategorik değişkenlerin birlikte girdi değişkenler olarak kullanılabileceği ve İki Aşamalı Kümeleme Analizi ile ideal küme sayısının belirlenmesinde BBK'nın ABK'ya kıyasla daha etkili olduğu sonucuna ulaşılmıştır. Yılmaz (2012) tarafından yapılan çalışmada üniversite öğrencilerinin internet kullanımını etkileyen amaçların öğrencilerin kümelere ayrışmasındaki rolü incelenmiştir. SPSS programında gerçekleştirilen analiz sonucunda, üniversite öğrencilerinin eğlence ve iletişim amacıyla internet kullanımı anlamında iki kümeye ayrıştığı sonucuna ulaşılmıştır. Aynı zamanda araştırma sonucuna göre, ilk kümenin çoğunluğu erkek olan ve internete büyük önem

veren öğrencilerden oluşurken, ikinci kümenin interneti daha az kullanan ve internete orta düzeyde önem veren öğrencilerden oluştuğu sonucuna ulaşılmıştır. Yıldırım ve Akın (2009) tarafından gerçekleştirilen çalışmada k-ortalamalar, İki Aşamalı Kümeleme Analizi ve CLARA Yöntemlerine ilişkin sonuçlar 3468 öğrenciden elde edilen veriler yardımıyla karşılaştırılmıştır. Araştırma sonucuna göre, her üç kümeleme yönteminin de büyük veri setleri üzerinde birbirine alternatif olarak kullanılabileceği belirlenmiştir. Tekin (2018) tarafından gerçekleştirilen çalışmada borsa İstanbul verileri Ward yöntemi, k-ortalamalar yöntemi ve İki Aşamalı Kümeleme Analizi kullanılarak SPSS programında analiz edilmiş ve sonuçlar karşılaştırılmıştır. Farklı yöntemlerle elde edilen küme sayısı benzerlik gösterirken, kümelerde yer alan hisse senetlerinin farklılık gösterdiği belirlenmiştir. İlgili araştırmalar bütüncül olarak değerlendirildiğinde, İki Aşamalı Kümeleme Analizinin SPSS programında gerçekleştirildiği ve İki Aşamalı Kümeleme Analizi, K-Ortalamalar Yöntemi ve ilgili araştırmalar kapsamında kullanılan diğer kümeleme yöntemi sonuçlarının ideal küme sayısı anlamında benzerlik göstermiştir. Bunun yanında, kümelerde yer alan birimler bakımından kullanılan yönteme göre farklılıklar olduğu belirlenmiştir. Ancak çalışmada K-Ortalamalar ve İki Aşamalı Kümeleme Analizi ile belirlenen ideal küme sayılarının farklı olması alanyazında yapılan çalışmalardan farklılık göstermektedir.

Bölüm 5
Sonuç ve Öneriler

Sonuçlar

Bu çalışmada PISA 2015 sınavına katılan ve sistematik örnekleme yöntemiyle belirlenen 34 OECD üyesi ülkeninin fen öğretimine ilişkin faktör puanları ve PISA fen okuryazarlığı ortalama puanlarına göre elde edilen kümeleme sonuçları Kohonen'in Öz Örgütlemeli Harita Yöntemi, K-Ortalamalar Yöntemi ve İki Aşamalı Kümeleme Analizi kapsamında incelenmiştir. Fen öğretimi boyutunun öğretmen yönetimindeki fen bilgisi öğretimi alt boyutu, fen bilgisi öğretmenlerinden alınan geri bildirim alt boyutu, fen bilgisi derslerine ilişkin uyarlanabilir öğretim alt boyutu ve sorgulama temelli fen bilgisi öğretimi alt boyutu için dört farklı faktör puanı ile öğrencilerin olası fen okuryazarlığı başarı puanı ortalaması olmak üzere toplam beş girdi değişkeni yardımıyla öğrencilerin farklı kümeleme yöntemlerine göre nasıl kümelendikleri belirlenmiştir.

Çalışmanın birinci alt probleminde Kohonen'in Öz Örgütlemeli Harita Yöntemiyle elde edilen sonuçlara göre ısı haritalarına bakarak kümelemede en etkili değişkenlerin sırasıyla fen bilgisi öğretmeninden algılanan geri bildirim, öğretmen yönetimindeki fen bilgisi öğretimi, olası fen başarı puan ortalaması, sorgulama temelli fen bilgisi öğretimi ve fen bilgisi dersilerine ilişkin uyarlanabilir öğretim olduğu belirlenmiştir. Başka bir ifadeyle öğrencileri kümelere ayırmada en etkili olan değişken algılanan geri bildirim iken en az etkili olan değişken uyarlanabilir öğretim olarak belirlenmiştir. Kohonen'in Öz Örgütlemeli Harita Yöntemiyle belirlenen ideal küme sayısının dört olduğu sonucuna ulaşılmıştır. Kümelerde yer alan öğrencilerin yaklaşık %60'ının 3.kümede, %21'inin 4.kümede, %15'inin 1.kümede ve %3'ünün ise 2.kümede oldukları belirlenmiştir. Elde edilen bu sonuca göre öğrencilerin büyük çoğunluğu üçüncü kümede yer alırken ikinci kümede yer alan öğrenci sayısının oldukça düşük olduğu tespit edilmiştir. Aynı zamanda birinci alt problem kapsamında üçüncü kümenin oluşmasında en etkili olan değişkenlerin sorgulama temelli fen bilgisi öğretimi ve fen başarısı olduğu belirlenmiştir. Elde edilen bu sonuç, Suarez (2011) tarafından yapılan çalışmanın sorgulama temelli fen bilgisi öğretimi başarısının materyal ve kaynak eksikliği ile ilişkili olduğu ve sorgulama temelli fen bilgisi öğretimi ile öğrenci başarısı arasında yüksek bir ilişkinin olduğu bulgusuyla benzerlik göstermektedir.

Çalışmanın ikinci alt probleminde K-Ortalamalar Yöntemi ile elde edilen küme sayıları ve kümelere ilişkin sonuçlar incelenmiştir. R programında K-Ortalamalar Yöntemiyle grup içi kareler toplamlarındaki değişimler, kümeler arası hata değerlerindeki değişimler, Gap istatistiğindeki değişimler ile ABK ve BBK'daki değişimler esas alınarak ideal küme sayısının dört olduğu belirlenmiştir. Elde edilen bu sonuç birinci alt problemde elde edilen sonuçlarla benzerlik gösterse de kümelere ilişkin dağılım grafiği incelendiğinde dört farklı renkle gösterilen kümelerin geçiş noktalarında renklerin bir miktar üst üste geldiği ve bu nedenle kümelerin tam olarak ayrışmadığı belirlenmiştir. Üst üste binme problemi olarak bu durumda bazı öğrencilerin hangi kümelerde yer aldıkları çalışma kapsamında ele alınan değişkenler yardımıyla net olarak belirlenememiştir. Kümelerde yer alan öğrenci sayılarının toplam öğrenci sayısına eşit olmaması yapılan kümeleme analizinde bir miktar hata olduğunu ve bu oranın yalaşık olarak %5'e eşit olduğunu göstermektedir. Birinci alt problemde elde edilen sonuca benzer olarak öğrencilerin en fazla üçüncü kümede ve sonrasında ikinci sırada dördüncü kümede yer aldıkları belirlenmiştir. Ancak K-Ortalamalar Yöntemiyle ikinci kümede yer alan öğrenci sayısının birinci kümede yer alan öğrenci sayısından daha fazla olduğu belirlenmiştir. Elde edilen bu sonuca göre en fazla öğrenci sayısına sahip ilk iki küme için kullanılan iki farklı kümeleme yöntemiyle benzer sonuçlar elde edilirken en az öğrenci sayısına sahip kümeler için kullanılan kümeleme yöntemine göre farklı sonuçlar elde edildiği görülmektedir. Alan yazın incelendiğinde, K-Ortalamalar Yöntemi sonucunda elde edilen ideal küme sayısı ile Kohonen'in Öz Örgütlemeli Harita Yöntemi ile elde edilen ideal küme sayısının benzerlik gösterdiği çalışmalara rastlanmaktadır (Baçāo, Lobo ve Painho, 2005; Kumar ve Vijayalakshimi, 2013; Lupaşcu ve Tegolo, 2011; Singh ve Kaur, 2017). Bununla birlikte alanyazın incelendiğinde kümeleme analizinde üst üste binme durumunun bir problem olarak nitelendirildiği çalışmalar göze çarpmaktadır (Zhang, Wang ve Zhang, 2007; Cleuziou, 2008; Goldberg, Hayvanovych ve Magdon-Ismail, 2010). İkinci alt problem kapsamında birinci kümenin oluşmasında en etkili olan değişkenlerin öğretmen yönetimindeki fen bilgisi öğretimi ve fen başarısı olduğu belirlenmiştir. Elde edilen bu sonuç Da Costa ve Araujo (2018) tarafından yapılan çalışmada, derslerin sıklıkla öğretmen yönetimindeki fen bilgisi öğretimi yöntemiyle

işlenmesinin fende daha yüksek başarı elde etmekle ilişkili olduğu bulgusuyla benzerlik göstermektedir.

Çalışmanın üçüncü alt probleminde İki Aşamalı Kümeleme Analizi ile elde edilen küme sayıları ve kümelere ilişkin sonuçlar incelenmiştir. Birinci ve ikinci alt problemde elde edilen analiz sonuçlarında daha sınırlı sonuçların yer aldığı İki Aşamalı Kümeleme Analizinde siluet değerleri ve BBK değerleri üzerinden ideal küme sayısı belirlenmeye çalışılmıştır. Ele alınan her iki ölçüte göre de ideal küme sayısının değişmediği ve ikiye eşit olduğu belirlenmiştir. İki Aşamalı Kümeleme Analizinde genel anlamda kümelerin ayrışmasında fen öğretimine ilişkin alt boyutların faktör puanları etkili olurken olası fen başarı puan ortalamasının etkili olmadığı belirlenmiştir. Elde edilen bu sonuç birinci alt problemde Kohonen'in Öz Örgütlemeli Harita Yönteminde kümelere ayırmada en etkili üçüncü değişken olarak belirlenen olası fen başarı puanının kümelemede en az etkiye sahip değişken olarak belirlenmesi bakımından farklılık göstermektedir. İdeal küme sayının iki olarak belirlenmesi bakımından elde edilen bu sonuç çalışmaının birinci ve ikinci alt problemleri ile farklılık göstermektedir. Alanyazın incelendiğinde, İki Aşamalı Kümeleme Analizi sonucunda elde edilen ideal küme sayısı ile K-Ortalamalar Yöntemi ile elde edilen ideal küme sayısının benzerlik göstermediği çalışmalara rastlanmaktadır (Shih, Jheng ve Lai, 2010; Ceylan, 2013).

Çalışma kapsamında elde edilen küme profilleri bireyler ve değişkenler arasında doğrudan bir neden-sonuç ilişkisi göstermemektedir. Elde edilen sonuçlar kümelere ayrılmada değişkenlerin kendi içerisinde karşılaştırmasını yapmaktadır. Bu nedenle çalışma kapsamında kullanılan kümeleme yöntemine göre bireylerin yer aldıkları kümelerde farklılıklar olabileceği belirlenmiştir. Bu sonucun ortaya çıkmasına veri setinin oldukça hetorejen olmasının neden olduğu düşünülmektedir.

Öneriler

Araştırmadan elde edilen bulgulara dayalı olarak aşağıda uygulayıcılar ve araştırmacılar için geliştirilen önerilerde bulunulmuştur.

Uygulayıcılar için öneriler. Alanyazında farklı programlar aracılığıyla gerçekleştirilen kümeleme analizlerinde her bir ülke için o ülkeye ilişkin ortalama puanlar üzerinden analizler gerçekleştirildiği göze çarpmaktadır. Bu çalışma kapsamında ortalama puanlar yerine o ülkeleri temsil eden öğrencilerin tamamı

kümeleme analizlerine dahil edilmiştir. Bu bakımdan ülkelere ilişkin ortalama puanlar yerine bireylerin her birinden elde edilen puanların analize dahil edilmesiyle daha duyarlı ve dolayısıyla daha güvenilir sonuçlar elde edileceği düşünülmektedir. Benzer çalışmalarda ülkelere ilişkin ortalama puanlar yerine bireylere ilişkin puanların kullanılması önerilebilir.

Araştırma kapsamında kullanılan yöntemler sonucunda elde edilen küme profillerinin daha belirgin olabilmesi amacıyla ileride yapılacak olan çalışmalarda daha fazla değişken ya da daha farklı değişkenlerin kullanılması önerilebilir.

Veri madenciliği yöntemlerinin kullanıldığı bu çalışmada 34 ülke yerine 9870 öğrenciden elde edilen veri ile analizler gerçekleştirilmiştir. Bu durum küçük veri yerine büyük veri ile çalışıldığı anlamına gelmektedir. Bu nedenle büyük veri ile çalışma yapmak isteyen araştırmacıların kümeleme analizlerinde veri madenciliği yöntemlerini kullanmaları önerilebilir.

Analize dahil edilen değişken sayısı bakımından R programı ile yapılan analizlerin daha hassas ve daha kesin sonuçlar üreteceği düşünüldüğünden analizlerde tek bir satırda ilgili birimi analize dahil etmek yerine en azından R programı yardımıyla o birime ilişkin birden fazla gözlem sonucunun analize dahil edilerek çıktıların da rapor edilmesi önerilebilir. Bu sayede kümeleme anlamında daha gerçekçi sonuçların elde edileceği düşünülmektedir.

Çalışma sonucunda farklı kümeleme yöntemlerinden elde edilen sonuçlar incelendiğinde R programından elde edilen çıktıların diğerlerine kıyasla daha fazla ve daha zengin olduğu belirlenmiştir. Özellikle siluet değerleri ile kalibrasyon grafiklerinin ideal küme sayısını belirlemede oldukça anlaşılır olması bakımından R programından elde edilen çıktıların daha fazla bilgi verdiği sonucuna ulaşılmıştır. Bu nedenle kümeleme analizinde elde edilen sonuçları rapor etmek için birden fazla değerlendirme ölçütünün kullanılması amacıyla araştırmacıların R programını kullanarak analizlerini gerçekleştirmeleri önerilebilir.

R kapsamında Kohonen'in Öz Örgütlemeli Harita Yönteminde çıktı olarak her bir ülkede yer alan bireylerin kümelere ilişkin dağılım grafiği elde edilebilmektedir. Bu grafikler yardımıyla her bir kümede yer alan öğrenci sayıları ve oranları ülke bazında kolaylıkla görülebilmektedir. Diğer programlarda analiz sonucunda bu ve benzeri

grafikler elde edilemediğinden araştırmacıların Kohonen'in Öz Örgütlemeli Harita Yöntemini R programıyla gerçekleştirmeleri önerilebilir.

Bu çalışmanın, çok büyük veri setlerine kümeleme analizi uygulayacak olan araştırmacıların Kohonen'in Öz Örgütlemeli Harita Yöntemini bir alternatif olarak göz önünde bulundurmaları anlamında araştırmacıları cesaretlendireceği düşünülmektedir.

İstatistiklere ve görselliğe dayalı daha zengin çıktılar elde edilmesi, açık kaynak kodlu ve ücretsiz olması sebebiyle araştırma kapsamında Kohonen'in Öz Örgütlemeli Harita Yöntemi için SPSS Modeler yerine R programı kullanılmıştır. Araştırmacıların Kohonen'in Öz Örgütlemeli Harita Kümeleme analizini R programıyla gerçekleştirmeleri önerilebilir..

R kapsamında K-Ortalamalar Yönteminde çıktı olarak kümeler arası hata değerleri, ABK ve BBK değişimi ve Gap istatistiğine ilişkin görseller elde edilebilmektedir. Bu görseller yardımıyla ideal küme sayısına ilişkin elde edilen deliller zenginleştirilebilmektedir. Diğer programlarda analiz sonucunda bu ve benzeri grafikler elde edilemediğinden araştırmacıların K-Ortalamalar Yöntemini R programıyla gerçekleştirmeleri önerilebilir.

Kohonen'in Öz Örgütlemeli Harita Yönteminde çıktı olarak dört kümeye ilişkin profil belirlenmiş, olası fen başarısı en yüksek olan öğrencilerin sorgulama temelli fen bilgisi öğretimini benimsedikleri belirlenmiştir. Ayrıca çalışma kapsamında K-Ortalamalar Yöntemi kapsamında da çıktı olarak dört kümeye ilişkin profil belirlenmiş, olası fen başarısı en yüksek olan öğrencilerin öğretmen yönetimindeki fen bilgisi öğretimini benimsedikleri belirlenmiştir. Gelecekte, bu çelişkili bulguların sebebini ortaya koymayı amaçlayan çalışmaların yapılması önerilebilir.

K-Ortalamalar Yöntemi kapsamında fen başarısı en yüksek olan öğrencilerin yer aldığı birinci kümeden sonra fen başarısı en yüksek olan öğrencilerin yer aldığı ikinci kümenin üçüncü küme olduğu ve üçüncü kümede yer alan öğrencilerin fen bilgisi öğretmenlerinden alınan geri bildirimi benimsedikleri belirlenmiştir. Bu sonuçtan yola çıkarak fen başarısını yükseltmek isteyen okulların fen bilgisi öğretmenlerinden alınan geri bildirime önem vermeleri önerilebilir.

Kohonen'in Öz Örgütlemeli Harita Yöntemi kapsamında fen başarısı en yüksek olan öğrencilerin yer aldığı üçüncü kümeden sonra fen başarısı en yüksek olan

öğrencilerin yer aldığı ikinci kümenin birinci küme olduğu ve birinci kümede yer alan öğrencilerin öğretmen yönetimindeki fen bilgisi öğretimini benimsedikleri belirlenmiştir. Bu sonuçtan yola çıkarak fen başarısını yükseltmek isteyen okullarda öğretmen yönetimindeki fen bilgisi öğretimine ağırlık verilmesi önerilebilir.

Araştırmacılar için öneriler. Bu çalışmanın, özellikle de sosyal bilimlerde çok fazla kullanılmayan Kohonen'in Öz Örgütlemeli Harita Yönteminin sosyal bilimler alanında farklı türde veri setleri söz konusu olduğunda kullanılması önerilebilir.

Alanyazın incelendiğinde K-Ortalamalar Yönteminin SPSS, Minitab, STATISTICA vb. programlar kapsamında gerçekleştirildiği göze çarpmaktadır. Bu çalışmanın, K-Ortalamalar Yöntemini kullanacak olan araştırmacıların bu yöntemi R programı ile gerçekleştirmeleri anlamında onları cesaretlendireceği düşünülmektedir.

Çalışma kapsamında Kohonen'in Öz Örgütlemeli Harita Yöntemi ve K-Ortalamalar Yöntemi sonucunda elde edilen ideal küme sayısının eşit olduğu belirlenmiştir. Benzer çalışmalarda kümeleme analizini kullanacak olan araştırmacıların ideal küme sayısını belirlemede Kohonen'in Öz Örgütlemeli Harita Yöntemi veya K-Ortalamalar Yönteminden birini kullanmaları önerilebilir.

Çalışma kapsamında girdi değişkenlerinin kümelerin oluşmasında ne düzeyde etkili olduklarına ilişkin alanyazında kullanılan diğer kümeleme yöntemlerinden farklı olarak Kohonen'in Öz Örgütlemeli Harita Yöntemi kapsamında ısı haritaları iki Aşamalı Kümeleme Analizi kapsamında ise değişkenlere ilişkin önem düzeyi grafiği elde edilmiştir. Bu çıktılar kullanılarak değişkenler kendi içlerinde karşılaştırılabilmektedir. Girdi değişkenlerinin kümeleme analizinde ne düzeyde etkili olduğunu belirlemek isteyen araştırmacıların Kohonen'in Öz Örgütlemeli Harita Yöntemi ve iki Aşamalı Kümeleme Analizini kullanmaları önerilebilir.

Çalışma sürecinde, kümeleme analizlerinin her ne kadar bulanık mantık kümeleme analizi yöntemiyle de gerçekleştirilmesi amaçlansa da yaklaşık 10.000 kişilik veri setinde bu analizin gerçekleştirilemediği belirlenmiştir. R programında bulanık mantıkla kümeleme analizlerinin gerçekleştirilebilmesi için binlerce çekirdeği, işlemci gücü 1000'in üzerinde ve çok fazla elektrik harcayan depo büyüklüğünde bilgisayarlarla gerçekleştirilebileceği düşünülmektedir. Küçük veri setleri ile kümeleme analizi gerçekleştirecek olan araştırmacıların R programında bulanık mantıkla kümeleme analizini gerçekleştirmeye çalışmaları önerilebilir.

Çalışmada farklı kümeleme yöntemleriyle elde edilen sonuçların farklılıklar gösterdiği belirlenmiştir. Elde edilen bu bulgulara dayalı olarak kümeleme yöntemlerine ilişkin çalışmalarda araştırmacıların veri madenciliğinde tek bir yöntemle çalışmak yerine en az üç farklı kümeleme yöntemi kullanarak elde edilen sonuçların geçerliğine ilişkin delil sunmaları önerilebilir.

Çalışmada PISA verileri kullanarak analiz yapıldığından ileride kümeleme analizleriyle elde edilen sonuçların incelenmesinin veya karşılaştırılmasının amaçlandığı çalışmalarda TIMMS, PIRRLS ve benzeri geniş ölçekli sınavlardan elde edilen veri setlerinin kullanılması önerilebilir. Bu sayede diğer geniş ölçekli sınavlara göre küme sayılarının nasıl bir değişkenlik göstereceği belirlenerek elde edilen sonuçların geçerliğinin arttırılacağı düşünülmektedir.

Çalışmada R programı kullanarak Kohonen'in Öz Örgütlemeli Harita yöntemi ve K-Ortalamalar Yöntemi kapsamında küme profilleri elde edilmiştir. . Benzer çalışmalarda kümeleme analizini kullanacak olan araştırmacıların küme profili belirlemede Kohonen'in Öz Örgütlemeli Harita Yöntemi veya K-Ortalamalar Yöntemini kullanmaları önerilebilir.

Kaynaklar

Acar, T. (2012). Türkiye'nin PISA 2009 sonuçlarına göre OECD'ye üye ve aday ülkeler arasındaki yeri. *Kuram ve Uygulamada Eğitim Bilimleri, 12*(4), 2561-2572.

Adams, R. J., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-Scale Assessments in Education,* 1, 5.

Akın, H. B., & Eren, Ö. (2012). OECD ülkelerinin eğitim göstergelerinin kümeleme analizi ve çok boyutlu ölçekleme analizi ile karşılaştırmalı analizi. *Marmara Üniversitesi E-Dergi Sistemi, 10*(37), 175-181.

Aksu, G. Güzeller, C. O., & Eser, M T. (2017). *PISA 2012 sonuçlarının duyuşsal özelliklere göre kümeleme çalışması. Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 32*(4), 838-862.

Ali, M. M. (2013). *Role of data mining in education sector. International Journal of Computer Science and Computing, 2*(4), 374-383.

ALMazroui, Y. A. (2013). A survey of data mining in the context of e-learning. *International Journal of Information Technology & Computer Science (IJITCS)*, *7*(3), 8-10.

Amari, S. (1980). Topographic organization of nerve fields. *Bull. Math. Biol*, 42, 339-364

Anderberg, M. (1973). *Cluster analysis for applications*. Academic Press, New York.

Andrews, N.O., & Fox, E.A. (2007). Recent Developments İn Document Clustering. Technical Report. *Department of Computer Science*, Virginia Tech, Blacksburg, VA.

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, *60*(3), 383-398.

Arı, E. S., Özköse, H. D. A., & Calp, M. H. (2016). İstanbul borsası'nda işlem gören firmaların finansal performanslarının kümeleme analizi ile değerlendirilmesi. *Bilişim Teknolojileri Dergisi*, *9*(1), 33-39.

Atalay, M., & Öztürk, Ş. (2016). Türkiye'deki illerin göç ve işsizlik istatistiklerine göre kümelenmesi. *2nd International Congres on Applied Sciences: Migration, Poverty and Employment*.

Atiya, A. F. (1990). An unsupervised learning technique for artificial neural networks. *Neural Networks*, 3, 707-711.

Baçăo, F., Lobo, V.S., & Painho, M. (2005). Self-organizing maps as substitutes for k-means clustering. *International Conference on Computational Science*.

Baker, R. S. J. D. (2010). Data mining for education. In McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopedia of Education* (3rd edition), 7, 112-118. Oxford, UK: Elsevier.

Baker, R., Growda, S.M., & Corbett, A.T. (2011). Automatically detecting a student's preparation for future learning: Help use is key. *4th International Conference on Educational Data Mining*.

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

Baepler P., & Murdoch C.J. (2010). Academic analytics and data mining in higher education. *Teach. Learn*, *4*(2), 17.

Barahate, S. R. (2012). Educational data mining as a trend of data miningıneducational system. Paper presented at the International Conference & Workshop on Recent Trends in Technology.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, 1, 295-311.

Becker, S., & Plumbley, M. (1996). Unsupervised neural network learning procedures for feature extraction and classification. *International Journal of Applied Intelligence*, 6, 185-203.

Berkhin, P. (2002). Survey of clustering data mining techniques. Technical Report, *Accrue Software, San Jose.*

Bhise, R. B., Thorat, S. S., & Supekar, A. K. (2013). Importance of data mining in higher education system. *IOSR Journel of Humanities And Social Science*, *6*(6), 18-21.

Bırtıl, F. S. (2012). *Kız meslek lisesi öğrencilerinin akademik başarısızlık nedenlerinin veri madenciliği tekniği ile analizi* (Doktora Tezi). Afyon Kocatepe Üniversitesi Fen Bilimleri Enstitüsü, Afyon.

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. Washington, D.C.

Bilen, O., Hotaman, D., Aşkın, O. E., & Büyüklü, A. H. (2014). LYS başarılarına göre okul performanslarının eğitsel veri madenciliği teknikleriyle incelenmesi: 2011 istanbul örneği. *Eğitim ve Bilim*, *39*(172), 78-94.

Bishop, C. M. (1995) *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Bishop, C. M., Svensen, M., & Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Comput.*, *10*(1), 215–234.

Bodner, T. E. (2008). What improves with increased missing data ımputations? *Structural Equation Modeling 15*(4), 651–75.

Burnham, K. P., & D. R. Anderson. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.

Cameron, R. C., & D. L. Miller (2015): A practitioner's guide to cluster-robust ınference. *Journal of Human Resources*, *50*(2), 317-372.

Ceylan, H. H. (2013). Perakende Sektöründe Konjoint Ve Kümeleme Analizi İle Fayda Temelli Pazar Bölümlendirme. *Yöntem ve Ekonomi*, *20*(1).

Ceylan, Z., Gürsev, S., & Bulkan, S. (2017). İki aşamalı kümeleme analizi ile bireysel emeklilik sektöründe müşteri profilinin değerlendirilmesi. *Bilişim Teknolojileri Dergisi*, *10*(4), 475-485.

Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. *19th International Conference on Pattern Recognition*, Tampa, FL, 1-4.

Çakmak, Z. (1999). Kümeleme analizinde geçerlilik problemi ve kümeleme sonuçlarının değerlendirilmesi. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, *1*(3), 187-205.

Çırak, G., & Çokluk, Ö. (2013). Yükseköğretimde öğrenci başarılarının sınıflandırılmasında yapay sinir ağları ve lojistik regresyon yöntemlerinin kullanılması. *Akdeniz İnsani Bilimler Dergisi*, *3*(2), 71-79.

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning*. USA: John Wiley&Sons.

Da Costa, P. D., & Araujo, L. (2018). *Quality of teaching and learning in science.* Luxembourg: Publications Office of the European Union

Delen, İ., & Bellibaş, M. S. (2015). Formative assessment, teacher-directed instruction and teacher support in Turkey: Evidence from PISA 2012. Mevlana International Journal of Education, *5*(1), 88–102.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1-11.

Dunham, M. H. (2003). *Data mining ıntroductory and advanced topics*. USA: Prentice Hall.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. *J. Cybern*. 3, 32–57.

Ersöz, F. (2009). OECD'ye üye ülkelerin seçilmiş sağlık göstergelerinin kümeleme ve ayırma analizi ile karşılaştırılması. *Türkiye Klinikleri Journal of Medical Sciences*, *29*(6), 1650-1659.

Fiedler, J. A., Mcdonald J. J. (1993). Market figmentation: Clustering on factor scores versus individual variables. *AMA Advanced Research Techniques Forum*.

Fovell, R. G., & Fovell, M. Y. C. (1993). Climate zones of the conterminous United States defined using cluster analysis. *Journal of Climate*, 6, 2103- 2135.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, *82*(3), 300-329.

Garson, D. G. (2014). *Cluser analysis (Statistical Associates Blue Book Series)*. Ashebero: Statistical Associates Publishing. Kindle edition.

Genolini, C., & Falissard, B. (2010). KmL: K-means for longitudinal data. *Computational Statistics*, 25, 317–332.

Goldberg, M. K., Hayvanovych, M., & Magdon-Ismail, M. (2010). Measuring similarity between sets of overlapping clusters. *2010 IEEE Second International Conference on Social Computing*. Minneapolis, MN, 303-308.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549-576.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6(4), 430.

Gurney, K. (1997). *An Introduction to Neural Network*. London: UCL Press Limited.

Gürsoy, U. T. Ş. (2009). *Customer* churn analysis in telecommunication sector. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 39(1), 35-49.

Hair Jr., J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis*. New Jersey: Prentice Hall.

Hämäläinen, W., Kumpulainen, V., & Mozgovoy, M. (2015). *Artificial intelligence applications in distance education.* Pennslyvania: IGI GLOBAL.

Hartigan, J. A., & M. A. Wong (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 28(1), 100–108.

Hastie T., Tibshirani R., & Friedman J. (2009) *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses related to achievement*. New York, NY: Routledge.

Haykin, S. (2009). *Neural networks and learning machines*. New Jersey: Prentice Hall.

Hore, P., Hall, L.O., & Goldgof, D.B. (2009). A scalable framework for cluster ensembles. *Pattern Recogn.*, 42(5), 676–688.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417-441, 498-520.

Inmon, W. H., & Hackathorn, R. D. (1994). *Using the data warehouse*. New York: John Wiley&Sons.

İnce, H., İmamoğlu, S. Z., & Keskin, H. (2013). Öz-düzenlemeli harita ağları ile k-ortalama kümeleme analizinin karşılaştırılması: Tüketici profilleme örneği. *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 28(4), 723-731.

Jain, A., & Dubes, R. (1988). *Algorithms for clustering data. New Jersey*: Prentice Hall

Jain, A. K. ve Dubes, R. C. (1998). *Algorithms for clustering data*. New Jersey: Prentice Hall.

Kalikov, A. (2006). *Veri madenciliği ve bir e-ticaret uygulaması (*Yüksek Lisans Tezi). Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Kanakana, G. M., & Olanrewaju, A. O. (2011). Predicting student performance in engineering education using an artificial neural network at tshwane university of technology. *International Conference on Industrial Engineering, Systems Engineering and Engineering Management for Sustainable Global Development*.

Kaplan, D., & Su, D. (2016). On matrix sampling and imputation of context questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics*, 41, 51–80.

Kaski, S., Kangas, J., & Kohonen, T. (1998). Bibliography of self-organizing map (SOM) papers: 1981-1997. *Neural Computing Surveys*, 1: 102-350.

Kaufman, L., & Rousseeuw, P. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Kayri, M. (2007). Two-step cluster analysis in researches: A case study. *Eurasian Journal of Educational Research*, 28, 89-99.

Kayri, M. (2008). Elektronik portfolyo değerlendirmeleri için veri madenciliği yaklaşımı. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, 5(1), 98- 110.

Kiang, M. Y. (2001). Extending the kohonen self-organizing map networks for clustering analysis. *Computational Statistics&Data Analysis*, 38(2001), 161-180.

Kjærnsli, M., & Lie, S. (2011). Students' preference for science careers: International comparisons based on PISA 2006. International Journal of Science Education, 33(1), 121–144.

Kleinberg, J. (2002). An impossibility theorem for clustering. In: Proceedings of the 16th Annual Conference on Neural Information Processing Systems, pp. 9–14.

Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., & Vısa, A. (2004). Combining data and text mining techniques for analysing financial reports. *Intelligent Systems In Accounting, Finance and Management*, 12,(1).

Klösgen, W., & Zytkow, J. (2002). *Handbook of data mining and knowledge discovery*. New York: Oxford University Press.

Koç, M., Karabatak, M. (2012). Sosyal ağların öğrenciler üzerindeki etkisinin veri madenciliği kullanılarak incelenmesi. *E-Journal of New World Sciences Academy*, 7(1), 155-164.

Kohonen T. (1984). *Self-organization and associative memory*. Berlin: Springer.

Kohonen, T. (2001). *Self-organizing maps*. Berlin: Springer-Verlagn.

Kohonen, T. (2014). *MATLAB Implementations and applications of the self-organizing map*. Helsinki: Unigrafia Oy.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review* 26(3), 159-190.

Kruskal, J. B. & Wish, M. (1978). *Multidimensional scaling*. Newbury Park: Sage Publications.

Kumar, S. A., & Vijayalakshmi, M. N. (2013). Discerning learner's erudition using data mining techniques. *International Journal on Intergrating Technology in Education*, 2(1), 9-14.

Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and k-means algorithm for market segmentation. *Computers&Operations Research*, 29(11),1475-1493

Langdon, D., Mckittrick, G., Beede, D., Khan, B., & Doms, M. (2011). STEM: Good jobs now and for the future, U.S. *Department of Commerce Economics and Statistics Administration*, 3(11), 2.

Larose, D.T., Larose, C. D. (2012). *Discovering knowledge in data: An ıntroduction to data mining*. New Jersey: John Wiley & Sons, Second Edition.

Linnakylä, P., & Malin, A. (2008). Finnish students' school engagement profiles in the light of PISA 2003. *Scandinavian Journal of Educational Research*, *52*(6), 583-602.

Little. R., & Rubin. D. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lupaşcu C. A., & Tegolo D. (2011). *Automatic unsupervised segmentation of retinal vessels using self-organizing maps and k-means clustering*. Heidelberg: Springer.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. California: University of California Press.

Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik,* 14, 85–100.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.

Merceron, A., & Yacef, K. (2007). Revisiting ınterestingness of strong symmetric association rules in educational data. *Workshop on Applying Data Mining in e-Learning*.

Milligan, G. (1996). Clustering validation: Results and implications for applied analyses. Singapore: World Scientific.

Minner, D., Levy, A. J., Century, J. (2010). Inquiry based science instruction-What is it and does it matter? Results from a research synthesis yerars 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474-496.

Mirkin, B. (1996). *Mathematical Classification and Clustering*. Dordrecht: Kluwer Academic Press.

Monette, D. R., Sullivan, T., De Jong, C. R. (1990). *Applied Social Research*. New York: Harcourt Broce Jovanovich, Inc.

Mulaik, A. (2009). *Foundations of Factor Analysis*. London: Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences Series.

Navarro, A. A. M., & Ger, P. M. (2018). Comparison of clustering algorithms for learning analytics with educational datasets. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5, 9-16.

Norusis, M. (2010). *PASW 18 statistical procedures companion*. London: Pearson

Önen, E. (2018). Öğrenci, öğretmen ve öğretimsel nitelikler açısından TIMSS-2015'e dayalı olarak öğrencilerin sınıflandırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 9(1), 64-84.

Oğuzlar, A. (2005). A new approach to clustering analysis: Self-organizing maps. *Ataturk University Journal of Economics and Administrative Sciences*, 19 (2).

Özbay, Ö. (2015). Veri madenciliği kavramı ve eğitimde veri madenciliği uygulamaları. *Uluslararası Eğitim Bilimleri Dergisi*, 2(5), 262-272.

Özşahin, M., & Yüreğir, O. H. (2012). Otomotiv sektörünün kendini örgütleyen haritalar ile finansal analizi. *Çanakkale Üniversitesi Fen ve Mühendislik Bilimleri Dergisi*, 28(2), 155-164.

Öztemel, E. (2006). *Yapay sinir ağları*. İstanbul: Papatya Yayıncılık.

Oğuzlar, A. (2004). *Veri Madenciliğine Giriş*. Bursa: Ekin Kitabevi.

Oja, M., Kaski, S. & Kohonen, T. (2003). Bibliography of self-organizing map (SOM) papers: 1998-2001. *Neural Computing Surveys*, 3, 1-156.

Özçalıcı, M. (2017). Özdüzenleyici haritalar yardımıyla piyasa bölümlendirmesi: türkiye ikinci el otomobil piyasası örneği. *Eskişehir Osmangazi Üniversitesi İktisadi İdari Bilimler Fakültesi Dergisi*, 12(2), 23–36.

Pan, W., Shen, X., & Liu, B. (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *Journal of Machine Learning Research*, *14*, 1865-1889.

Penn, B. S. (2005). Using self-organizing maps to visualize highdimensional data. *Computers & Geosciences*, 31(5), 531-544.

Punj, G., & W. Stewart (1983). An ınteraction framework of consumer decision making. *Journal of Consumer Research*, 10, 181-196.

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: a didactic. *Frontiers in Psychology*, 9, 2231.

Ramaswami, M. & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. *International Journal of Computer Science Issues*, 7(1), 10-18.

Reilly, C., Wang, C., & Rutherford, M., (2005). A rapid method fort he comparison of clusteranalysis. *StatisticaSinica*, 15, 19-33.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. England: Cambridge University Press.

Ritter, H. (1991). Asymptotic level density for a class of vector quantization processes. *IEEE Trans. Neural Networks*, 2, 173-175.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33, 125-146.

Romero, C., Ventura, S., Espejo, P. G., & Hervas, C. (2008). Data mining algorithms to classify students. *First International Conference on Educational Data Mining*.

Romero, C., & Ventura, S. (2013). Data mining in education. Wiley ınter disciplinary reviews. *Data Mining and Knowledge Discovery*, 3(1), 12–27.

Rundle-Thiele, S., Kubacki, K., Tkaczynski, A., & Parkinson, J. (2015). Using two-step cluster analysis to ıdentify homogeneous physical activity groups. *Marketing Intelligence & Planning*, 33(4), 522-537.

Sarıman, G. (2011). Veri madenciliğinde kümeleme teknikleri üzerine bir çalışma: K means ve k-medoids kümeleme algoritmalarının karşılaştırılması. *Süleyman Demirel University Journal of Natural &Applied Sciences*, 15-3, 192-202.

Sarkar, D. (2008). Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5

Schroeder, C., Scott, T., Tolson, H., Yuang, Y. T., & Lee, H. Y. (2007). A meta analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching*, *44*(10), 1436-1460.

Sharma, R., & Singh, H. (2013). Data mining in education sector. *International Journal of Electronics & Data Communication*, 2(1), 4-8.

Shih, M. Y., Jheng, J. W., & Lai, L. F. (2010). A two-step method for clustering mixed categroical and numeric data. *Tamkang Journal of Science and Engineering*, 13(1), 11-19.

Siemens, G., & Baker, R. S. J. D. (2012). Learning analytics and educational data mining: *2nd International Conference on Learning Analytics and Knowledge*.

Singh, D., & Kaur, A. (2017). Comparative analysis of k-means and kohonen-som data mining algorithms based on student behaviors in sharing information on facebook. *International Journal Of Engineering And Computer Science*, 6(4), 20990-20993.

Suarez, M. L. (2011). *The relationship between inquiry-based science instruction and student achievement* (Doctoral dissertation). University of Southern Mississippi, USA.

Şentürk, A. (2006). *Veri madenciliği kavram ve teknikler*. Bursa: Ekin Kitabevi.

Şekerkaya, A., & Cengiz, E. (2010). Kadın tüketicilerin alışveriş merkezi tercihlerinin belirlenmesi ve bir pilot araştırma. *Öneri*, *9*(34), 41–55.

Tair, M., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. *International Journal of Information and Communication Technology Research*, *2*(2), 140-146.

Taşdemir, M. (2012). *Öğrenci başarısına etki eden faktörlerin regresyon analizi ile tespiti (Yüksek Lisans Tezi)*. Dicle Üniversitesi, Diyarbakır.

Taşkın, Ç., & Emel, G. G. (2010). Veri madenciliğinde kümeleme yaklaşımları ve kohonen ağları ile perakendecilik sektöründe bir uygulama. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *15*(3), 395- 409.

Tatlıdil, H. (1992). *Çok değişkenli istatistiksel analiz*. Ankara: Hacettepe Üniversitesi Yayınları.

Taylan, O., & Karagözoglu, B. (2009). An adaptive neuro-fuzzy model for prediction of student's academic performance. *Computers & Industrial Engineering*, *57*(3), 732-741.

Tekin, B. (2018). Ward, k-ortalamalar ve iki adımlı kümeleme analizi yöntemleri ile finansal göstergeler temelinde hisse senedi tercihi. *Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *21*(40).

Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*. Washington: American Psychological Association.

Tiwari, M., Singh, R., & Vimal, N. (2013). An empirical study of application of dm techniques for predicting student performance. *International Journal of Computer Science and Mobile Computing*, *2*(2), 53-57.

Tkaczynski, A., Rundle-Thiele, S., & Beaumont, N. (2010). Destination segmentation: A recommended two-step approach. *Journal of Travel Research*, 49, 139–152.

Tsuda, K., & Kudo, T. (2006). Clustering graphs by weighted substructure mining. in 23rd International Conference on Machine Learning.

Usami, S. (2014). Constrained k-means on cluster proportion and distances among clusters for longitudinal data analysis. *Japanese Psychological Research*, *56*(4), 361-372.

Vellido, A., Castro, F., & Nebot, A. (2010). *Handbook of educational data mining*. London: Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Vedder-Weiss, D., & Fortus, D. (2012). Students' declining motivation to learn science: A follow up study. *Journal of Research in Science Teaching*, *49*(9), 1057–1095.

Wehrens R., & Buydens L. M. C. (2007). Self- and super-organizing maps in r: the kohonen package. *Journal of Statistical Software*, *21*(5), 1-19.

Wehrens, R., & Kruisselbrink, J. (2018). Flexible self-organizing maps in kohonen 3.0. *Journal of Statistical Software*, *87*(7), 1-18.

Wells, R. D. (1999). Factor scores and factor structure. *In Advances İn Social Science Methodology*, 5,123-138.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann.

Wu, J. (2012). *Advances in k-mean clustering: A data mining thinking*. Hedilberg: Springer Science & Business Media.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Trans. Neural Networks*. *16*(3), 645–678.

Yıldırım, İ., & Akın, Y. (2009). Veri madenciliğinde kümeleme teknikleri ile ortaöğretim öğrencilerinde şiddet eğiliminin tespiti üzerine bir analiz. *Finans Politik & Ekonomik Yorumlar*, *46*(530).

Yılmaz, M. B. (2012). Profiles of university students according to ınternet usage with the aim of entertainment and communication and their affinity to ınternet. *International Online Journal Education Science*, *4*(1), 225-242.

Zeileis, A., & Croissant, Y. (2010). Extended model formulas in R: Multiple parts and multiple responses. *Journal of Statistical Software*, *34*(1), 1-13.

Zhang, T., Ramakrishnon, R., & Livny, M. (1996). BIRCH: Method for very large databases. *International Conference on Management of Data*.

Zhang, S., Wang, R., & Zhang, X. (2007). Identification of overlapping community structure in complex networks using fuzzy -means clustering. *Physica A: Statistical Mechanics and its Applications*, *374*(1), 483-490.

EK-A: Etik Komisyonu Onay Bildirimi

Form: 40

Tez Çalışması Etik Komisyon İzin Muafiyeti Formu

20 / 01 / 2017

Hacettepe Üniversitesi
Eğitim Bilimleri Enstitüsü
Eğitim Bilimleri Anabilim Dalı Başkanlığı'na

Tez Başlığı / Konusu:	ULUSLARARASI ÖĞRENCİ DEĞERLENDİRME PROGRAMI 2015 ÖRNEKLEMİNDE VERİ MADENCİLİĞİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Yukarıda başlığı/konusu gösterilen tez çalışmam:

1. İnsan ve hayvan üzerinde deney niteliği taşımamaktadır,
2. Biyolojik materyal (kan, idrar vb. biyolojik sıvılar ve numuneler) kullanılmasını gerektirmemektedir.
3. Beden bütünlüğüne müdahale içermemektedir.
4. Gözlemsel ve betimsel araştırma (anket, ölçek/skala çalışmaları, dosya taramaları, veri kaynakları taraması, sistem-model geliştirme çalışmaları) niteliğinde değildir.

Hacettepe Üniversitesi Etik Kurullar ve Komisyonlarının Yönergelerini inceledim ve bunlara göre tez çalışmamın yürütülebilmesi için herhangi bir Etik Komisyondan/Kuruldan izin alınmasına gerek olmadığını; aksi durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Taha ESER
(Öğrencinin Adı Soyadı, İmzası)

Öğrenci Bilgileri

Adı Soyadı	**Mehmet Taha ESER**
Öğrenci No	N13245164
Anabilim Dalı	Eğitim Bilimleri
Programı	Eğitimde Ölçme ve Değerlendirme
Statüsü	☐ Yüksek Lisans ☒ Doktora ☐ Bütünleşik Dr.

Danışman Görüşü ve Onayı

Hazır veri kullanıldığı için etik kurul iznine gerek yoktur.

Yrd. Doç. Dr. Derya ÇOBANOĞLU AKTAN
(İmza)
(Danışmanın ünvanı, Adı ve Soyadı)

EK-B: Etik Beyanı

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı bütün bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin bütününü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

05/08/2019

Mehmet Taha ESER

EK-C: Yüksek Lisans/Doktora Tez Çalışması Orijinallik Raporu

05/08/2019

HACETTEPE ÜNİVERSİTESİ

Eğitim Bilimleri Enstitüsü

Eğitim Bilimleri Ana Bilim Dalı Başkanlığına,

Tez Başlığı: Uluslararası Öğrenci Değerlendirme Programı 2015 Verilerinin Veri Madenciliğinde Kümeleme Yöntemleriyle İncelenmesi

Yukarıda başlığı verilen tez çalışmamın tamamı (kapak sayfası, özetler, ana bölümler, kaynakça) aşağıdaki filtreler kullanılarak **Turnitin** adlı intihal programı aracılığı ile kontrol edilmiştir. Kontrol sonucunda aşağıdaki veriler elde edilmiştir:

Rapor Tarihi	Sayfa Sayısı	Karakter Sayısı	Savunma Tarihi	Benzerlik Oranı	Gönderim Numarası
01/08/2019	118	187844	20/06/2019	%10	1156838900

Uygulanan filtreler:

1. Kaynaklar hariç
2. Alıntılar dâhil
3. 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan eder, gereğini saygılarımla arz ederim.

Ad Soyadı: Mehmet Taha ESER

Öğrenci No.: N13245164

Ana Bilim Dalı: Eğitim Bilimleri

Programı: Eğitimde Ölçme ve Değerlendirme

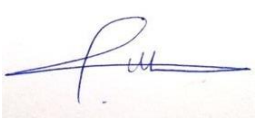
Statüsü: ☐ Y.Lisans ☒ Doktora ☐ Bütünleşik Dr.

İmza

DANIŞMAN ONAYI

UYGUNDUR.

Dr. Öğretim Üyesi Derya ÇOBANOĞLU AKTAN



EK-Ç: Thesis/Dissertation Originality Report

05/08/2019

HACETTEPE UNIVERSITY

Graduate School of Educational Sciences

To The Department of Department of Educational Sciences

Thesis Title: Examination of the Program for International Student Assessment 2015 Data by Clustering Methods in Data Mining

The whole thesis that includes the *title page, introduction, main chapters, conclusions and bibliography section* is checked by using **Turnitin** plagiarism detection software take into the consideration requested filtering options. According to the originality report obtained data are as below.

Time Submitted	Page Count	Character Count	Date of Thesis Defense	Similarity Index	Submission ID
01/08/2019	118	187844	20/06/2019	10%	1156838900

Filtering options applied:

1. Bibliography excluded
2. Quotes included
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Educational Sciences Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Name Lastname: Mehmet Taha ESER

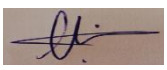
Student No.: N13245164

Department: Educational Sciences

Program: Educational Measurement and Evaluation

Status: ☐ Masters ☒ Ph.D. ☐ Integrated Ph.D.

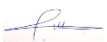
Signature



ADVISOR APPROVAL

APPROVED

Assistant Professor Derya ÇOBANOĞLU AKTAN



EK-D: Yayımlama ve Fikrî Mülkiyet Hakları Beyanı

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım haklan bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

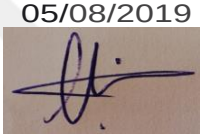
Yükseköğretim Kurulu tarafından yayınlanan "**Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına ilişkin Yönerge**" kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

X Enstitü/ Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihinden itibaren 2 yıl ertelenmiştir. (1)

o Enstitü/Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren … ay ertelenmiştir. (2)

o Tezimle ilgili gizlilik kararı verilmiştir. (3)

05/08/2019



Mehmet Taha ESER

"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"

(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma *sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü Üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

(2) Madde 6. 2. *Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç; imkânı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanın önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü Üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

** Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.*