T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANA BİLİM DALI
İSTATİSTİK BİLİM DALI

TÜRKİYE'DE BİLİŞİM TEKNOLOJİLERİ KULLANIMI KARMA VERİ SETİNİN KÜMELEME ANALİZİ İLE İNCELENMESİ

Yüksek Lisans Tezi

Derin SAVAŞAN

İSTANBUL, 2023

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANA BİLİM DALI
İSTATİSTİK BİLİM DALI

TÜRKİYE'DE BİLİŞİM TEKNOLOJİLERİ KULLANIMI KARMA VERİ SETİNİN KÜMELEME ANALİZİ İLE İNCELENMESİ

Yüksek Lisans Tezi

Derin SAVAŞAN

DANIŞMAN
Dr. Öğr. Üyesi Özlem ERGÜT

İSTANBUL, 2023

ÖZET

TÜRKİYE'DE BİLİŞİM TEKNOLOJİLERİ KULLANIMI KARMA VERİ SETİNİN KÜMELEME ANALİZİ İLE İNCELENMESİ

Bilişim teknolojilerindeki gelişmelerin katlanarak artan hızı, toplumun yapı taşı olan bireyden başlayarak, küresel ölçekte toplumsal yapı üzerinde büyük bir etkiye sahiptir. Bu bakış açısı ile toplumlarda meydana gelen ekonomik, sosyal ve kültürel etkilerin ortaya çıkarılması büyük önem taşımaktadır. Bu çalışmanın amacı, farklı değişkenler ile hane ve fert bazında gözlemleri, bilişim teknolojileri kullanım şekillerine göre kümeleyerek, elde edilen kümelerin özelliklerinin incelenmesidir. Analizde, Türkiye İstatistik Kurumu tarafından Türkiye'de hane ve bireylere uygulanan 'Hanehalkı Bilişim Teknolojileri Kullanım Araştırması' anketinin 2019 ve 2021 yıllarında toplanan verileri kullanılmıştır. Veri setinin büyüklüğü ve sayısal değişkenler ile kategorik değişkenleri bir arada bulundurması göz önüne alınarak, veri setinin analizinde KAMILA algoritması kullanılmıştır. Veri setinin koronavirüs (Covid-19) pandemisi öncesi ve sonrasına ait olması, bilişim teknolojileri ve ürünleri kullanımının hem hane hem de fert düzeyinde küme özelliklerinin karşılaştırılması ve farklılıkların belirlenmesi açılarından faydalı olmuştur. Hane düzeyinde yapılan analizde, pandemi öncesi ve sonrası kümelerin özelliklerinde değişikliklerin olduğu gözlemlenmiştir. Fert düzeyinde yaplan iki analizde de, benzer bir sonuçla pandemi öncesi ve sonrası kümelerin özelliklerinde farklılıkların oluştuğu ve küme profillerinin değiştiği belirlenmiştir.

Anahtar Kelimeler: Karma Tipte Veri, Kümeleme Analizi, Bilgi Teknolojileri Kullanma Eğilimleri

ABSTRACT

AN INVESTIGATION ON THE USE OF INFORMATION TECHNOLOGIES IN TURKEY WITH MIXED TYPE DATA CLUSTER ANALYSIS

The exponentially increasing speed of developments in information technologies has a great impact on the social structure on a global scale, starting with the individuals who are making up the society. With this point of view, revealing the economic, social, and cultural effects that occur in societies has a great importance. The aim of this study is to examine the profile of the clusters that obtained by the observations on the basis of household and individual usage patterns of information technologies with different variables. In the analysis, the data collected in 2019 and 2021, under the name "Household Information Technologies Usage Survey" applied to households and individuals in Turkey by the Turkish Statistical Institute were used. Considering the size and type of variables in the dataset, the KAMILA algorithm was used in the analysis of the data set. The fact that the data set belongs to pre and post coronavirus (Covid-19) pandemic has been beneficial in terms of determining the effects on both household and individual levels in terms of the use of information technologies and products. In the analysis conducted for household level, it was observed that there were changes in the profile of the clusters before and after pandemic. Likewise, it was determined that there were differences in the profiles of the clusters before and after pandemic in the analysis accompanied at individual level.

Keywords: Mixed Type Data, Clustering Analysis, Trends in Using Information Technologies

ÖNSÖZ

Dünyayı tartışmasız olarak etkileyen güncel konulardan biri koronavirüs (Covid-19) pandemisidir. Covid-19 salgını, dünya genelinde bireyler, işletmeler ve ülkeler için büyük belirsizliklere, zorluklara ve endişelere neden olmuştur. Pandeminin olumsuz etkilerinin azaltılması gayretiyle pek çok alanda önlemler alınmış ve düzenlemeler getirilmiştir. Bu dönem eğitim, çalışma, iletişim ve yaşam şekillerinde temel değişiklikleri zorunlu hale getirmiş ve dönüşümün bilinmeyen bir şekilde başladığı kritik bir dönüm noktası haline gelmiştir.

Bilgisayar teknolojileri ve iletişim teknolojilerinin bir arada kullanılmasını vurgulayan bilişim teknolojileri, yakın tarihte hızlı bir ivme ile gelişmiştir. Geliştirilen teknolojiler, Covid-19 pandemisi boyunca ortaya çıkan ihtiyaçlar doğrultusunda, bireylerin kullanımıyla evrilmiş ve günlük hayatın önemli bir parçası haline gelmiştir. Bu çalışmada, Covid-19 pandemisinin toplum düzeyinde yarattığı etkilerin incelenmesi amaçlanmıştır. Bu amaç doğrultusunda Türkiye İstatistik Kurumu'nca derlenen 'Hanehalkı Bilişim Teknolojileri Kullanım Araştırması' anket verileri, hane ve fert bazında bilişim teknolojilerinin kullanımının sosyo-demografik ve ekonomik faktörleri açısından incelenmiştir. Bu çalışmaya konu olan verilerin 2019 ve 2021 yıllarına ait olması sebebiyle, elde edilen sonuçlar pandemi öncesi ve sonrası için karşılaştırma imkanı sağlamaktadır.

Tez konusunun belirlenmesi ve veri setine ulaşmam konusunda yol gösteren, çalışmalarım süresince değerli bilgi, kaynak ve katkılarını esirgemeyen sayın hocam Dr. Öğr. Üyesi Özlem ERGÜT'e, bölümde eğitime başlamadan önce sorduğum sorulara verdiği yanıtlarla bana rehber olan sayın hocam Prof. Dr. Ahmet Mete ÇİLİNGİRTÜRK'e, pandemi sürecinde özveri ile programın devamlılığını sağlayan sayın hocalarım Prof. Dr. Dilek ALTAŞ KARACA, Prof. Dr. İlknur Esen YILDIRIM ve Prof. Dr. Selay GİRAY YAKUT'a sonsuz teşekkürlerimi sunuyorum.

Derin SAVAŞAN

İstanbul, 2023

İÇİNDEKİLER

Sayfa No

KISALTMALAR

BT:	Bilişim Teknolojileri
EM:	Beklenti Maksimizasyonu (Expectation Maksimization)
HHB:	Hanehalkı Birey Sayısı
IBBS:	İstatistiki Bölge Birimleri Sınıflandırma
ISCO 08:	International Standard Classification of Occupation Uluslararası standard Meslek Sınıflaması
ISDN:	Integrated Services Digital Network
KAMILA:	KAy- means for MIxed LArge data
TÜİK:	Türkiye İstatistik Kurumu

TABLOLAR LİSTESİ

Sayfa No

ŞEKİLLER LİSTESİ

Sayfa No

1. GİRİŞ

Bilgisayar teknolojileri ve iletişim teknolojilerinin bir arada kullanılmasını vurgulayan bilişim teknolojileri, yakın tarihte hızlı bir ivme ile gelişmiş ve günümüzde de bu gelişmeyi sürdürmektedir. Bireylerin içinde bulundukları koşullara, ellerinde bulunan imkanlara göre zaman içerisinde farklı davranışlar geliştirmesi sonucu, bilişim teknolojilerinin kullanımı bireylerin ve oluşturdukları toplumların gelişmişlik düzeylerini belirleyen önemli unsurlardan biri haline gelmiştir. Bilgiye ulaşma hızının ve yollarının her geçen gün daha kolay ve çeşitli hale geldiği günümüzde, bilişim teknolojilerinin kullanımı günlük hayatta daha fazla yer almaya devam etmektedir. Teknolojik gelişmeler ışığında, bireylere sunulan hizmetlerin çoğalması ve dolayısı ile yaşamı kolaylaştırılmasının sonucu olarak, insanlar yaşamın çeşitli alanlarında sunulan imkanları kullanarak daha kısa sürelerde sonuç elde etmek istemektedir. Geçmişte fiziksel olarak yapılan günlük faaliyetlerin bir kısmı (bankacılık işlemleri, görüntülü görüşme, dava dosyası takibi, mal veya hizmet satışı, eğitim alma veya verme, giyilebilir teknolojilerle ölçüm yapma vb.) artık çevrimiçi olarak, bulunulan yerden ve kişiselleştirilmiş saat dilimlerinde yapılabilmektedir. Bilişim teknolojilerinin kullanımı günümüzde üç boyutlu ve içeriğini kullanıcılarının oluşturdukları sanal dünyaya kadar evrilmiştir.

Teknolojik gelişmelerin hayatımıza getirdiği kolaylıklardan doğru faydalanmak konusunda toplumsal olarak bilinç geliştirmek önemini korumakta ve bu yanı ile farklı disiplinler için araştırma ve tartışma konusu olmaya devam etmektedir. Bu gelişmelerin sonuçlarından biri olan veri depolama ve işleme kapasitelerinin katlanarak artması, daha büyük hacimdeki verilerin anlık olarak veya sonrasında işlenmesi sonucu, veri kaynağındaki eğilimleri analiz edilip, sonuçlar elde edilmesine ve bu sonuçlara göre yeni kararlar alınmasına olanak sağlamaktadır. Bu bağlamda toplumlarda meydana gelen ekonomik, sosyal ve kültürel etkilerin de ortaya çıkarılmasında büyük önem taşımaktadır.

Dünyayı tartışmasız olarak etkileyen güncel konulardan biri koronavirüs (Covid-19) pandemisidir. Virüs, 2019' un sonlarında önce Çin'de ortaya çıkmış ve ardından tüm dünyaya hızla yayılmıştır. Covid-19 salgını, dünya genelinde bireyler, işletmeler ve ülkeler için bütük bir belirsizliğe, zorluk ve endişelere neden olmuştur. Bu dönem eğitim, çalışma, iletişim ve yaşam şekillerinde temel değişiklikleri zorunlu hale getirmiş ve dönüşümün bilinmeyen bir şekilde başladığı kritik bir dönüm noktası haline gelmiştir. Covid-19 pandemisinin toplum düzeyinde yarattığı etkileri incelemek için, pandemi öncesi ve sonrası çerçevesi benzer verilere ulaşılması önemlidir. Bu amaca yönelik, Türkiye İstatistik Kurumu'nun 2004 yılından bu yana (2006 senesi hariç) her yıl düzenli olarak Türkiye'de hane ve bireylerine uygulanan anket aracılığıyla 'Hanehalkı Bilişim Teknolojileri Kullanım Araştırması' adı altında toplamakta olduğu veriler araştırmacılar için oldukça önemli ve zengin içerikte bir kaynaktır. Çalışmanın amacı farklı değişkenler ile gözlemlerin, hane ve fert bazında bilişim teknolojilerinin

kullanımının sosyo-demografik ve ekonomik faktörler açısından analiz edilmesi ve gözlemlerin aynı kümede sınıflanarak, küme özelliklerinin belirlenmesidir. Bu çalışmaya konu olan veriler 2019 ve 2021 yıllarına aittir. Veri setlerinin 2019 ve 2021 senesine ait olması sebebiyle, elde edilen sonuçlar pandemi öncesi ve sonrası için karşılaştırma imkanı sağlayacaktır.

Belirlenen amaç doğrultusunda seçilen değişkenler yardımıyla gözlemlerin gruplara ayrılmasında kümeleme analizi kullanılmıştır. Çok değişkenli analiz tekniklerinden olan kümeleme analizi ile ilgili olarak yapılan çalışmalarda veri setindeki değişkenlerin tipi, yapılacak kümeleme analizinin kararlaştırılmasında belirleyici olmaktadır. Bu çalışmada kullanılacak veri seti kategorik değişkenler ile oransal ölçekli sayısal değişkenler içermektedir. Farklı ölçeklerle ölçülmüş karma tipteki değişkenleri içeren veri setleri ile yapılan akademik çalışmalarda, k- ortalamalar, k-ortaylar, k-prototipler gibi algoritmaların sıklıkla kullanıldığı gözlenmektedir. Bu çalışmada karma yapıda ve özellikle büyük veri setlerinin analizi için oluşturulmuş olan KAMILA (KAy- means for MIxed LArge data) algoritması, R Studio açık kaynaklı programlama dili ortamında kullanılmıştır.

Bu çalışmada Kümeleme Analizi yöntemleri incelenecek, kullanım alanları ve diğer çok değişkenli analiz yöntemleri ile arasındaki ilişkilerine yer verilecektir. Kümeleme Analizinde önemli bir aşama olan uzaklık ve birliktelik ölçülerini deyatlandırılacaktır. Kümeleme analizi yöntemlerinin, R Studio açık kaynak programlama ortamında sıkça kullanılan algoritmaları hakkında bilgi verilecektir. Çekirdek yoğunluk tahmini, çoklu nominal model, KAMILA algoritmasının modeli ve ve tahmin gücü konularına yer verilecektir. KAMILA algoritması ile uygulaması yapılacak kümeleme analizi ile hane bazında gözlemler, hanede bulunan kişi sayısı, hane toplam aylık geliri, hanede bulunan bilişim ekipmanları, hanede kullanılan internet bağlantısı türleri ve istatistiki bölge birimleri değişkenlerine göre kümelenecek ve kümelerin benzerlikleri incelenerek özellikleri ortaya konmaya çalışılacaktır. Fert bazında yapılacak analiz de ise, iki sorunun cevabı aranacaktır. Bu sorulardan ilki internet üzerinden mal ve hizmet alışverişi yapan bireylerin yaş, cinsiyet ve e-ticaret alışkanlıklarına göre gözlemlerin kümelenmesi ve küme özelliklerinin incelenmesidir. İkinci olarak fertlerin, taşınabilir cihazlar ile internetteki faaliyetlerinin, yaş, internet kullanım sıklığı, eğitim durumları ve meslek gruplarına göre sınıflandırılması ve oluşan sınıfların özelliklerinin araştırılmasıdır.

2. KÜMELEME ANALİZİ

Bu bölümde küme tanımı, kümeleme analizinin amacı, aşamaları, kümeleme analizinde kullanılan uzaklık, benzerlik ve birliktelik ölçüleri, kümeleme analizinin yöntemleri ve R Studio programındaki uygulamalarından bahsedilecektir.

Kümeleme analizi veriyi oluşturan gözlem birimlerinin, seçilen değişkenlere göre benzerlikleri bakımından incelenerek ayrılmasını sağlayan çok değişkenli istatistiksel bir yöntemdir. Ayırma işleminde amaç birbirine en çok benzeyen birimleri bir araya getirerek ait oldukları alt grupları belirlemek; birbirine en benzemeyen birimlerin ise farklı alt gruplarda bulunmasını sağlamaktır. Birbirlerine en çok benzeyen birimlerin bir araya gelmesi ile oluşan alt grupların her biri küme olarak değerlendirilmektedir. Kümelemede, grupların örtüşmesinden ziyade, birbirinden ayrışması istenmektedir. Birbiri ile örtüşen kümelerde, birimler birden fazla gruba ait olabilmekte, ayrışan kümelerde ise her birim bir gruba ait olmaktadır.

Bir kümeleme şeması, verideki benzerlik ve farklılıkların modellerini tanımlanması yoluyla kümelerin isimlendirilmelerini sağlayan; büyük ve karmaşık çok değişkenli verileri ise sınıflandırmak için uygun bir yöntem sağlamalıdır (Everitt, Hothorn 2011:163).

Kümeleme Analizi ham veri matrisindeki gözlemlerin, bazen de değişkenlerin sahip oldukları özellikler çerçevesinde kümelemek amacıyla geliştirilmiş yöntemler topluluğu olarak da tanımlanabilir (Alpar 2021:319). Her bir gözlem birimi için, analize dahil edilen değişkenlerin sağladıkları bilgiler, bir markaya ait ürünlerin özellikleri, hastalıklara dair belirtilerin gözlenip gözlenmeme durumu veya bireylerin online hizmetlerden yararlanma araçları ve sıklığı olabilmektedir. Kümeleme analizinde oluşturulacak kümelerin sayısı analiz öncesi çoğunlukla bilinmemekte ve küme sayısının belirlenmesi, analiz aşamlarından biri olmaktadır.

Kümeleme Analizi gözlemlerin uzaklık veya benzerlik ölçülerine göre homojen gruplara ayıran; böylece araştırılan konu ile ilgili özet bilgi elde edilmesini sağlayan, farklı alanlarda sıkça kullanılan çok değişkenli analiz yöntemlerinden biridir. Kümeleme Analizi'nde gözlemlerin küme içi benzerlikleri maksimum, kümeler arası benzerlik minimum olacak biçimde gruplandırılması amaçlanmaktadır (Ergüt 2020: 73). Diğer bir ifade ile, küme içi benzerliklerin maksimum olduğu homojen gruplar, kümeler arası değerlendirildiğinde heterojen gruplar oluşturmaktadır.

Kümeleme Analizi genellikle en az üç temel aşamadan oluşmaktadır. İlk adım, veri içerisinde gerçekte kaç kümenin var olduğunu belirlemek için gözlem birimleri arasındaki benzerlik veya ilişkinin ölçülmesidir. İkinci adım, gözlem birimlerinin gruplara bölündüğü kümeleme işlemidir. Son adım ise, belirlenen küme yapılarının özelliklerinin belirlenmesidir (Hair vd 2010: 18). İlk adımda benzerlikleri

belirlemek amacıyla benzerlik/ uzaklık veya ilişkinin ölçüsü şeçilmekte, sonraki adımda kullanılacak kümeleme tekniğine karar verilip, küme sayısı ve küme üyelikleri belirlenmektedir. Son adımda ise elde edilen kümelerin özellikleri incelenerek araştırmanın amacına yönelik özetleyici bilgiler çıkarılması amaçlanmaktadır.

Benzerlik veya homojenlik tanımı analizden analize değişmektedir ve çalışmanın amacına bağlıdır. Bir iskambil destesi düşünüldüğünde 52 oyun kartı, bir dizi farklı şema kullanılarak gruplandırılabilir. Bir şemada tüm kırmızı kartlar bir grupta ve tüm siyah kartlarda başka bir grupta olabilir. Her biri oyunun amacına bağlı olan bir dizi farklı gruplama şemasına sahip olunabileceği açıktır (Sharma 1996: 185). Buna göre bir kümeyi oluşturan birimler birbiriyle benzeşirken, diğer kümelerin birimleri biririnden farklılaşacaktır.

Kümeleme Analizi kullanım yerleri incelendiğinde; bir yatırım bankacılığı firmasının finansal analistinin, devralmalar için ana hedef olan bir grup firmayı belirlemekle ilgilenmesi; bir pazarlama yöneticisinin, test pazarlaması için kullanılabilecek benzer şehirleri belirlemekle ilgilenmesi; bir siyasi adayın kampanya yöneticisinin, önemli noktalarda benzer görüşlere sahip seçmen gruplarını belirlemekle ilgilenmesi gibi senaryoların her biri, belirli özellikler bakımından birbirine benzeyen varlık veya özne gruplarının belirlenmesiyle ilgilidir (Sharma 1996: 185). Bu yönü ile Kümeleme Analizi belirsizlik koşullarının ve karmaşık yapıların bulunduğu alanlarda, değerlendirme ve çözümleme amacıyla kullanılabilmektedir.

Kümeleme Analizi yardımıyla, incelenen p değişken açısından aykırı/aşırı değer olarak nitelendirilebilecek gözlemler belirlenebilmektedir (Alpar 2021:320). Örneğin hanede kullanılan teknolojik araçların kullanımının (cep telefonu, tablet, dizüstü bilgisayar, masaüstü bilgisayar, internete bağlanan TV, oyun konsolu gibi) bireylerin eğitim durumu, cinsiyeti, mesleği ve yaşı değişkenleri ile incelenmesinde, bireylerden bir veya bir kaçından oluşan küme yapıları aykırı / uç değerler olarak isimlendirilebilmektedir.

Kümeleme Analizi'nin diğer çok değişkenli analiz yöntemleri arasındaki ilişkiler incelendiğinde;

Gözlem birimlerinin gruplandırılmasında kullanılması nedeniyle kümeleme ve ayırma analizleri arasında benzerlik olmakla birlikte, iki yöntem arasında önemli farklılıklar da bulunmaktadır. Ayırma analizinde küme sayısı bilinmekte, bu sayı analiz süresince değişmemekte ve araştırmacıdan gözlem birimlerini bu kümelere sınıflandırması istenmektedir. Ayrıca ayırma analizinden elde edilen ayırma fonksiyonu, daha sonraki analizlerde kullanılabilmektedir (Tatlıdil, 2002: 329). Diğer bir ifade ile, ayırma analizinde elde edilen fonksiyonlar, analize yeni gözlem birimlerinin eklenmesi durumunda, bu gözlem birimlerinin hangi kümye ait olacağı konusunda atama yapılmasına yardımcı olmaktadır.

Kümeleme Analizi'nde amacın veride var olan durumun ve buna bağlı olarak küme sayısının belirlenmesi olması sebebiyle, analize eklenen yeni gözlem birimlerinin hangi kümeye ait olacağı konusunda atama yapmaya yardımcı olması söz konusu değildir.

Kümeleme Analizi'nde, ayırma analizinde olduğu gibi verilerin normal dağılımlı olması gerektiği varsayımı olmakla birlikte, normallik varsayımı prensipte kalmaktadır (Tatlıdil, 2002: 329). Normallik, doğrusallık ve sabit varyanslılık gibi diğer istatistiksel yöntemler için yaşamsal önem taşıyan varsayımların Kümeleme Analizi'ndeki önemi çok azdır. Bu nedenle Kümeleme Analizi'nin uygulanmasında daha çok örneklemin evreni temsil edip etmediği ve çoklubağlantılı değişkenlerin olup olmadığı konuları üzerine yoğunlaşmaları önerilmektedir. Kümeleme Analizi, diğer birçok çok değişkenli analizde olduğu gibi aşırı değerlere karşı duyarlıdır. Kümeleme Analizi'nde örneklem büyüklüğü istatistiksel güç çerçevesinde ele alınmaz. Bu kapsamdaki en önemli nokta, incelenecek örneklemin evrendeki alt grupları yansıtacak büyüklükte ve özellikte olmasıdır (Alpar 2021:321-322).

Uzaklık ölçülerinin kullanıldığı kümeleme yöntemleri, kullanılan ölçü birimi farklılıklarına duyarlıdır. Bunun sebebi tanım aralığı geniş olan değişkenin uzaklık veya benzerlik ölçüleri üzerindeki etkisinin daha fazla etkisi olmasıdır. Böyle durumlarda verilerin analiz öncesi uygun olan dönüştürme ve standartlaştırma işlemlerine tabii tutulması uygun olmaktadır.

Kümeleme Analizi, verinin kendi yapısında doğal olarak bulunan benzerliklerin sınıflandırılmasını önermesi sebebiyle faktör analizi ile karşılaştırılabilir. Kümeleme Analizi'nde gözlem birimleri sınıflandırılırken, Faktör Analizi öncelikle değişkenlerin gruplanması ile ilgilidir (Hair, 2010: 508). Bu yönüyle analizler birbirlerinden farklılaşmaktadır.

Kümeleme Analizi değişkenleri, bağımlı (kriter) ve bağımsız (tahmin edici) değişkenler şeklinde ikiye ayırmaz (Nakip, Yaraş 2017: 545). Bu yönüyle Kümeleme Analizi, Faktör Analizi'ne benzemektedir.

Bu bilgilere ek olarak, Faktör Analizi'nde gruplandırma verideki değişkenliğe (korelasyona) göre modellenirken, Kümeleme Analizi mesafeye (yakınlığa) göre gruplandırma yapmaktadır (Hair, 2010: 508).

Kümeleme Analizi tanımlayıcı yöntemlerden biridir. Gözlem birimlerinin kaç küme içerisinde gruplandırılacağı ve hangi gruba atanacağı, şeçilen uzaklık, benzerlik veya birliktelik ölçüsüne ve kümeleme yöntemine göre değişiklik gösterebilmektedir. Bu sebeple, aynı veri setinin kullanıldığı durumlarda dahi, kümeleme analizinin farklı sonuçlar vermesi gözlenebilir. Diğer bir ifade ile aynı veri setinden birbirinden farklı kümelere ulaşmak mümkün olmaktadır. Bu sebeple sonuçlar yorumlanırken, seçilen uzaklık, benzerlik veya birliktelik ölçüsü ve kümeleme yöntemindeki farklılıklar ve araştırmanın amacı göz önünde bulundurulmalıdır.

2.1. Kümeleme Analizi'nin Aşamaları

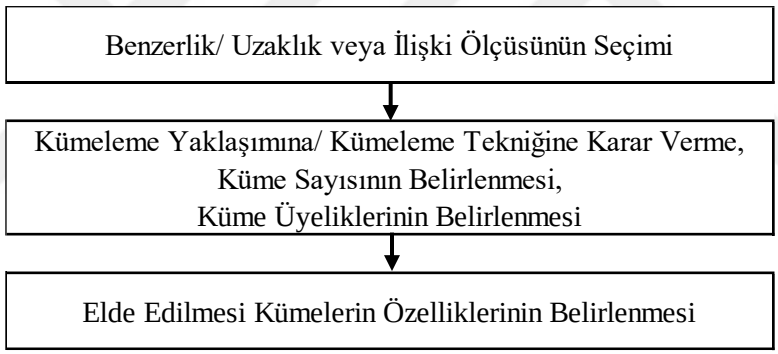
Kümeleme Analizi üç temel aşamadan oluşmaktadır.

1.Benzerlik/ uzaklık veya ilişki ölçüsünün seçimi: Bu adımda birimlerin benzerlik, uzaklık veya ilişkilerini belirlemek amacıyla benzerlik / uzaklık veya ilişki ölçüsünün seçimi yapılmaktadır.

2. Kümeleme yaklaşımına/ kümeleme tekniğine karar verme, küme sayısını belirleme ve küme üyeliklerinin belirlenmesi: Bu adımda kullanılacak kümeleme yaklaşımına (çoğunlukla hiyerarşik ve hiyerarşik olmayan yöntemler) veya kümeleme tekniğine karar verilmektedir. Sonrasında küme sayısı ve küme sayısına uygun olarak küme atamaları suretiyle gözlem birimlerinin gruplandırması tamamlanmaktadır.

3.Elde edilen kümelerin özelliklerinin belirlenmesi: Bu adımda elde edilen küme üyelikleri incelenerek, araştırmanın amacı doğrultusunda küme özellikleri belirlenmektedir.

Bu adımlardan sonra, küme özellikleri gözönüne alınarak, elde edilen burgular araştırmacı tarafından yorumlanmaktadır. Kümeleme Analizi'nin temel aşamaları Şekil 2.1'de verilmektedir.



Şekil 2.1: Kümeleme Analizi'nin Temel Aşamaları

Kümeleme Analizi'nde önemli iki nokta, uzaklık/ benzerlik veya birliktelik/ ilişki ölçüsünün belirlenmesi ile uygun kümeleme tekniğinin belirlenmesi adımıdır. Yapılan seçimler doğrultusunda, değişkenlerin aldığı değerlerin incelenmesi ve gerekli olduğu durumlarda dönüştürme ve standartlaştırma işlemlerinin tamamlanması da analiz öncesinde önem arz etmektedir. Kümeleme Analizi'nin sonucu bahsi geçen tüm işlemlerin uygulanıp uygulanmama durumuna göre farklılıklar gösterebilmektedir.

Kümeleme Analizi'nde kullanılacak veri matrisinde yer alacak olan değişkenlerin konunun kuramsal ve uygulamaya yönelik yanları dikkate alınarak seçilmesi önerilmektedir. Diğer tüm çok değişkenli çözümlemeler gibi Kümeleme Analizi'nde değişken seçiminde yapılacak yanlış seçimleri ortadan kaldırma becerisine sahip değildir (Alpar 2021:320). Analizde kullanılacak değişkenlerin

seçilmesinde, konu ile ilgili olarak daha yapılmış akademik çalışmalarla birlikte, konu hakkında uzman sayılabilecek bilgiye sahip kişilerden yardım alınması da mümkündür. Bir veya birkaç gözlem biriminden oluşan küme yapılarının elde edilmesi durumunda, bu yapıların aşırı/ aykırı gözlem birimlerinden oluşup oluşmadığının belirlenmesi de analiz sonucunun sağlıklı bir şekilde yorumlanmasında önem taşımaktadır.

Çok değişkenli istatistiksel çözümleme n gözlemli ve p değişkenli (**nxp** boyutlu) X veri matrisi ile ilgilenmektedir. Kümeleme Analizi'nin (aynı zamanda çok boyutlu ölçekleme ve faktör analizinin) başlangıç noktası, veriden elde edilen bu tür matrislerdir (Alpar 2021:165). Veri matrisinden yer alan n gözlem biriminin p değişkene göre uzaklıkları, uzaklık matrisi (distance matrix) adı verilen **D** matrisi ile gösterilir. Gözlem birimlerinin birbirleri ile olan benzerlik (similarity, Sim) matrisi **S** ile gösterilir. Benzerlik matrisinin elemanları, **D** matrisinin elemanlarına göre belirlenir. Gözlem birimlerinin birbirlerinden farklıklarını belirten (dissimilarity) **D** matrisi, **S** (Sim) matrisinin elemanları aracılığı ile hesaplanır (Özdamar 2004:284).

Uzaklık ve benzerlik ölçüleri bazı çok değişkenli yöntemler için bir alt yapı oluşturmasının yanı sıra, yalnızca tanımlayıcı amaçlar çerçevesinde de kullanılabilmektedir (Alpar 2021:165).

Kümeleme Analizi'nin temel amacı birbirine yakın gözlemlerin belirlenmesi ve aynı küme içerisinde gruplanması olduğundan, uzaklık ve benzerlik ölçüleri gözlemler için hesaplanmakta ve **nxn** boyutlu uzaklık veya benzerlik matrisi elde edilmektedir. Benzer şekilde, faktör analizinde **pxp** boyutlu bir benzerlik matrisi olan korelasyon matrisinden yola çıkılmaktadır (Alpar 2021:165).

Gözlemler için elde edilen uzaklık ve benzerlikler, n tane gözlem için n(n-1)/2 tane uzaklık benzerlik söz konusu olmakta ve **nxn** boyutlu bir matriste özetlenirken, değişkenler için elde edilen uzaklık ve benzerlikler **pxp** boyutlu bir matris ile özetlenmektedir (Alpar 2021:166).

Veri matrisinde yer alan n birimin p değişkene göre uzaklıkları, Uzaklık Matrisi adı verilen **D** matrisi ile gösterilmektedir.

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ \mathbf{d_{21}} & 0 & d_{23} & \cdots & d_{2n} \\ \mathbf{d_{31}} & \mathbf{d_{32}} & 0 & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{d_{n1}} & \mathbf{d_{n2}} & \mathbf{d_{n3}} & & 0 \end{bmatrix}$$

Burada dij = d(xi,xj), xi ve xj gözlem vektörleri arasındaki uzaklık değerini göstermektedir. Bir nesnenin kendisine olan uzaklığı sıfır olacağından matriste esas köşegenin tüm elemanları sıfır

olmaktadır (Ergüt 2020:74). nxn boyutlu **D** Matrisi simetrik bir matristir. Birinci ve ikinci gözlem birimleri arasında uzaklığı gösteren d_{21} ve d_{12} uzaklıkları birbirine eşit olmaktadır.

İncelenen özellik açısından iki nesne birbirine yakınsa bunların benzer olduğunu söylemek mümkün olabilmektedir. Nesnelerin (gözlem birimlerinin) birbirleri ile olan benzerlik düzeylerini Benzerlik Matrisi (**S**) ile gösterilmesi mümkündür.

$$S = \begin{bmatrix} 1 & s_{12} & s_{13} & \cdots & s_{1n} \\ s_{21} & 1 & s_{23} & \cdots & s_{2n} \\ s_{31} & s_{32} & 1 & \cdots & s_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & s_{n3} & & 1 \end{bmatrix}$$

$s_{ij} = s(x_i,x_j)$, i'inci ve j'inci nesneler arasındaki benzerliği göstermek üzere, nesneler arasındaki mesafenin yakın olması benzerliğin yüksek, uzak olması ise nesnelerin benzerliğinin düşük olduğuna işaret etmektedir (Ergüt 2020: 74).

2.2. Uzaklık ve Benzerlik Ölçüleri

Kümeleme Analizi'nde gözlem birimleri arasındaki yakınlığın ölçümünde çeşitli yöntemler kullanılmaktadır. Bunlar, uzaklık, benzerlik ve birliktelik ölçüleridir (Sharma 1996: 218).

2.2.1. Nicel Veriler İçin Uzaklık ve Benzerlik Ölçüleri

Bilimsel ve matematiksel bakış açısından uzaklık, iki nesnenin birbirinden ne kadar uzak olduğunun nicel bir derecesi olarak tanımlanır. Uzaklık kelimesinin eş anlamlısı benzemezliği içerir. Metrik (nicel olarak ölçülebilir) özellikleri karşılayan bu uzaklık ölçülerine basitçe metrik denirken, diğer metrik olmayan mesafe ölçülerine bazen sapma denmektedir (Cha 2007: 300).

Nesneler birbirine ne kadar benzemezse uzaklık değeri büyük, birbirine ne kadar benzerse uzaklık ölçüsünün alacağı değer de o kadar küçük olacaktır. Temel olarak benzerlik nicel olup, iki nesne veya iki özellik arasındaki ilişkinin kuvveti olarak açıklamak mümkündür (Çilingirtürk 2011:166). Diğer bir anlatımla, gözlem birimleri birbirine ne kadar benzerse, uzaklık değeri o kadar küçük; gözlem birimleri birbirine ne kadar benzemezse, uzaklık değeri o kadar büyük olmaktadır. Uzaklık metrik olarak ölçülebilmekte ve yorumlanabilmektedir. Benzerliği iki gözlem biriminin yapısal olarak benzerliği veya aralarındaki ilişkinin kuvveti şeklinde yorumlamak da mümkündür.

Kümeleme algoritmaları, bir gözlem çiftinin benzerliğinin değerlendirilmesi için çeşitli ölçüler kullanır. Ölçülerden her biri, veri tipi, ölçek ve amaca bağlı olarak benzerliğin özel bir yönünü gösterecek şekilde değişik yollarla hesaplanmaktadır.

2.2.1.1. Öklit (Euclidean) Uzaklık Ölçüsü

En çok kullanılan uzaklık ölçülerinden biridir. p değişkenli bir yapıda i. ve j. gözlem arasındaki genelleştirilmiş Öklit uzaklığı aşağıdaki formül ile hesaplanabilmektedir.

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

Bu formülde,

x_{ik}:i. gözlemin k. değişken değeri,

x_{jk}:j. gözlemin k değişken değeri,

p: değişken sayısıdır.

Değişkenlerden birinin ölçüm birimi diğerine göre büyük olduğu durumlarda, bu değişken öklit uzaklığı üzerinde etkili olacaktır (Alpar 2021:168-169). Bu sebeple değişkenlerin aldığı değerlerin analiz öncesi standartlaştırılması ve bu şekilde ölçeklerden kaynaklanan farklılıkların giderilmesi önem taşımaktadır.

2.2.1.2. Kareli Öklit Uzaklık Ölçüsü

Kareli öklit uzaklığı, öklit uzaklığının karesi alınarak hesaplanmaktadır.

$$d_{ij} = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$$

2.2.1.3. Manhattan City Block Uzaklık Ölçüsü

Manhattan City Block, gözlem birimleri arasındaki mutlak uzaklıkların toplamı alınarak hesaplanmaktadır.

$$d_{ij} = \sum_{k=1}^{p}|x_{ik} - x_{jk}|$$

2.2.1.4. Minkowski Uzaklık Ölçüsü

Manhattan City Block uzaklık ölçüsü, aşadığa verilen formülde, m=2 için Öklit uzaklık ölçüsü ile aynı sonucu vermektedir.

$$d_{ij} = \left[\sum_{k=1}^{p}|x_{ik} - x_{ij}|^m\right]^{1/m}$$

2.2.1.5. Mahalanobis D^2 Uzaklık Ölçüsü

Mahalanobis kare uzaklığı iki birim değeri veya noktası arasındaki uzaklığı ölçmede, değişkenler arasındaki kovaryans veya korelasyon katsayısını da dikkate alan bir uzaklık ölçütüdür (Albayrak 2006 :40). Örneklem varyans kovaryans matrislerinin homojen olduğu varsayımı altında iki grup arasındaki Mahalanobis uzaklığı aşağıdaki eşitlik ile hesaplanabilmektedir (Alpar 2021:182).

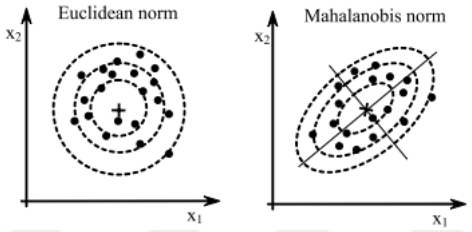
$$D_{ij}^2 = (\mu_i - \mu_j)' S^{-1} (\mu_i - \mu_j)$$

Bu formülde,

μ_i : i. grubun ortalama vektörü,

μ_j : j. grubun ortalama vektörü,

S^{-1} : pxp boyutlu varyans – kovaryans matrisinin tersini göstermektedir.



Şekil 2.2: Öklid Uzaklığı ve Mahalanobis Kare Uzaklığı

Kaynak: Abonyi, Feil (2007), *Cluster Analysis for Data Mining and System Identification*, Germany: Birkhauser Verlag AG, s.7

Şekil 2.2'de görülebileceği üzere, öklid uzaklığı için üç halka çizilmiştir. Üçüncü halkanın merkezden üç standard sapma uzaklıkta olduğu söylenebilir. Mahalanobis kare uzaklığı ise gözlem birimlerinin elips merkezine olan uzaklığının karesi olmaktadır (Alpar 2021:97).

2.2.1.6. Hotelling T2 Uzaklık Ölçüsü

İki kümenin ortalama vektörlerinin karşılaştırılmasında kullanılmaktadır.

$$T^2 = \left(\frac{n_1 n_2}{n_1 + n_2}\right) \sum_{k=1}^{p} (\mu_{ik} - \mu_{jk})' S^{-1} (\mu_{ik} - \mu_{jk})$$

2.2.1.7. Pearson Korelasyon Katsayısı

Sürekli değişkenler arasındaki doğrusal ilişkiyi göstermek amacıyla kullanılan bir ölçüdür. İ. ve j. gözlemler için Pearson korelasyon katsayısı r_{ij} ile gösterilmek üzere formülü,

$$\boldsymbol{r_{ij}} = \frac{\sum_{k=1}^{p} (x_{ik} - \overline{x_{i.}})(x_{jk} - \overline{x_{j.}})}{\left[\sum_{k=1}^{p} (x_{ik} - \overline{x_{i.}})^2 \sum_{k=1}^{p} (x_{jk} - \overline{x_{j.}})^2\right]^{1/2}}$$

ile ifade edilmektedir. Burada $\bar{xi}.$, $i.$ kümedeki değişkenlerin ortalamasını; $\bar{xj}.$, j. kümedeki değişkenlerin ortalamasını göstermektedir (Camkıran 2017:13)

2.2.1.8. Cosine Benzerlik Ölçüsü

x_i, i'inci gözlemin x değişkeni değerini, x_j, j'inci gözlemin x değişkeni değerini, p değişken sayısını göstermek üzere, formülü;

$$\text{Benzerlik}_{ij} = \frac{\sum_{i,j}^{p} x_i \; x_j}{\sqrt{\left(\sum_{i=1}^{p} x_i^2 \sum_{j=1}^{p} x_j^2\right)}}$$

şeklinde verilebilmektedir.

2.2.2. Birliktelik Ölçüleri

Kategorik verilerde, cinsiyet, bir ürüne sahip olup olmama, bir ürünü kullanıp kullanmama gibi sayısal sonuçlu olmayan değişkenlerde, birim çiftleri arasındaki benzerliğin ya da benzemezliğin belirlenmesinde kullanılan ölçülere birliktelik ölçüleri denilmektedir. Ölçüm değerleri değişkenlerin varlığı ya da yokluğu ilkesine göre hesaplanmakta, bir değişkenin varlığı genellikle 1, yokluğu 0 ile gösterilmektedir. Ölçülerin hesaplanmasında öncelikle 2x2 kontenjans (birliktelik) ya da diğer adıyla çapraz sınıflandırma tablosu oluşturulması gerekmektedir.

Tablo 2.1: İki Sonuçlu p Değişkenli Veri Örneği

	Değişkenler				
Gözlem	**1**	**2**	**3**	**..**	**p**
i	0	0	1	..	1
j	1	0	0	..	0

Tablo 2.2: İki Sonuçlu p Değişkenli Veriye Ait Kontenjans Tablosu

	Gözlem j		
Gözlem i	**1**	**0**	Toplam
1	a	b	a+b
0	c	d	c+d
Toplam	a+c	b+d	a+b+c+d = p

a: her iki gözlemde 1 değerini olan değişkenlerin sayısı

b ve c: 1-0 eşleşmelerin sayısı

d: her iki gözlemde 0 değerini olan değişkenlerin sayısı

2.2.2.1. Basit Eşleştirme Katsayısı

Bu sayı p değişken içerisinde birbirine uygun olan hücre frekanslarının (0-0 ve 1-1) oranını göstermektedir (Oktay 2017:28)

Benzerlik $_{ij} = \frac{a+d}{a+b+c+d} = \frac{a+d}{p}$

2.2.2.2. Rogers ve Tanimoto Katsayısı

Bu ölçüde birlikte uyuşum gösteren çiftler (1-1 ve 0-0) pay ve paydada dikkate alınırken, birlikte uyuşum göstermeyen çiftlere (0-1 veya 1-0) iki katı ağırlık verilmektedir. Aldığı değerler 0-1 arasında değişim göstermektedir (Alpar 2021:174).

Benzerlik $_{ij} = \frac{a+d}{a+d+2(b+c)}$

2.2.2.3. Yule'nin Q Benzerlik Katsayısı

Dağılım aralığı (-1, 1) olmaktadır. Katsayı -1 olduğunda değişkenler arasında ters yönlü kusursuz bir ilişkinin olduğu, katsayı 1 olduğunda değişkenler arasında doğru yönlü kusursuz bir ilişkinin olduğu ve katsayılar 0 olduğunda değişkenler arasında herhangi bir ilişkinin olmadığı belirtilir (Oktay 2017:21-23).

Benzerlik $_{ij} = \frac{ad-bc}{ad+bc}$

2.2.2.4. Jaccard Katsayısı

Jaccard Katsayısı benzerlik oranı olarak da bilinmektedir (Oktay 2017:29).

Benzerlik $_{ij} = \frac{a}{a+b+c}$

2.3. Kümeleme Yöntemleri

Uzaklık ve benzerlik matrislerinin elde edilmesinin ardından, kümelemede kullanılacak yönteme karar verilmesi adımı gelmektedir. Bu yöntemler hiyararşik kümeleme ve hiyerarşik olmayan kümeleme olmak üzere iki grupta toplanmaktadır. Hiyerarşik kümeleme, aşamalı kümeleme olarak da isimlendirilmektedir. Hiyerarşik olmayan kümeleme ise aşamalı olmayan kümeleme olarak da karşımıza çıkmaktadır.

2.3.1. Hiyerarşik Kümeleme Yöntemleri

Kümeleme yöntemlerinden en sık kullanılan, gözlemlerin bir dizi iç içe veya hiyerarşik olarak sınıflandırması ile sonuçlanan yöntemdir. Her gözlemin tek üyeli bir küme oluşturduğunun kabul

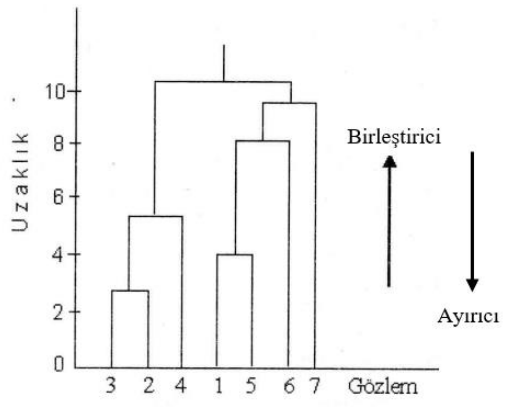
edildiği aşama ile başlamakta ve tüm gözlemlerin bir araya geldiği aşamada son bulmaktadır (Der, Everitt 2009:320). Bu bağlamda, hiyerarşik kümeleme yönteminde gözlem birimleri tek bir adımda belirli sayıda sınıfa bölünememektedir. En uygun sınıf sayısı, araştırmacı tarafından incelenerek karar verilmektedir.

Hiyerarşik kümeleme yöntemleri birleştirici (agglomerative) ve ayırıcı (divisive) olmak üzere iki grupta incelenmektedir.

Birleştici yöntemlerde, her gözlem başlangıçta tek başlarına ayrı birer küme olarak kabul edilmektedir. Daha sonraki adımda, en yakın iki küme (iki gözlem) yeni bir küme içerisinde birleştirilmektedir. Böylece her adımda küme sayısı azalmaktadır. Uygulamanın sonunda tüm gözlemler bir küme içinde gruplanmaktadır.

Ayırıcı yöntemlerinde, birleştirici yöntemlerde işleyen süreç tersine doğru yürütülmektedir. Bu sürecin başlangıcında, tüm gözlemleri içeren büyük bir küme söz konusudur. Daha sonraki adımlarda, en farklı gözlemler birbirinden ayrılarak daha küçük kümeler oluşturmaktadır. Bu süreç, her gözlem kendi başına ayrı bir küme oluşturuncaya kadar devam etmektedir.

Bu iki yöntemin yönü Şekil 2.3'teki ağaç diyagramı üzerinde gösterilmektedir. Bir ağaç gözlem birimlerinin k küme sayısını göstermek üzere, k adet gruba bölünmesinin iç içe geçmiş bir dizisi olarak tanımlanabilir. Burada k, 1'den n'e kadar herhangi bir değeri alabilmektedir. Hiyerarşik yapı genellikle dendogram adı verilen, iki boyutlu bir diyagramla temsil edilmektedir (Chatfield, Collins 1992:219). Şekil 2.3'te gösterildiği üzere iki yöntemin işleyişi birbirinin tersi yönünde olmaktadır. Ayırıcı yöntemler yukarıdan aşağıya doğru ilerlerken, birleştirici yöntemler aşağıdan yukarıya doğru ilerlemektedir.



Şekil 2.3: Birleştirici ve Ayırıcı Yöntemler için Ağaç Diyagramı

Kaynak: Alpar, R. (2021), *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*, Ankara: Detay Yayıncılık, s.324

Hiyerarşik kümeleme yöntemlerinin okunuşunun ve yorumunun kolay oluşu üstün tarafı olmakla birlikte, en sakıncalı tarafı ise sabit olmayışı ve güvenilirliğinin az oluşudur (Nakip, Yaraş 2017:552).

Farklı hiyerarşik kümeleme algoritmaları, kümeler arasındaki mesafelerin nasıl hesaplandığına göre farklılıklar göstermektedir (Sharma 1996: 188).

Kümeleri oluşturmak için sıklıkla kullanılan hiyerarşik kümeleme yöntemleri Tek Bağlantı (Single Linkage-En Yakın Komşuluk) Yöntemi, Tam Bağlantı (Complete Linkage-En Uzak Komşuluk) Yöntemi, Ortalama Bağlantı (Average Linkage) Yöntemi, Merkez (Centoid) Yöntemi, Meydan Yöntemi ve Ward Yöntemidir.

2.3.1.1. Tek Bağlantı Yöntemi

Tek bağlantı yöntemine [Single Linkage (SLINK)] en yakın komşuluk (Nearest Neighbour) yöntemi denilmektedir. Kümeleme sürecinin başında, uzaklık matrisindeki en küçük uzaklığa (en yakın) sahip iki gözlem (ya da benzerlikler matrisindeki en büyük benzerlik) dikkate alınmakta ve birinci küme oluşturulmaktadır. Daha sonra bir sonraki en küçük uzaklık bulunmaktadır. Bu süreç tüm gözlemler bir kümede toplanana kadar devam etmektedir.

2.3.1.2. Tam Bağlantı Yöntemi

Tam bağlantı yöntemi [Complete Linkage (CLINK)] en uzak komşuluk (Furthest Neighbour) yöntemi olarak da anılmaktadır. İlk aşamada, uzaklıklar matrisindeki en küçük uzaklık ile kümelemeye başlanmakta; ancak daha sonra, kümeleme sürecinde oluşturulacak yeni iki küme arasındaki uzaklık olarak kümelerdeki gözlemler arasındaki en büyük uzaklık (en az benzerlik) dikkate alınmaktadır.

2.3.1.3. Ortalama Bağlantı Yöntemi

Ortalama kümeleme yöntemindeki kümeleme süreci tek bağlantı ve tam bağlantı yöntemlerine benzemekle birlikte; kümelemenin kriteri bir kümedeki tüm gözlem birimlerinden elde edilen ortalama uzaklığın diğer kümedeki tüm gözlem birimlerine olan ortalama uzaklığıdır. Bu yöntemin bir özelliği küçük küme içi değişkenliğe sahip kümeleri birleştirme eğilimine sahip olmasıdır. Aşırı değerlerden en az etkilenen yöntemlerden biri olma özelliğine sahiptir.

Ortalama bağlantı yöntemi, tek bağlantı yönteminin birbirine en yakın komşulardan; tam bağlantı yönteminin en uzak komşulardan başlanarak kümeleme yapması ile karşılaştırıldığında, bu iki uç yöntem arasında sonuçlar vermesi sebebiyle bir alternatif olmaktadır.

2.3.1.4. Merkez Yöntemi

Bir kümeyi oluşturan gözlemlerin ortalamaları esas alınmaktadır. Kümede sadece bir gözlemin olduğu durumlarda, onun değeri merkez kabul edilmektedir. Aşırı değerlerden en az etkilenen hiyerarşik kümeleme yöntemidir (Alpar 2021: 325- 333).

2.3.1.5. Medyan Yöntemi

Merkez yönteminde iki kümenin büyüklüklerinin farklı olması durumunda, yeni kümenin merkezi daha büyük olan kümeninkine daha yakın olmaktadır. Merkez yönteminin dezavantajlarını azaltmak için önerilen bu yöntemde, yeni grubun merkezi grupların büyüklüklerinden bağımsızdır (Ergüt 2020: 82).

2.3.1.6. Ward Yöntemi

Ward'ın hiyerarşik kümeleme yöntemi, grupları birbirine bağlama yollarından ziyade grup içi kareler toplamına dayanmaktadır. Bir toplama algoritması kullanılır. Her aşamada, grup içi kareler toplamında mümkün olan en küçük artışı sağlayan iki grup birleştirilerek grup sayısı azaltılmaktadır (Chatfield, Collins 1992:224).

2.3.2. Hiyerarşik Olmayan Kümeleme Yöntemleri

Hiyerarşik olmayan kümeleme yöntemleri, k küme sayısını göstermek üzere, n sayıda gözlemin, k adet kümede gruplanması için tasarlanmıştır. Küme sayısı hakkında bir önbilgi varsa hiyerarşik olmayan kümeleme yöntemlerinin kullanılması önerilmektedir. Hiyerarşik olmayan yöntemler ilk bölünmeye karşı çok duyarlıdır.

Hiyerarşik olmayan yöntemlerde adımlar:

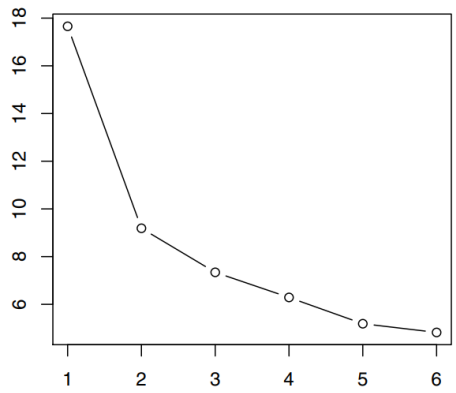
1. İlk adımda k adet kümeye ait merkezler seçilmektedir.
2. İkinci adımda her gözlem birimi en yakın olduğu kümeye atanmaktadır.
3. Bu adımda, önceden belirlenmiş durdurma kuralına göre her gözlem, k kümeden birine yeniden atanabilir.
4. Son adımda, yeniden atama yapılacak gözlem kalmadığında veya belirlenen durdurma kuralı gerçekleşince algoritmanın çalışması durmaktadır. Aksi halde ikinci adıma geçilmektedir (Sharma 1996: 202).

Hiyerarşik olmayan kümeleme yöntemlerinin hiyerarşik kümeleme yöntemlerine göre bazı avantajları şöyle sıralanabilir:

1. Hiyerarşik yöntemler daha küçük veri setleri için uygunken, hiyerarşik olmayan yöntemler çok daha büyük veri setlerine (n>1000) kolaylıkla uygulanabilmektedir. Bu yöntemlerin başlangıcında hiyerarşik yöntemlerdeki gibi gözlem sayısı boyutlarında benzerlik veya uzaklık matrisi hesaplanmamaktadır.

2. Hiyerarşik olmayan kümeleme yöntemlerinin hiyerarşik kümeleme yöntemlerime göre bir diğer avantajı verideki aykırı değerlere karşı daha az duyarlı olmasıdır.

Hiyerarşik olmayan kümeleme yöntemlerinde küme sayısı dirsek grafiği (elbow graph) yardımıyla belirlenmektedir. Şekil 2.4'te görseli bulunan dirsek grafiğinde, x ekseni küme sayılarını gösterirken, y ekseni gözlem birimlerinin küme için uzaklıklarının karesinin toplamını göstermektedir. Küme sayısı seçimi yapılırken, dirsek grafiğindeki eğimin azalıp, yataya yaklaştığı küme sayısı tercih edilir. Ancak bu değer araştırmanın amacı, eğimin ne kadar belirgin olduğu gibi etkenler göz önüne alınarak araştırmacı tarafından değerlendirilmektedir.



Şekil 2.4: Dirsek grafiği

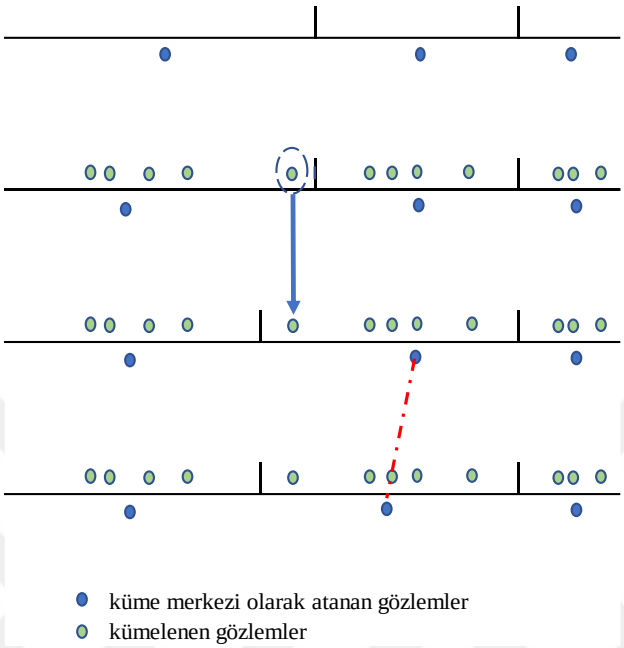
Kaynak: Everitt, B., Hothorn, T. (2011), An Introduction to Applied Multivariate Analysis with R, s.181

En bilinen hiyerarşik olmayan kümeleme yöntemi k-ortalama kümeleme yöntemidir (Alpar 2021: 340).

Hiyerarşik olmayan yöntemlerden biri olan K-ortalamalar yönteminde, her bir gözlemin kendi kümesine olan mesafeleri toplamı minimize edilecek şekilde k adet kümeye bölünmektedir. K-ortalamalar yönteminde 'k' oluşturulacak küme sayısını, 'ortalamalar' kümeyi oluşturan gözlemlerin ağırlıklı ortalamasını ifade etmektedir.

Bu yöntemde ilk aşama küme merkezlerinin belirlenmesidir. İlk önce k adet gözlem tesadüfi olarak seçilmekte ve bu gözlemlerin her biri, bir kümenin merkezini veya orta noktasını temsil etmektedir. İkinci aşamada, küme merkezleri belirlendikten sonra her bir gözlem hangi küme ortalamasına yakın olduğu belirlenmekte ve gözlem birimi o kümeye atanmaktadır. Üçüncü aşamada

her atamadan sonra küme ortalamaları yeniden hesaplanır. Dördüncü aşamada ise kümeler arasında gözlem geçişi durana kadar bir önceki adım tekrarlanmaktadır. Kümeler arası gözlem geçişi Şekil 2.5'deki gibi görselleştirilmiştir.



Şekil 2.5: Sayısal bir değişkenin k-ortalamalar yöntemi ile kümelenmesinin çizimi

K-ortalamalar yöntemi başlangıç noktasının seçimine karşı hassas bir yöntemdir. Başlangıç küme merkezi olarak atanan gözlemler, farklı kümeleme sonuçlarının oluşmasına sebebiyet verebilmektedir. Başlangıç noktalarının rastgele seçilmesi sonucunda elde edilen çözümlemelerin sıklıkla farklı sonuçlar üretmesi hiyerarşik olmayan yaklaşımların önemli bir sorunudur. Hiyerarşik olmayan kümeleme algoritmalarının farklı başlangıç noktalarına göre farklı sonuçlar üretebilmekte ancak analiz sonucu oluşan kümelenmeler arasındaki farklılıklar fazla olmamaktadır.

2.3.3. Modele Dayalı Kümeleme Analizi Yöntemleri

Model tabanlı kümeleme, verilen bazı varsayılan karma modelleme yapısı kullanılarak kümelendiği, kümeleme için istatistiksel bir yaklaşımdır. Tarihsel olarak Gauss karışım modeli, model tabanlı kümeleme literatürüne hakim olmuştur. Yeni model tabanlı kümeleme çalışmaları, eliptik olmayan dağılımların karışımlarına odaklanmıştır (Tang, Browne, McNicholas, 2015:84).

Küme yapısı için kabul edilebilir bir modelin bulunduğu durumlarda, modele dayalı kümeleme analizi istatistiksel olarak açıklanabilen çözümler verebilmektedir. Kümeleme Analizi için model olarak sonlu karışım yoğunluklarının (finite mixture densities) kullanılmasıyla kümeleme

problemi, varsayılan karışımın parametrelerini tahmin etmek ve ardından tahmin edilen parametreleri küme üyeliğinin sonsal olasılıklarını hesaplamak için kullanmak haline gelmektedir. Sonlu karışım yoğunlukları, kümeleme işlemi için istatistiksel bir model sağlamakta ve sonlu karışım modellerine dayalı kümeleme analizleri, model tabanlı kümeleme yöntemleri olarak da bilinmektedir (Everitt, Hothorn 2011:185).

2.4. Kümeleme Analizi Yöntemlerinin R Uygulamaları

Bu bölümde R Studio programlama dilinde geliştirilmiş kümeleme algoritmalarından bahsedilecektir. İlk olarak R programlama dili ve R Studio çalışma ortamı için özet bilgilere yer verilmiştir.

R programlama dili, istatistiksel hesaplama ve grafikler için ücretsiz bir yazılım ortamıdır. Çok çeşitli UNIX platformları, Windows ve MacOS üzerinde çalışabilmektedir.[1]

R programlama dili, çok çeşitli istatistiksel (doğrusal ve doğrusal olmayan modelleme, klasik istatistiksel testler, zaman serisi analizi, sınıflandırma, kümeleme vb) ve grafik teknikler sağlamaktadır.[2]

Mevcut R programlama dili, dünyanın her yerinden gelen katkılarla ortak bir çabanın sonucudur. R ilk olarak Robert Gentleman ve Auckland Üniversitesi İstatistik Departmanı'ndan Ross Ihaka tarafından yazılmıştır.[3]

R programı https://cran.r-project.org/mirrors.html adresinden ücretsiz olarak indirilip, bilgisayar kurulumu gerçekleştirilebilmektedir. Kapsamlı R arşiv ağına ulaşabilmek için, ülkeler listesinde Türkiye için sağlanan sunuculardan birinin kullanılması önerilmektedir. Bu çalışmada kullanılan kümeleme algoritması KAMILA ve tüm grafikler, arka planında R programı çalışan R Studio çalışma ortamı kullanılarak çalıştırılmıştır. R Studio istatistiksel hesaplamalar ve grafikler için geliştirilmiştir. R Studio programı https://posit.co/downloads/ adresinden ücretsiz olarak indirilebilmekte ve kurulumu tamamlanabilmektedir. Bu çalışmada kullanılan R Studio 2021.09.1 + 372 versiyonudur.

Bu çalışmada kullanılan paketler kamila[4], ggplot2[5], cluster[6] ve factoextra[7] olarak listelenebilir.

[1] https://www.r-project.org/
[2] https://www.r-project.org/about.html
[3] https://www.r-project.org/contributors.html
[4] https://cran.r-project.org/web/packages/kamila/index.html
[5] https://cran.r-project.org/web/packages/ggplot2/index.html
[6] https://cran.r-project.org/web/packages/cluster/index.html
[7] https://cran.r-project.org/web/packages/factoextra/index.html

Kamila paketi hiyerarşik olmayan kümeleme algoritmalarından k-ortalamalar ile büyük ve karma yapıdaki veri setlerinin kümeleme analizine olanak sağlamaktadır. Kamila algoritması R Studio ortamında Alexander Foss ve Marianthi Markatou tarafından geliştirilmiştir.

ggplot2 paketi grafik dilbilgisini kullanarak detaylı ve çok çeşitli görselleştirmelerin oluşturulması sağlamaktadır. Bu çalışmada tahmin gücü grafiklerininin oluşturulması aşamasında kullanılmıştır.

Factoextra paketi çok değişkenli veri analizlerinin çıktısını alarak görselleştirme sağlamaktadır. Bu çalışmada küme grafiklerini oluşturmak için kullanılmıştır.

R Studio açık kaynaklı ücretsiz bir programlama dili olması, istatistiksel analizlerin kolaylıkla yapılabilmesine olanak sağlaması ve büyük veri setleri ile kolaylıkla çalışılabilmesi sebebiyle son dönemlerde yaygın olarak kullanılmaya başlanmıştır.

R programında hiyerarşik kümeleme algoritmalarını kullanarak çok değişkenli gözlemleri kümelemek için, **stats** paketinde yer alan ***hclust*** fonksiyonundan veya **Cluster** paketinde bulunan ***agnes*** fonksiyonundan yararlanılabilir (Bulut 2018:379). ***hclus*t** fonksiyonu gözlem birimleri arasında uzaklık ölçüleri yardımıyla hesaplanan **D** (farklılıklar/ benzemezlik) matrisini kullanarak hiyerarşik kümeleme analizinin yapılmasını sağlamaktadır. Kullanılan kümeleme teknikleri "single" tek bağlantı, "complete" tam bağlantı, "average" ortalama bağlantı "centroid", merkez (küresel) ortalama, "median" medyan (ortanca) bağlantı, "ward.d" ve "ward.D2" olmak üzere iki farklı ward kümeleme yaklaşımıdır. **Cluster** paketinde yer alan ***agnes*** (Agglomerative Nesting Hierarchical Clustering/ Birleştirici Hiyerarşik Kümeleme) fonksiyonunda kullanılan uzaklık ölçüleri "öklid" ve "manhattan" uzaklıklarıdır. **hclust** fonksiyonuna benzer olarak **agnes** fonksiyonu da gözlem birimleri arasında uzaklık ölçüleri yardımıyla hesaplanan D uzaklık matrisi aracılığıyla oluşturulan benzemezlik (farklılıklar) matrisini kullanarak hiyerarşik kümeleme analizinin yapılmasını sağlamaktadır. **Cluster** paketinde yer alan **diana** (DIvisive ANAlysis Clustering/ Ayırıcı Kümeleme Analizi) fonksiyonu ise uzaklık ölçüsü olarak "Öklid" ve "Manhattan City Block" uzaklıklarını kullanmaktadır ve benzemezlik matrisi yardımıyla ayırıcı hiyerarşik kümeleme analizi yapılmasını sağlamaktadır.

Hiyerarşik olmayan kümeleme yöntemlerinden k- ortalamalar yöntemi ile yapılacak hiyerarşik olmayan kümeleme çalışmasında **stats** paketinde yer alan ***kmeans*** fonksiyonundan yararlanılmaktadır

Hiyerarşik olmayan kümeleme algoritmalarından bir diğeri **Cluster** paketi içerisindeki ***PAM*** (Partioning Around Medoids/ Medoidler/ Medoidler Etrafindan Kümeleme)'dir. ***PAM*** algoritması, gözlem birimleri arasından k tane temsili birimin veya medoidin aranmasına dayanmaktadır. Kümedeki en merkezi nokta olan temsilciye medoid denir. Algoritma tekrar tekrar çalışarak medoidlerin en iyi

seçimini yapmaya çalışır. Her gözlem birimi en yakın medoide atanarak kümeler oluşturulur. Amaç, benzemezliklerin toplamını en aza indiren k tane temsili nesneyi bulmaktır. ***PAM*** uzaklık ölçüsü olarak "öklid" ve "manhattan" uzaklıklarını kullanmaktadır ve benzemezlik matrisi yardımıyla ayırıcı hiyerarşik kümeleme analizi yapılmasını sağlamaktadır.

Cluster paketi içeraisinde yer alan diğer bir algortima ***clara***'dır. ***Clara*** (Clustering Large Applications/ Büyük Uygulamaların Kümelenmesi) veri kümesinin temsilcisi olarak küçük bir kısmının seçildiği örneklemeye ve medoidlerin ***PAM*** algoritması kullanılarak seçilen bu örneklemden belirlenmesine dayanır. Alt veri kümesinden k temsili medoid seçildikten sonra, verinin tümüne ait her bir gözlem birimi en yakın medoide atanır. ***Clara*** algoritması ***PAM***'e kıyasla daha büyük veri setlerinde kullanılabilmektedir. Burada, sabit boyutlu örneklem alt veri kümelerini dikkate alması ve bu yolla zaman ve depolama gereksinimlerinin azalması rol almaktadır.[8] (Reference Manuel: Cluster, s:14-15).

Hiyerarşik olmayan kümeleme yöntemlerinden bir diğeri **clustMixType** paketi içerisindeki ***k-proto*** algoritmasıdır. Sayısal ve kategorik değişkenler içeren karma tipte veri setleri için, k-ortalamalar algoritmasının uzantısı olarak geliştirilmiş ayırıcı kümeleme yöntemidir. Algoritma, k-ortalamalar algoritmasına benzer olarak, küme prototiplerini tekrarlı olarak hesaplamakta ve küme atamalarını gerçekleştirmektedir. Küme prototipleri, sayısal değişkenler için küme ortalaması ve kategorik değişkenler için modları kullanmaktadır.

Hiyerarşik ve hiyerarşik olmayan kümeleme yöntemleri, gözlemler arası uzaklıklara dayanmaktadır. Farklı olarak, karma çok değişkenli normal dağılım modeline dayalı **EM** (Expectation Maximization) algoritması mevcuttur. Modele dayalı yaklaşım verinin karma bir dağılımdan geldiğini varsayar (Bulut 2018:394). **Fpc** paketinde yer alan ***flexmixedruns*** (Fitting mixed Gaussian/multinomial mixtures with flexmix) algoritması sürekli değişkenlerin Gauss dağılımlarıyla modellendiğini ve kategorik değişkenlerin bağımsız çok terimli dağılımlarla modellendiği gizli bir sınıf karışımı (kümeleme) modeline uyduğunu varsaymaktadır. Uyum, EM algoritması ile hesaplanan maksimum olabilirlik tahminiyle yapılmaktadır. [9]

ClustMD paketinin içerisinde bulunan ***ClustMD*** (Model Based Clustering for Mixed Data) algoritması karma tipteki veriler için model tabanlı bir kümeleme yöntemidir ve gizli değişken modeli kullanılarak geliştirilmiştir. Veri sürekli, ikili, sıralı veya nominal değişkenlerden olabilmektedir. Uyum, EM algoritması ile hesaplanan maksimum olabilirlik tahminiyle yapılmaktadır (McParland, Gormley, 2015:1).

[8] https://cran.r-project.org/web/packages/cluster/index.html
[9] https://www.rdocumentation.org/packages/fpc/versions/2.2-9/topics/flexmixedruns

Mclust paketi içerisinde yer alan ***mclust*** (Model-Based Clustering) algoritması parametreleştirilmiş sonlu Gauss karışım modellerine dayalı bir kümeleme algoritmasıdır.[10] (Reference Manuel: mclust, s:77). Modeller ve küme sayısı, hiyerarşik birleştirici kümeleme yöntemi ile belirlenmektedir. Optimum model BIC kriterine göre seçilmektedir. Belirlenen küme sayısı ve kümeleme modeline göre gözlem birimleri EM algoritması ile kümelenmektedir (Bulut 2018: 397).

Bulanık kümelemede, birimlerin kümelere hangi üyelik derecesi ile atandığını belirlemek temel amaçtır. Veri noktası ile küme merkezlerinin arasındaki uzaklığın hesaplanması ile üyelik dereceleri elde edilmektedir (Camkıran 2017:45). Bu derecelere üyelik olasılıkları denmekte ve bir gözlemin tüm kümeler için küme üyelik olasılıkları toplamı 1'e eşit olmaktadır (Bulut 2018:410). Bulanık kümeleme analizinde kullanılan algoritmalardan biri **cluster** paketi içerisindeki ***fanny*** (Fuzzy Analysis Clustering/ Bulanık Kümeleme Analizi) algoritmasıdır. Bulanık kümelemede, her gözlem birimi çeşitli kümelere "yayılır", i'inci gözlem biriminin v'inci kümeye aidiyeti u_{iv} ile gösterilmektedir. Üyeliklerin değeri negatif değer almamakta ve bir gözlem birimi için toplamı 1'e eşit olmaktadır (Reference Manuel: Cluster, s:39).

DBSCAN (Density-based Spatial Clustering of Applications with Noise), rastgele şekle sahip kümeleri keşfetmek için tasarlanmış, yoğunluğa dayalı bir kümeleme algoritmasıdır. **DBSCAN** paketindeki, ***DBSCAN*** algoritması Ester ve diğerleri (1996) tarafından açıklanan orijinal algoritmayı takip etmektedir (Reference Manuel: dbscan, s:5-6).[11] Algoritma, rastgele belirlenen bir noktasından hareketine başlamaktadır. Algoritmanın iki önemli belirleyicisi mevcuttur. Bunlar gözlemin etrafında tanımlanacak olan çemberin yarıçapını ifade eden ε ("eps") ve bir kümede en az kaç gözlem olacağını ifade eden "MinPts" değerleridir (Bulut 2018: 412). **DBSCAN** algoritması rastgele seçilen bir gözlemin merkezde bulunduğu eps çapında bir daire içerisinde, MinPts değeri ile belirlenen sayıda gözlem içerir. Eğer kümedeki gözlem sayısı MinPts değerinden küçük ise, kümelemenin merkezi diğer gözlemler üzerine taşınmakta ve algoritma tüm gözlemler kümelenene kadar devam etmektedir. Bazı gözlem birimleri herhangi bir kümeye ait olmayabilir ve bu gözlem birimleri gürültü (noise) olarak adlandırılmaktadır. **DBSCAN** algoritması **fpc** (Flexible Procedures for Clustering) paketi içerisinde de mevcuttur. DBSCAN paketindeki uygulama, fpc paketindeki uygulamaya göre önemli ölçüde daha hızlıdır (Reference Manuel: dbscan, s:5-6).

Kümeleme Analizi Yöntemlerinin R uygulamaları başlığı altında bilgileri verilen algoritmaların özeti Tablo2.3'te bulunabilir.

[10] https://cran.r-project.org/web/packages/mclust/index.html
[11] https://cran.r-project.org/web/packages/dbscan/index.html

Tablo 2.3: Yaygın Kullanılan R Kümeleme Algoritmaları

Kümeleme Yöntemi		R Paketi	Komut
Aşamalı Kümeleme Yöntemleri	Tek Bağlantı (Single Linkage)	stats	hclust(d,method = "single")
	Tam Bağlantı (Complete Linkage)	stats	hclust(d,method = "complete")
	Ortalama Bağlantı (Average Linkage)	stats	hclust(d,method = "average")
	Merkez (Centroid)	stats	hclust(d,method = "centroid")
	Meydan (Median)	stats	hclust(d,method = "median")
	Ward (Ward)	stats	hclust(d,method = "ward.D") hclust(d,method = "ward.D2")
Aşamalı Olmayan Kümeleme Yöntemleri	k-Ortalamalar(k-means)	stats	kmeans(veri,k)
	Medoid (PAM)	cluster	pam(veri,k)
Modele Dayalı Kümeleme	EM (Expectation Maximization)	mclust	Mclust(veri,G=k, model)
Bulanık (Fuzzy) Kümeleme	Bulanık c ortalamalar	cluster	fanny(veri,k)
Yoğunluğa Dayalı Kümeleme	DBSCAN	fpc	dbscan(veri,eps,MinPts)

Makine öğrenimi bağlamında kümeleme yöntemleri, denetimli öğrenme ve denetimsiz öğrenmeye başlıkları altında toplanmıştır. K- ortalamalar yöntemi, makine öğrenmesi açısından değerlendirildiğinde, denetimsiz (unsupervised) bir tekniktir. Denetimsiz kümeleme tekniklerinde, gözlemler için önceden belirlenmiş etiketler söz konusu değildir. Bu yöntemde amaç veri seti içerisinde her bir gözlem biriminin, kendisine en yakın / benzer gözlem birimleri ile birlikte olacak şekilde k adet kümeye bölünmesidir.

3. KAMILA

Karma Tip Veri Kümeleme Yöntemi

Günümüzde pek çok alanda saklanabilen verinin miktarı artmakta ve bu verilerden anlamlı ve gelişime yönelik analizlerin yapılma yolları araştırılmaktadır. Farklı kaynaklardan beslenen büyük veri setlerinin her geçen gün daha yaygın hale gelmesi ve bilgisayarların hesaplama kapasitelerinin artması, sayısal ve kategorik değişkenlerin bir arada olduğu karma tipteki veri setlerinin analiz edilmesine duyulan ihtiyaç ve uygulama olanaklarını arttırmaktadır. Kümeleme algoritmaları ile ilgili yapılan literatür çalışmalarında çoğunluğun homojen yapıdaki, diğer bir anlatım ile tüm değişkenlerin sürekli ya da tüm değişkenlerin kategorik olduğu veri setleri için uygulama yapıldığı görülmektedir. Sürekli ve kategorik değişkenlerin bir arada olduğu karma tipteki veri setlerinde, araştırmacıların değişkenlerin türünü tek tipte olacak şekilde dönüştürüp homojen yapıda veri setleri ile analizlere devam etmeleri sıkça karşılaşılan bir durumdur.

Karma tipteki büyük veri setlerinin çoğalması, bu yapıdaki veri setleri için kullanılabilecek kümeleme algoritmaların geliştirilmesi ihtiyacını doğurmuştur (Foss vd. 2016: 419). KAMILA (KAy-means for Mixed Large data sets) algoritması, sağlam ve ölçeklenebilir bir yöntem olarak 2016 senesinde Foss, A., Markatou, M., Ray, B. ve Heching, A. tarafından geliştirilmiştir.

Kamila Kümeleme Algoritması, yaygın olarak kullanılan iki kümeleme algoritmasının en iyi özelliklerini birleştirmeyi hedeflemiştir. Bu algoritmalar k-ortalamalar (k-means) ve Gauss nominal çok değişkenli karma modeller (Gaussian multinominal mixture models) algoritmalarıdır. Çok değişkenli küresel dağılımların ortak yoğunluklarının değerlendirmesi için verimli bir yöntem olarak, çekirdek yoğunluğu tahmini kullanılmaktadır (Foss vd. 2016:429-430). Veri setinin, sonlu karma yoğunluk dağılımını izleyen, rastgele sürekli ve kategorik değişkenlerden oluştuğu varsayılmaktadır (Foss, Markatou 2018:6).

Bu bölümde çekirdek yoğunluk tahmini, çoklu nominal model, KAMILA algoritması ve tahmin gücü açıklanacaktır.

3.1. Çekirdek Yoğunluk Tahmini

Belli bir anakütleden alınan örneklem yardımıyla, anakütlenin parameter değeri (θ) olarak kabul edilecek bir sayı veya aralık belirlemeye tahmin denmektedir. Sözü edilen parametre, bir anakütleyi diğer anakütlelerden ayırmaya yarayan aritmetik ortalama, varyans, oran vb. ölçülerinin genel bir ifadesidir. Tahminci ($\hat{\theta}$) ise, eldeki örneklemlerden anakütle parametresinin nasıl

hesaplanacağını gösteren formül olarak tanımlanabilir (Altaş, 2013:1). Parametrik olmayan tahminin odak noktası parametrik tahminden farklıdır. Parametrik tahminde, verilen bir yoğunluk fonksiyonu f(x) için, θ'nın en iyi tahmincisi $\hat{\theta}$ elde etmeye vurgu yapılır. Parametrik olmayan durumda ise, yoğunluk fonksiyonu f(x)'in iyi bir tahmincisi olan $\hat{f}(x)$ fonksiyonun doğrudan elde edilmesi vurgulanır (Scott, 1992:33). Parametrik yaklaşımda verinin, parametrik yapısı bilinen bir anakütleden geldiği varsayılmaktadır.

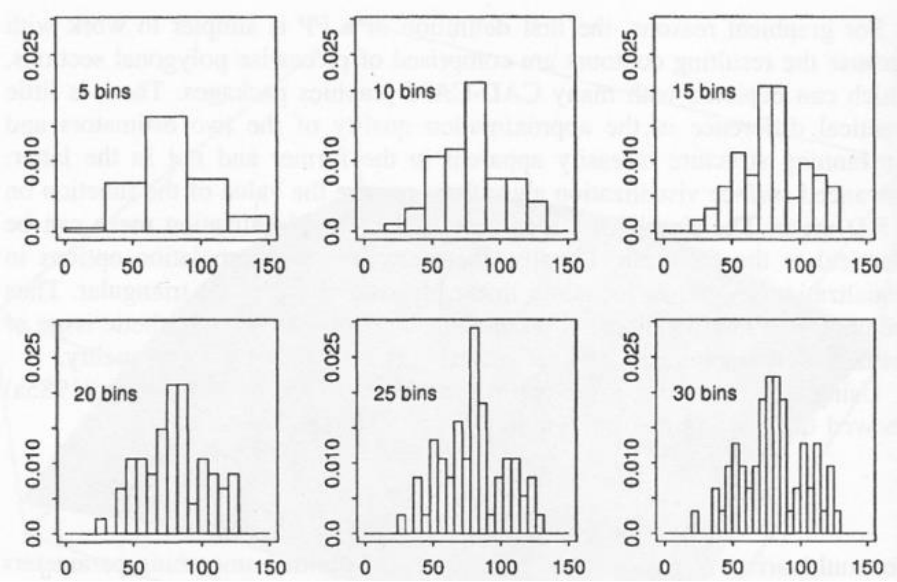
Yoğunluk fonksiyonu kestirimi basitçe dağılımı bilinmeyen bir veri seti için yoğunluk fonksiyonu oluşturulması problemi olarak tanımlanabilmektedir. Yoğunluk fonksiyonu kestirimi için parametrik ve parametrik olmayan yaklaşımlar mevcuttur. Parametrik olmayan yaklaşım sadece verilere dayanır ve "verinin kendi adına konuşmasına izin verir" (Erçelik 2019:1). Diğer bir anlatımla, parametrik olmayan eğriler verilerdeki yapı tarafından belirlenmektedir. Parametrik eğriler ise, veriye ait dağılım yapısının bilinmesini veya bilinen bir dağılımdan yola çıkılarak tahmin yapılması ilkesine dayanmaktadır.

İlgilenilen rassal değişken X'in sürekli olması durumunda olasılık yoğunluk fonksiyonu *f(x)* ile göstermekte, a ve b herhangi iki reel sayı olmak üzere *f(x)* fonksiyonunun belirlenmesi, X'in dağılımının tanımını ve X ile ilişkili olasılıkların aşağıdaki verilen formülden hareketle bulunmasını sağlamaktadır (Akay, Uyar 2017:54).

$$P\,(a < X < b) = \int_a^b f(x)dx \qquad \text{a < b için}$$

Parametrik olmayan yaklaşımda, anakütlenin olasılık dağılımının bilinen bir dağılıma uyduğu varsayımı yapılmamaktadır. Bu yönüyle parametrik olmayan tahmin yöntemleri, parametrik yöntemlere göre daha esnek tahmin imkanı sağlamaktadır.

Parametrik olmayan yöntemlerde gözlem birimlerinin oluşturduğu histogram, yoğunluk tahmincisinin resmi olarak tanımlanabilmektedir. Histogram çiziminde kullanılan bölünme (bin) değeri, çizilecek histogramın şeklinde etkili olmaktadır.



Şekil 3.1: Farklı Bölünme Değerleri İçin Oluşan Histogramlar

Kaynak: Scott, D.W. (1992), *Multivariate Density Estimation*, New York: John Wiley & Sons, s.110

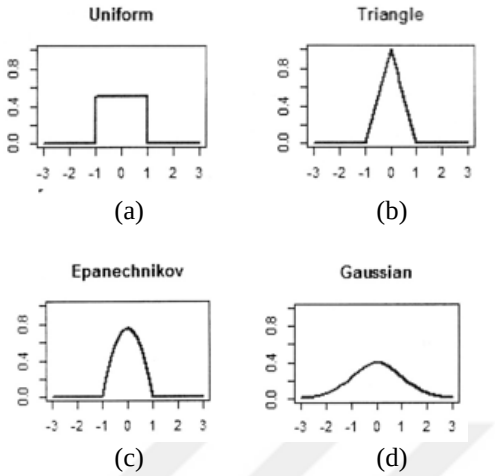
Şekil 3.1'de aynı veriye ait farklı bölünme değerleri (bin) için çizdirilen histogramlar verilmektedir. Bölünme sayısı arttıkça, grafik daha detaylı olmakta ve dağılımın şekli değişmektedir.

Çekirdek (kernel), parametrik olmayan tahmin yöntemlerinde kullanılan ağırlıklandırma fonksiyonudur ve K harfi ile gösterilmektedir. Yoğunluk fonksiyonunun çekirdek tahmininde belirlenen noktanın sağındaki ve solundaki gözlemlerin söz konusu noktaya olan uzaklıklarına göre eşit ağırlık vermenin daha uygun olması nedeniyle, kernel fonksiyonu genellikle simetrik bir yoğunluk fonksiyonu olarak tanımlanır (Çağlayan Akay, Kangallı Uyar 2017: 75). Çekirdek fonksiyonu pozitif, sürekli bir olasılık yoğunluk fonksiyonu olmakla birlikte, - ∞ ve+∞ arasındaki tüm olası değerlerinin toplamı 1'e eşittir. n gözlem biriminin oluşturduğu örneklem için yoğunluk fonksiyonu tahmincisi :

$$\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) \quad (1)$$

olarak yazılabilmektedir. Bu eşitlikte K çekirdek fonksiyonu ve h_n, histogramdaki bölme genişliğine benzer, pozitif değerler alan düzleştirme parametresidir (Kvam, Vidakovic 2007:208).

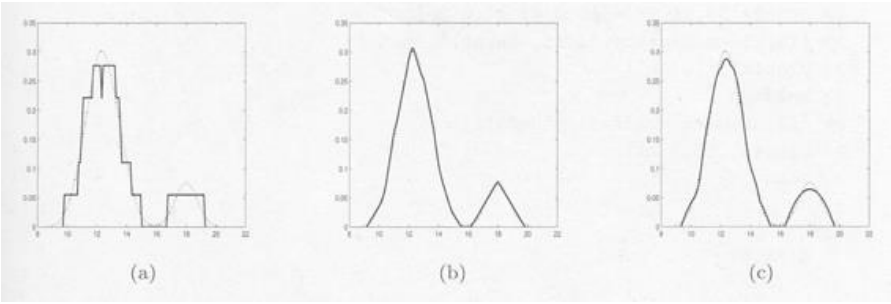
K değeri eğrinin şeklini belirlerken, h_n değeri çekirdeğin yayılımını kontrol etmektedir (Kvam, Vidakovic 2007:209). Şekil 3.2'de dört temel çekirdek fonksiyonu verilmiştir. Temel çekirdek fonksiyonları simetrik yapıdadır.



Şekil 3.2: (a) Kutu, (b) Üçgen, (c) Epanechnickov, (d) Normal Çekirdek Fonksiyonları

Kaynak: Çağlayan Akay, E., Kangallı Uyar, S. G. (2017), *R Uygulamalı Nonparametrik Ekonometri,* İstanbul: Der Yayınları, s.82

Şekil 3.3'te gözlem sayısı n=7 iken, yoğunluk tahmini için kullanılan üç çekirdeğin çizimleri verilmektedir. Gözlem birimleri aynı olduğu halde, seçilen çekirdeğin şekline göre yoğunluk tahminleri için farklı grafikler oluşmaktadır.



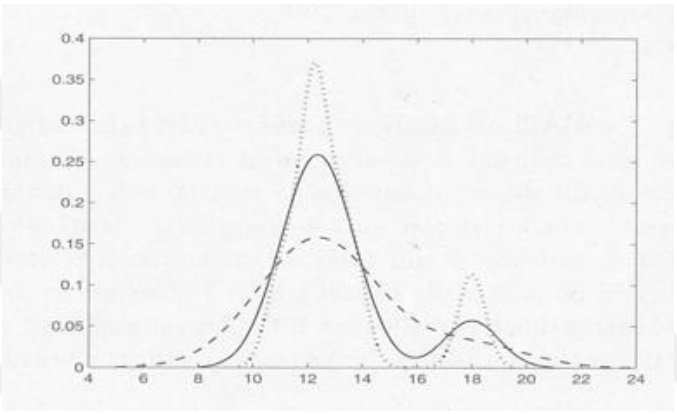
Şekil 3.3: (a) Kutu, (b) Üçgen, (c) Epanechnickov Çekirdekleri için Yoğunluk Tahmini

Kaynak: Kvam, P.H, Vidakovic, B. (2007), *Nonparametric Statistics with Applications to Science and Engineering,* New Jersey: John Wiley & Sons, s.211

Yoğunluk fonksiyonu tahminlerinde önemli bir adım, bant genişliğinin seçimi olmaktadır. Farklı bant genişlikleri ile yoğunluk fonksiyonu elde etme fikrinin temeli, gözlemleri uygun bir eğriye uydurmak ve buna göre elde edilmiş fonksiyonları kullanmaktır.

Bant genişliği, diğer adıyla düzgünleştirme parametresinin seçimi tahminlerin sapma ve varyansı arasındaki dengeyi sağlaması sebebiyle, çekirdek seçiminden daha önemli olduğu düşünülmektedir.

Aynı gözlem birimleri ve çekirdek seçimi için, farklı bant genişliği değerleri kullanıldığında, elde edilen yoğunluk tahminleri değişiklik göstermektedir. Şekil 3.4'te n=7 iken, seçilen normal çekirdeğin farklı bant genişlikleri için ortaya çıkardığı yoğunluk tahminleri arasındaki farklılık gözlenebilmektedir.



Şekil 3.4: Gözlem sayısı n=7 iken, farklı bant genişliklerinde yoğunluk tahminleri

Kaynak: Kvam, P.H, Vidakovic, B. (2007), *Nonparametric Statistics with Applications to Science and Engineering,* New Jersey: John Wiley & Sons, s.211

Düzgünleştirme parametresi çok büyük seçildiğinde elde edilen sonuçlarda varyans azalmasına rağmen sapma artmaktadır (Çağlayan Akay, Kangallı Uyar 2017: 74).

Çekirdek tahmincisi,

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) = \frac{1}{nh} \sum_{i=1}^{n} K(\psi_i) \qquad (2)$$

olarak da tanımlanmaktadır. Bu formülde $\psi_i = (\frac{x_i - x}{h})$ olmakta ve ağırlık fonksiyonu olarak isimlendirilmektedir (Çağlayan Akay, Kangallı Uyar 2017: 129).

Veri setinde bulunan her bir gözlem birimi için, gözlem biriminin çekirdekten uzaklığına göre bir ağırlık verilmektedir. Verilen ağırlıklar, çekirdekten uzaklaştıkça küçük değerler almakta, çekirdeğe yaklaştıkça büyük değerler almaktadır.

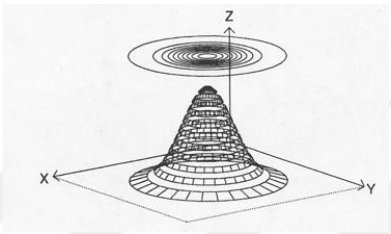
Çekirdek yoğunluk fonksiyonunun tahmincisi,

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} (\omega_i) \quad (3)$$

$$\omega_i = \omega_{ni}(x) = \frac{1}{h} K\left(\frac{x_i - x}{h}\right) \quad (4)$$

olarak da yazılabilmektedir (Çağlayan Akay, Kangallı Uyar 2017: 130).

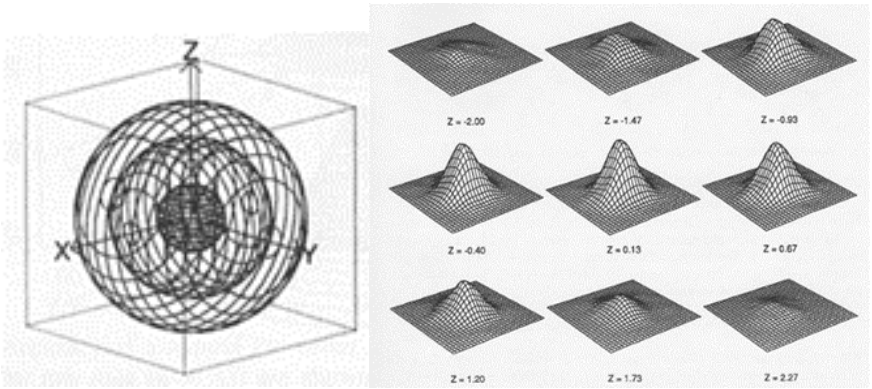
Bant genişliği (bandwidth) h, düzgünleştirme parametresi (smoothing parameter) veya pencere genişliği (window- width) olarak da anılabilmektedir.



Şekil 3.5: İki Değişkenli Veriye Ait Histogram ve Normal Çekirdek Fonksiyonu ile Oluşturulan Kontür Grafiği

Kaynak: Scott, D.W. (1992), *Multivariate Density Estimation*, New York: John Wiley & Sons, s.21

Şekil 3.5' te normal çekirdek fonksiyonu ile görselleştirilen, iki değişkenli veriye ait kontür grafiği görseli verilmektedir. *x* ve *y* eksenlerinde görselleştirilen çekirdek fonksiyonu, *z* ekseni de eklenerek, kontür grafiğini oluşturmaktadır.

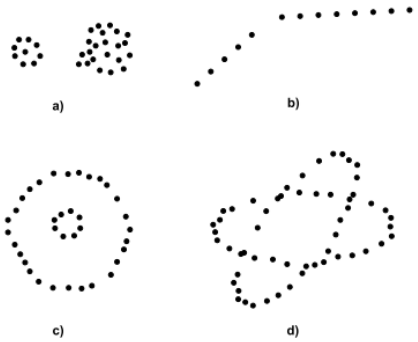


(a) (b)

Şekil 3.6: Üç Değişkenli Veriye Ait Kontür ve Histogram Dilimleri Görselleri

Kaynak: Scott, D.W. (1992), *Multivariate Density Estimation*, New York: John Wiley & Sons, s.22-23

Şekil 3.6’ da (a) normal dağılıma sahip üç değişkenli veriye ait kontür grafiği görseli verilmektedir. (b) sekmesinde üç değişkenli normal dağılıma sahip veri setine ait standartlaştırılmış z değerleri için histogram dilimleri dizisi görselleri yer almaktadır.



Şekil 3.7: Çeşitli Küme Şekilleri

Kaynak: Abonyi, Feil (2007), *Cluster Analysis for Data Mining and System Identification*, Germany: Birkhauser Verlag AG, s.2

Şekil 3.7’de farklı şekillerde küme görselleri yer almaktadır. Kümeler (a) küresel, (b) ince uzun (elognated) veya doğrusal, (c) elips şeklinde veya (d) yüksek boyutlu analoglar şeklinde olabilmektedir (Abony, Feil 2007:1). Algoritmalar, verinin doğal yapısında bulunan grupları tespit etmek üzere kurgulanmaktadır.

3.2. Çoklu Nominal Model (Multinominal Model)

Yaş, gelir, süre gibi metrik ölçüme sahip reel sayılar kümesinde değerler alabilen sürekli değişkenler için; ortalama, varyans, medyan ve çeyrek değerleri hesaplanarak elde edilebilmektedir. Bu değerler değişkenin özellikleri konusunda araştırmacıya bilgi sağlamaktadır. Çoklu nominal değişkenlerde ise, değişkenler evet/ hayır, kadın/ erkek, küçük/ orta / büyük gibi analiz öncesi sınıflandırılmış değerlerden birini almaktadır. Değişkene ait cevap evet/ hayır gibi iki seçenekten oluşuyorsa ikili (binomial), küçük/ orta / büyük gibi ikiden fazla seçeneğe sahipse, çoklu nominal (multinominal) olarak isimlendirilmektedir. Çoklu nominal modelde amaç, veriyi tanımlamak ve anlam çıkarabilmek amacıyla olasılık yoğunluk fonksiyonları yardımıyla parametrelerin belirlenmesine dayanmaktadır. Bu çalışmada hem ikili (bir cihaza sahip olup olmama durumunu ölçen değişkenler) hem de çoklu nominal (eğitim durumu, meslekler gibi) değişkenler yer almaktadır.

İki seçenekli cevaba sahip bir değişken için, değişkeni iki seviyeli olarak isimlendirmek mümkündür. Böyle bir durumda birinci seçeneğin olma ihtimali ile ikinci seçeneğin olma ihtimallerinin toplamı bire eşit olmaktadır. Benzer şekilde ikiden fazla yanıt seçeneğine sahip bir değişken için, cevapların oluşma ihtimalinin toplamı yine bire eşit olmaktadır. Matematiksel olarak ifade edildiğinde,

$$\sum_{j=1}^{c} p_j = 1$$

x = 0, 1, …, m — m sayıda bağımsız deneme sayısı,

j = 1,…, c — incelenen j olayının (j değişkeninin) gözlenen c sayıda sonucu olduğunu göstermek üzere,

θ — j olayının bilinen gerçekleşme oranını belirtmektedir.

j olayının gerçekleşme olasılığı 0 ile 1 arasında bir değer almakta ve $0< p_j < 1$ olarak gösterilmektedir. Bir değişken için olası tüm cevapların veya sonuçların olma ihtimalinin toplamının bire eşit olması durumu aşağıdaki gibi ifade edilmektedir.

$$P(x; p, \theta) = \text{prob}\,(X_i = x) = \sum_{j=1}^{c} p_j \; P_j \; (x; \theta_j) \quad (5)$$

$$Pj(x; \theta_j) = \binom{m}{x} \theta_j^x \; (1 - \theta j)^{m-x} \quad (6)$$

5 numaralı eşitlik ile verilen ifade, *p* ve θ olmak üzere iki parametreye sahip olasılık yoğunluk fonksiyonudur (Everitt, Hand 1981:89). Bu fonksiyon yardımıyla yoğunluk grafiği oluşturulabilmektedir.

Her biri farklı bir olasılık yoğunluk fonksiyonuna sahip değişkenlerden oluşan ve sonlu karma yoğunluk (finite mixture density) olarak bilinen çıktı ile sonuçlanan model, değişkenlerin hepsinin beraber değerlendirildiği durumda verinin içinde doğal olarak bulunan gruplar olduğunu varsayar. Sonlu karma yoğunluk modeli seçildiğinde yapılacak analiz, kümeleme analizi parametrelerinin tahmin edilmesi olarak ele alınmaktadır. Sonlu karma yoğunluk modeli ile yapılan kümeleme analizi, model tabanlı kümeleme analizi olarak da anılmaktadır (Everitt, Hothorn 2011:185-186). Model tabanlı kümeleme algoritmaları, güçlü ayırıcı kümeleme yaklaşımlarındandır. Verilerin temeldeki olasılık dağılımlarının karışımından oluştuğu varsayımıyla, olasılıksal bir yaklaşım kullanarak, gözlemlenen veriler ile matematiksel modeler arasındaki uyumu optimize etmeye çalışır.

Verileri temsil etmek için birçok karışım modeli belirlenebilir ancak Gauss karışım modeli en yaygın kullanılan temsil modelidir (Jiao vd. 2022:1). Gauss karışım modellerinde, veriler karma modelleme yaklaşımı ile kümelenmektedir (Tang, Browne, McNicholas 2015:84). K-ortalamalar algoritması kümeleri daireler şeklinde sınıflama eğiliminde iken, Gauss karışım modelleri kümelerin beraber değişimlerini daha esnek bir şekilde gözönüne almaktadır. K-ortalamalar algoritmasında, her bir gözlem biriminin küme merkezlerine uzaklığı göz önünde bulundurulurken; Gauss karışım modellerinde, her gözlem biriminin her küme için hesaplanan kümeye aidiyet olasılığı dikkate alınmaktadır (Wang vd. 2021:95).

Gauss Karışım Modelleri (Gaussian Mixture Models) için maksimum olasılık tahmini yapmak amacıyla kullanılan geleneksel algoritma beklenti maksimizasyonu (EM- expectation maximization) olmaktadır (Arı 2013:101).

3.3. Kamila Algoritması

KAMILA algoritması sürekli ve kategorik tipteki verilerin bir arada olduğu büyük veri setlerinin kümeleme analizi için geliştirilmiştir.

Uygulama yapılacak veri setine ait örnek gösterim Şekil 3.8'de verilmektedir. Örnek gösterime göre, veri seti n sayıda gözleme sahip, P sayıda sayısal değişken ve Q sayıda kategorik değişken içermektedir.

	Sürekli Değişkenler						Kategorik Değişkenler					
Gözlemler	V_1	V_2	.	.	.	V_P	W_1	W_2	.	.	.	W_Q
1												
2												
3												
.												
.												
.												
i												
.												
.												
.												
.												
n												

Şekil 3.8: Kamila Algoritması Uygulanacak Veri Setinin Örnek Gösterimi

Kamila algoritması ile ilgili olarak, veri setinin n adet bağımsız gözlem ve (P+Q) boyutlu, aynı dağılıma sahip değişkenlerden ve G sayıda kümeden oluştuğu varsayılmaktadır. Bu varsayımda,

N Gözlem sayısı,

P Sürekli değişkenlerin sayısı,

Q Kategorik değişlenlerin sayısı,

V Sürekli rassal değişkenlerin P boyutlu vektörü,

W Kategorik rassal değişkenlerin Q boyutlu vektörü,

Lq q = 1,2,..,Q olacak şekilde q. kategorik değişkene ait seviye (level) sayısıdır.

Herhangi bir küme için sürekli ve kategorik değişkenlerin bağımsız oldukları varsayılmaktadır.

g'inci kümenin sınıflandırmasında kullanılan birim bileşen yoğunluk fonksiyonu,

$f_{V,g}(v; \mu_g, \sum_g)$

şeklinde modellenmektedir. Bu modelde,

g Küme üyeliğine ait indeks numarası (küme indeksi),

μ_g g'inci kümenin merkezi,

$\sum_g$ g'inci ağırlıklandırma matrisidir.

g'inci kümenin sınıflandırmasında kullanılan, birim bileşen olasılık kütle fonksiyonu,

$$f_{\mathbf{w},g}(\mathbf{w}) = \prod_{q=1}^{Q} m\left(w_q;\ \theta_{gq}\right) \quad (7)$$

şeklinde modellenmiştir. Bu modelde,

m(;) multinominal olasılık kütle fonksiyonu,

θ_{qg}, q'uncu kategorik değişken için g'inci bileşene ait multinominal parametre vektörüdür.

g'inci küme için yerel bağımsızlık varsayımı altında, ortak yoğunluk fonksiyonu,

$$f_{\boldsymbol{V},\boldsymbol{W},g}\left(\boldsymbol{v},\boldsymbol{w};\ \mu_g,\Sigma_g,\theta_{gq}\right) = f_{\boldsymbol{V},g}\left(\boldsymbol{v};\ \mu_g,\Sigma_g\right)\prod_{q=1}^{Q} m\left(w_q;\ \theta_{gq}\right) \quad (8)$$

şeklinde olmaktadır (Foss, Markatou 2018: 6-7).

KAMILA bilinmeyen parametreleri EM algoritmasina benzer yinelemeli bir süreçle tamhin edilmektedir. t'inci iterasyonda $\hat{\mu}_g^{(t)}$, g'inci kümenin merkezinin ve $\hat{\theta}_{gq}^{(t)}$, g'inci kümenin q'uncu ayrık rastgele değişkenin multinominal parameter tahmincisidir. Yinelemeli tahmin prosedürü, bölünme ve tahmin olmak üzere iki geniş adımdan oluşmaktadır. Bölünme adımı, her gözlemin bir kümeye atanmasını ve tahmin adımı, yeni kümeyi kullanarak ilgilenilen parametreleri yeniden tahmin etmektedir. .t'inci iterasyondaki $\hat{\mu}_g^{(t)}$ve $\hat{\theta}_{gq}^{(t)}$ için, i'inci gözlem biriminin her $\hat{\mu}_g^{(t)}$' lerden öklid uzaklığı ($d_{ig}^{(t)}$) hesaplanır ve $r_i^{(t)}$= $\min_g(d_{ig}^{(t)})$ belirlenmek suretiyle, en küçük uzaklık mesafesi gözönüne alınarak küme ataması gerçekleştirilir. En küçük mesafenin hesaplanması için kullanılan çekirdek yoğunluk fonksiyonu aşağıda verilmektedir.

$$\hat{f}_R^{(t)}(r) = \frac{1}{Nh^{(t)}}\sum_{l=1}^{n} k\left(\frac{r - r_l^{(t)}}{h^{(t)}}\right) \quad (10)$$

Bu eşitlikte k, çekirdek fonksiyonu ve $h^{(t)}$, t'inci iterasyondaki bant genişliği olmaktadır. $\hat{f}_R^{(t)}$ fonksiyonu, veri setindeki sürekli değişkenler için $\hat{f}_{\boldsymbol{V}}^{(t)}$ fonksiyonunun oluşturulmasında kullanılmaktadır. Q kategorik ve bağımsız değişken için, i'inci gözlemin g'inci kümede olma olasılığı

$$c_{ig}^{(t)} = \Pi_{q=1}^{Q} m(w_{iq;} \hat{\theta}_{gq}^{(t)}) \quad (11)$$

eşitliği ile hesaplanabilmektedir. Burada m(;) multinominal olasılık kütle fonksiyonudur (Foss, Markatou 2018:7).

$$H_{,}^{(t)}(g) = \log[\hat{f}_V^{(t)}(d_{ig}^{(t)})] + \log[c_{ig}^{(t)}] \quad (12)$$

i'inci gözlem birimi hesaplanana $H_{,}^{(t)}$ (g) değerini maximize eden g'inci kümeye atanmaktadır (Foss vd. 2016:432).

Bağımsız olayların birlikte olma olasılıkları, her olayın ayrı ayrı meydana gelme olasılıklarının çarpımına eşit olduğundan, $H_{,}^{(t)}$ (g)' nin alacağı değerin hesaplanmasında matematiksel olarak log (a . b) = log a + log b dönüşümü kullanılmıştır.

Beklenti maksimizasyonu (EM, Expectation Maximization) algoritması, maksimum olabilirlik (maximum likelihood) tahmini yapmak için kullanılan popular bir algoritmadır (Arı 2013:51). Algoritma her yenilemesinde iki adım gerçekleştirir. E (expectation) adımında, gözlem birimlerini içeren uygun bir fonksiyonun öngürülmesi ve M (maximization) adımında bu fonksiyonun maximize edilmesidir (Kvam, Vidakovic 2007:307).

Olabilirlik fonksiyonlarının (likelihood functions) doğası, verileri sabit tutarken farklı parametre değerleri için fonksiyonun nasıl değiştiğininin gözlemlenmesidir. Daha büyük olasılık değerleri, veriler tarafından nispeten daha iyi desteklenen parametre değerlerine karşılık gelmektedir. Farklı parameter değerleri için olasılıkların görece büyüklüklerinin değerlendirilmesi önem taşımaktadır. Olabilirlik, olasılık değildir; bu sebeple tüm değerlerinin toplamının 1'e eşit olması beklenmemektedir (Bilder, Loughin 2015:496-497).

KAMILA algoritması, sürekli değişkenler için k-ortalamalar algoritmasında olduğu gibi, parametrik varsayımlar gerektirmemektedir. Gauss karışım modellerinde olduğu gibi, KAMILA algoritması sürekli ve kategorik değişkenlerin katkılarını ağırlık belirtmeden başarıyla dengeleyebilmektedir. Bununla birlikte Gauss karışım modellerinin varsayımını gevşeterek, uygun olan yoğunluk tahmincisini veriden hesaplanan yoğunluk tahmincisine dayandırmaktadır (Foss vd. 2016: 429).

KAMILA algoritması mevcut yöntemlerler ile karşılaştırıldığında, dört sebepten dolayı farklılaşmaktadır. Bunlardan ilki, değişkenler sahip oldukları ölçeklerde kullanılabilmekte ve dönüştürme işlemine ihtiyaç duyulmamaktadır. Verinin dönüştürülmesi sebebiyle oluşan bilgi kaybı bu yolla önlenebilmektedir. İkinci olarak, sürekli ve kategorik değişkenlere eşit ağırlık verilmesini, bu şekilde değişkenlerin eşit etkiye sahip olmasını sağlanmaktadır. Üçüncü olarak, algoritma sınırlayıcı

parametrik varsayımları barındırmamakta, kümelerin biçimini geniş eliptik formda dağılacak şekilde genelleyebilme ve son olarak algoritmanın uygulanmasında, kümelemeye girecek değişkenler için araştırmacı tarafından belirlenecek ağırlıklandırmaya ihtiyaç duyulmamaktadır.

Kategorik değişkenler için kukla değişken kullanılarak dönüşüm sıklıkla başvurulan bir yöntemdir. Kukla değişken kullanımı, verinin boyutunu arttırmakta ve bu durum kategorik değişkenlerin sayısı ve seviyeleri arttığında, problemlere sebebiyet verebilmektedir.

KAMILA algoritması ile kümeleme analizi yapılırken, değişkenlerin ağırlıklandırılması gözlemlerin küme içi dağılımının en az, kümeler arası dağılımının en fazla olması eş zamanlı sağlanacak şekilde yapılmaktadır. Ağırlıklandırmalar, küme içi dağılım ile kümeler arası dağılımın oranının, sürekli ve kategorik değişkenler için ayrı ayrı hesaplanması ve elde edilen oranların test edilerek, sürekli ve kategorik değişkenlerin dağılım oranlarının çarpımının en az olacak şekilde belirlenmesi yoluyla belirlenmektedir.

KAMILA algoritmasının çalışma şekli, sürekli ve kategorik değişkenlerin araştırmacı tarafından ağırlıklandırmasına gerek duyulmayan ve k- ortalamalar kümeleme yönteminin semiparametrik bir genellemesi olarak belirtilebilir.

KAMILA algoritmasının R Studio ortamı için geliştirilen paketinde (package 'kamila') listelenen ve bu çalışmada kullanılan argümanları aşağıdaki şekilde açıklanabilmektedir:

conVar Sürekli değişkenler için veri çerçevesi

catFactor Kategorik değişkenler için veri çerçevesi. Analiz öncesi kategorik değişkenlerin faktör olarak tanımlanması gerekmektedir.

numClust algoritma tarafından hesaplanan küme sayısı

numInit belirlenen iterasyon sayısı

maxIter her çalışma için belirlenen en fazla iterasyon sayısı

calcNumClust küme sayısını belirlemek için seçilen method. Bu çalışmada "ps – prediction strength" metodu kullanılmıştır.

numPredStrCvRun prediction strength /tahmin gücü methodu seçildiğinde kullanılmaktadır.

predStrThresh tahmin gücü methodu için eşik değeri.

KAMILA algoritması sürekli değişkenler için çekirdek yoğunluğu tahmin tekniği kullanırken, kategorik değişkenler için multinominal model kullanmaktadır.

Küresel simetrik dağılıma sahip ve merkezi μ olan radial çekirdek yoğunluk tahmincisi için,

$$f_x\,(\mathrm{x}) = \frac{f_R\ (r)\,\Gamma\,(\frac{p}{2}+1)}{pr^{p-1}\pi^{p/2}} \tag{9}$$

olarak yazılabilmektedir. Burada r = $\sqrt{(x-\mu)^T(x-\mu)}$ ve R = $\sqrt{(X-\mu)^T(X-\mu)}$ olmaktadır. r (yarıçap), [0, ∞) değerlerini alabilmektedir. $\hat{f}_R$, R'nin olasılık yoğunluk fonksiyonudur (Foss, Markatou 2018:7).

3.4. Tahmin gücü (Prediction Strength)

Kamila algoritması kullanılarak yapılan uygulamada, küme sayısı tahmin gücü algoritması kullanılarak belirlenmektedir. Tahmin gücü algoritması Tibshirani ve Walther (2005) tarafından geliştirilmiştir. Algoritmanın temel fikri, kümelemenin denetimli sınıflandırma problemi olarak değerlendirilmesi ve gerçek sınıf etiketlerinin tahmin edilmesinin sağlanmasıdır. Elde edilen tahmin gücü değeri, verilerin kaç gruba ayrılabileceğini ve yapılan ayrımın ne kadar iyi olduğunu anlamak için değerlendirilmektedir. Yanıt değişkeni olmadan verinin yapısını elde etmeye olanak sağlayan kümeleme analizi, bu yönüyle denetimsiz makine öğrenmesi alanında kullanılan önemli bir araçtır. Kümeleme analizinde, kullanılan yöntem kümeleme sonucunu değiştirebilmekte, bu sebeple en doğru veya en iyi ayrılmış olarak tanımlanabilecek kümelerden bahsetmek mümkün olmamaktadır. Kümeleme analizinin temel zorluğu olan uygun küme sayısının belirlenmesi, kullanılan pek çok yöntemde, küme içi dağılımın en az olacak şekilde elde edilmesi hedeflenmektedir. Geliştirilen tahmin gücü algoritmasında, küme sayısının tahmini bir model seçim problemi olarak ele alınmaktadır. Küme içi hata kareleri toplamı yerine tahmin hatasına odaklanılarak, sonuçların doğrudan yorumlanabilir ve gözlem birimleri için tahmin hatasının öngörülebilmesi sağlanmaktadır.

Tahmin gücünün hesaplanmasında kullanılan yaklaşım, kümeleme sonuçlarını karşılaştırabilmek için veri seti içerisinden bağımsız olarak test ve eğitim veri setlerinin çekilmesidir. Test ve eğitim setleri, k adet kümeye ayrılmaktadır. Sonrasında, eğitim seti ile belirlenen küme merkezleri, test setinin kümelenmesi için kullanılmakta ve test seti için bulunan ilk kümeleme sonuçları ile karşılaştırılmaktadır. Kümeleme sonucunda aynı kümede sınıflandırılan ortak üyelikler belirlenmeye çalışılmaktadır. Her test kümesi için, her iki kümeleme sürecinde de aynı kümede sınıflandırılan gözlemlerin oranı hesaplanmaktadır. Tahmin gücü, k test kümesi için hesaplanan oranların aldığı en küçük değer olmaktadır. Küme sayısı bir olduğunda, test kümesi her iki kümeleme sonucunda aynı sınıfta olacağından, tahmin gücü bire eşit olmaktadır. Tahmin gücü değeri için belirlenen eşik değer 0.8 - 0.9 civarında olduğunda, kümeler arası ayrımın iyi olduğu gözlemlenmektedir (Tibshirani, Walther 2005: 511-515).

4. UYGULAMA

Bu bölümde 2019 ve 2021 yıllarına ait Türkiye İstatistik Kurumu'nca, Türkiye'deki bilişim ve iletişim teknolojileri kullanımına yönelik olarak toplanmış ve 'Hanehalkı Bilişim Teknolojileri Kullanım Araştırması Mikro Veri Seti' adı altında derlenmiş olan veri setleri, belirlenen değişkenlerin sürekli ve kategorik değişkenlerden oluşmasından dolayı, bu tipteki verilerin kümelenmesine fırsat tanıyan Kamila algoritması ile kümelenmiştir. Öncelikle Türkiye'de bilişim teknolojileri kullanımına dair literatür taraması sunulmuştur. Sonrasında analizden elde edilen bulgulara yer verilmiştir.

4.1. Türkiye'de Bilişim Teknolojileri Kullanımına Ait Literatür Taraması

Kümeleme analizi sosyal bilimler alanında sıkça faydalanılan, çok değişkenli istatistiksel bir yöntemdir. Bu çalışmanın kapsamı dikkate alınarak, literatürde bilişim teknolojilerinin kullanımı alanında yapılmış çalışmalara ve bu çalışmada uygulaması yapılan KAMILA kümeleme algoritmasının farklı alanlardaki uygulamalarına yer verilecektir.

Arıcıgil Çilan ve Kuzu (2013) çalışmalarında, Türkiye'deki kişisel e-ticaret uygulamalarının demografik faktörlerle ilişkisini oranlar, frekans tabloları ve parametrik olmayan testler kullanarak kategorik veri analizi yöntemleri ile araştırmışlar, bu doğrultuda Türkiye İstatistik Kurumu'nun 2012 yılı Hanehalkı Bilişim Teknolojileri Kullanım Araştırması mikro veri setlerini kullanmışlardır. Elde edilen bulguları, tüketiciler açısından algılanan riskler, eğitim seviyesinin önemi ve yaş faktörü açılarından değerlendirmişlerdir.

Arıcıgil Çilan vd. (2013) çalışmalarında, TÜİK'in 2012 yılı Hanehalkı Bilişim Teknolojileri Kullanım Araştırması veri setlerini kullanarak, Türkiye'de internet kullanımının profilini tanımsal istatistik ölçülerde belirmeyi ve Türkiye'de fertlerin internet kullanım faaliyetlerine göre kaç kümede toplanabileceğini Gizil Sınıf Analizi ile incelemeyi amaçlamışlardır, Araştırma sonucunda, 16-74 yaş arasında yer alan ve son üç ay içerisinde internet kullanan nüfusun kişisel internet kullanım faaliyetlerinde üç sınıfa ayrılabileceği belirlenmiştir. Cinsiyet ve eğitim değişkenlerine göre farklılıklar da bulgulara dahil edilmiştir.

Anıl ve Köksal (2016) çalışmalarında, TÜİK'in 2014 yılı Hanehalkı Bilişim Teknolojileri Kullanımı Araştırması verilerini kullanarak, aşamalı regresyon analizi uygulamışlardır. Gelir, eğitim, yaş, cinsiyet gibi değişkenlerin hem internet erişiminde hem de kullanımında etkili olduğu sonucuna ulaşılmıştır. Hanede bulunan ilkokul çağındaki çocuk sayısının internet erişiminde etkin olduğu, gelir, yaş ve eğitim seviyelerinin internet kullanımında farklılıklar yarattığı ortaya konmuştur.

Fidan (2017) çalışmasında, 2011 – 2014 seneri arasında,Türkiye İstatistik Kurumu'nun yayınladığı Hanehalkı Bilişim Teknolojileri Kullanımı Anketi ve Nüfus ve Demografi İstatistikleri veri setlerini kullanarak, Türkiye'de Düzey1 bölgelerinde yaşayan kişilerin bilgisayar ve internet kullanım düzeyleri arasındaki farklılıkları Gini katsayıları ile belirlemeyi amaçlamışlardır. Araştırma sonuçlarında, bilgisayar ve internet kullanımlarında bölgeler arasında düşük seviyede sayısal bölünmenin bulunduğu, TR9'un en yüksek, TR11'in en düşük sayısal bölünme seviyesine sahip olduğu tespit edilmiştir.

Görgün Baran ve Erdem (2017) çalışmalarında, 2016 senesine ait Hanehalkı BT Kullanımı Araştırması veri kullanılmış ve Türkiye'de 16-74 yaş bireylerde bilgisayar ve internet kullanım yeteneklerinin yaş, cinsiyet, eğitim, çalışma durumu, kullanım sıklıkları, kullanma amaçları ve yaşadıkları istatistiki bölgelere göre etkilerini logistic regresyon yöntemi ile analiz etmişlerdir ve Türkiye bağlanımda toplumsal cinsiyetin bilgisayar kullanım yeteneklerine sahip olma olasığının erkeklerin lehine olduğu sonucuna ulaşmışlardır.

Kara vd. (2017) çalışmalarında, 2004-2014 yılları arasında cinsiyet, yaş ve eğitim durumu değişkenleri ile aktif internet kullanımına yönelik eğilimleri belirlemek amacıyla betimsel istatistik yöntemlerini kullanarak kritik yılları tespit etmeyi amaçlamışlardır. Bu amaç doğrultusunda değişkenleri birbirleriyle karşılaştırmışlardır. Çalışma bulgularına göre ülkemizde eğitim seviyesi arttıkça aktif internet kullanım oranı da artmakta ve erkekler, kadınlara oranla daha sık internet kullanmaktadır. 16-24 yaş ve 25-34 yaş aralığındaki genç yetişkin fertlerin aktif internet kullanım oranlarının, diğer yaş gruplarına göre analiz yapılan senelerin içerisinde her zaman daha yüksek olduğu belirlenmiştir.

Selim ve Balyaner (2017) çalışmalarında, Türkiye İstatistik Kurumu'nun 2013 yılı Hanehalkı BT Kullanımı Araştırması verilerini kullanarak, Türkiye'de 6-15 yaş arası çocukların ve yetişkinlerin sahip olduğu bilişim teknolojileri ürünleri sayısını belirleyen faktörlerin sayma veri modeli ile incelenmesini amaçlamışlardır. Analizden elde edilen sonuçlarına göre, Türkiye'de 6-15 yaş arası çocukların ve yetişkinlerin sahip olduğu bilişim teknolojileri ürünleri sayısının doğudan batıya ve kırsal kesimden kentsel kesime doğru artış göstermektedir.

Sezer vd. (2019) çalışmalarında, Türkiye'de bilgisayar ve internet kullanım alışkanlıklarının neler olduğunu ve yanlış kullanım sonucunda ortaya çıkarabileceği sorunları belirlemeyi amaçlamışlardır. Türkiye İstatistik Kurumu tarafından derlenmiş olan veriler ile yapılmış araştırmalardan elde edilmiş bulgular betimsel tarama modeli ile değerlendirilmiş, bulgular çalışmanın amacına uygun olarak tekrar tablolaştırılarak frekans değerleriyle sunulmuştur. Yapılan çalışmanın sonucunda Türkiye'de internet ve bilgisayar kullanım oranlarının hızla arttığı ve buna bağlı olarak

kişisel arası ilişkilerde, ebeveyn çocuk etkileşiminde, iş ve eğitim yaşamında problem yaşayan fertlerin sayısında artış olduğu belirlenmiştir.

Coşkun vd. (2019) çalışmalarında, Türkiye İstatistik Kurumu'nun 2016 yılı Hanehalkı Bilişim Teknolojileri Kullanım Anketi verilerini kullanarak, hanehalkının internet hizmetlerine sahip olma durumunu incelemişler ve hanehalkı internet kullanımını etkileyen faktörlerin karar ağaçları ile analiz edilmesini amaçlamışlardır. Analiz sonucunda önemli faktörlerin hane cep telefonu sahipliği, hane bilgisayar kullanımı, hanede 0-25 yaş arasında fertlerin bulunup bulunmaması, hane tablet sahipliği, hane dizüstü bilgisayar sahipliği, hanehalkı büyüklüğü, hanehalkı reisinin yaşı, hane smart TV sahipliği ve hane aylık geliri olduğunu belirlemişlerdir.

Marangoz vd. (2019) çalışmalarında, Türkiye İstatistik Kurumu'nun 2016 yılı Hanehalkı Bilişim Teknolojileri Kullanım Anketi verilerini kullanarak, tüketicilerin internet üzerinden gerçekleştirdikleri alışveriş davranışlarını ve bu davranışları etkileyen faktörleri belirlemeyi amaçlamışlardır. Bu hedefle genelleştirilmiş doğrusal modellerden yararlanmışlardır. Araştırma sonucunda, eğitim ve gelir düzeyi arttıkça internetten harcama miktarında artış olduğunu; hanehalkı büyüklüğünün ise internetten harcama miktarı üzerinde olumsuz etkiye sahip olduğunu belirlemişlerdir.

Alkan vd. (2022) çalışmalarında, Covid-19 döneminde bireylerin e-ticaret sıklığını etkileyen faktörlerin belirlenmesini ve tüketici profillerinin ilgili dönemdeki alışveriş alışkanlıklarına etkisinin tespit edilmesini amaçlamışlardır. Türkiye İstatistik Kurumu'nun 2021 yılı Hanehalkı Bilişim Teknolojileri Kullanım Anketi verilerini, genelleştirilmiş sıralı logistic regresyon yöntemi ile incelemişlerdir. Elde edilen bulgularda, genç yaşta, üniversite mezunu, kadın, profesyonel meslek gruplarında çalışan ve yüksek gelirli, aktif sosyal medya paylaşımı yapan, internet bankacılığı kullanan, dizüstü bilgisayar sahibi ve batı bölgelerdeki illerde ikamet eden bireylerin diğer gruplara göre Covid-19 döneminde daha sık alışveriş yaptığını belirlemişlerdir.

Alkan ve Ünver (2022) çalışmalarında, Türkiye'de Z kuşağının e-ticaret kullanımını tercih etmelerinde etkili olan ekonomik ve sosyo-demografik özelliklerin logistik regresyon analizi kullanarak araştırılmasını amaçlamışlardır. Bu amaçla, TÜİK'in 2021 yılı Hanehalkı Bilişim Teknolojileri Kullanımı Araştırması verilerini kullanmışlar ve çalışmanın sonucunda eğitim durumu, yaş. gelir düzeyi, işteki durum, cinsiyet, istatistiki bölge, cep telefonu sahipliği ve hanehalkı büyüklüğü değişkenlerinin e-ticaret kullanımıyla ilişkili olduğunu sonucuna ulaşılmıştır.

Demirel (2022) çalışmalarında, TÜİK'in 2020 yılı Hanehalkı BT Kullanım Araştırması verilerini kullanarak, logit modeli ile bilgisayar, bivariate probit modeliyle taşınabilir bilgisayar (PC), laptop ve tablet sahipliğini etkileyen faktörleri ayrı ayrı analiz etmişlerdir. Logit modeli ile yapılan çalışmada yaş ve hane büyüklüğünün, bilgisayara sahip olmayı olumsuz etkilediği sonucu elde

edilmiştir. Gelir, eğitim düzeyi, çalışıyor olma durumu, internet kullanım sıklığı ve online eğitim almanın ise bilgisayara sahip olma durumunu olumlu etkilediğini tespit etmişlerdir. Bivarite probit modelleriyle yapılan çalışmalarda ise, yüksek gelir ve eğitim düzeyinin taşınabilir bilgisayar sahipliğini olumsuz etkilediği sonucuna ulaşılmıştır. Benzer şekilde yüksek gelir ve eğitim düzeyinin, laptop ve tablet sahipliğini olumlu; online eğitim almanın, taşınabilir bilgisayar ve taşınabilir bilgisayar sahipliğini olumlu; tablet sahipliğini olumsuz etkilediği sonuçları elde edilmiştir.

Ecemiş ve Coşkun (2022) çalışmalarında, 2014-2021 Dönemi Hanehalkı Bilişim Teknolojileri Kullanımı Araştırma verilerinden faydalanarak, çok kriterli karar verme yöntemleriyle istatistiki bölge düzeyinde bilişim teknolojilerinin kullanımına yönelik gelişim performanslarını ölçebilecek bir model ile internet erişimi, hanede kullanılan bağlantı türleri ve internet kullanımı kriterleriyle analiz yapmışlardır. Elde edilen bulgulara göre 2014 senesinde internet erişimi en önemli kriter iken, bu kriter 2021 senesinde mobil bant geliştiği olarak bulunmuştur. Çalışma kapsamında en fazla gelişim gösteren bölgelerin ise sırasıyla Güneydoğu Anadolu Bölgesi, Batı Anadolu Bölgesi, Akdeniz Bölgesi, Orta ve Orta Doğu Anadolu Bölgeleri olduğu sonucuna ulaşılmıştır.

Foss vd. (2016), karma veri türündeki veri setleri için mevcut kümeleme yöntemlerinin, güçlü parametrik varsayımlar olmadan sürekli ve kategorik değişkenlerin katkısını adil bir şekilde dengeleyemediğini tespit etmişler, bu sorunun giderilmesini amaçlamışlar ve sorunu ele alan kümeleme yöntemi olarak Kamila'yı geliştirmişlerdir. Çalışmalarında teorik incelenme ile birlikte simülasyon ve gerçek veriler üzerinden uygulama yapmışlardır.

Bilgiç (2019), farklı ölçekler ile ölçülmüş karma tipteki değişkenlerden oluşan süpermarket alışveriş verisinin, KAMILA, k-ortalamalar, k-ortaylar ve k-prototipler algoritmaları ile kümeleme analizini gerçekleştirmişlerdir. KAMILA kümeleme analizi ile farklı demografik özelliklere ve satın alma davranışlarına sahip müşterilerden, altın segment olarak isimlendirilebilecek segmentteki müşterilerin başarıyla tespit edilebildiğini belirtmişlerdir.

Ahmad ve Khan (2019) çalışmalarında, karma veri setleri için kullanılan kümeleme yöntemlerini beş ana başlıkta (ayırıcı, hiyerarşik, model tabanlı, sinir ağları ve diğer) değerlendirmişler, bu yöntemlerin güçlü ve zayıf yönlerini analiz ederek, detaylı bir derleme yapmışlardır.

Mbuga ve Tortora (2021) çalışmalarında, spektral kümeleme yöntemini karma tip veri setlerinde kullanmaya uygun olacak şekilde genişletmeyi ve bu yeni yöntemde sürekli ve kategorik değişkenler için geleneksel olarak kullanılan öklid mesafesi tabanlı benzerlik ölçüsünü farklı ölçülerle değiştirmeyi amaçlamışlardır. Bu amaç doğrultusunda spektral kümelemenin performansını, karma tip veri kümeleme yöntemi olan k-prototypes ve KAMILA ile karşılaştırmışlardır.

4.2. Veri Seti ve Değişkenler

Bilişim teknolojilerindeki gelişmelerin katlanarak artması sebebiyle toplumun yapı taşı olan bireylerden başlanarak, toplumsal ve küresel ölçeklerde bu gelişmelerden kaynaklanan değişimlerin analiz edilmesi ve etkilerinin ortaya çıkarılması önem taşımaktadır. Bu amaçla Türkiye İstatistik Kurumu'nun yayınladığı, Hanehalkı Bilişim Teknolojileri Kullanımı Araştırması Mikro Veri Seti pek çok araştırmacıya zengin bir veri kaynağı olarak kullanılma imkanı vermektedir. Bu çalışmada 2019 ve 2021 senelerine ait veri setleri kaynak olarak kullanılmış ve analiz edilmiştir. Analizlerin temel amaçlarını iki başlık altında derlemek mümkündür.

1- Hane bazında yapılacak araştırmada, gözlemlerin hanede bulunan birey sayısı, hane aylık geliri (TL), hanede bulunan bilişim ekipmanları, hanede kullanılan internet bağlantısı türleri ve hanenin bulunduğu istatistiki bölgeler değişkenleri için analiz edilerek, gözlemlerin kümelenmesi incelenecektir.

2- Fert bazında iki farklı analiz yapılması planlanmıştır. Bunlardan ilkinde gözlemlerin yaş, cinsiyet ve e-ticaret alışkanlıklarına göre kümelenmesi amaçlanmıştır. İkinci analizde ise, gözlemlerin yaş, internet kullanım sıklığı, eğitim durumu, meslek, taşınabilir cihazlar üzerinden yapılan internet faaliyetleri değişkenleri ile değerlendirilip, ortaya çıkacak grupların özellikleri incelenecektir.

İstatistiki Bölge Birimleri Sınıflaması Düzey 1 ile belirlenen 12 Bölgeye ait kodlar ve tanımlar Tablo 4.1'de verilmektedir.

Tablo 4.1: İstatistiki Bölge Birimleri Sınıflaması Düzey 1 (12 Bölge)

Renklendirme araştırmacı tarafından yapılmıştır.

Kod	Tanım	Kod	Tanım
TRA	Kuzeydoğu Anadolu	TR4	Doğu Marmara
TRB	Ortadoğu Anadolu	TR5	Batı Anadolu
TRC	Güneydoğu Anadolu	TR6	Akdeniz
TR1	İstanbul	TR7	Orta Anadolu
TR2	Batı Marmara	TR8	Batı Karadeniz
TR3	Ege	TR9	Doğu Karadeniz

Bölge sınıflandırmasına ait görsel, Türkiye İstatistiki Bölge Birimleri Sınıflandırması Haritası Şekil 4.1'de yer almaktadır.

Şekil 4.1: Türkiye İstatistiki Bölge Birimleri Sınıflandırması (Türkiye İBBS) Haritası

Kaynak: İstanbul: Yıldız Teknik Üniversitesi, Meryem Hayır Kanat, Coğrafya Alan İncelemeleri Bölge Kavramı ve NUTS Bölgeleri, https://avesis.yildiz.edu.tr/search?scope=All&q=mhayir, s.7

4.3. Yöntem

Bu çalışmanın amacı, karma yapıdaki büyük veri setlerinin sınıflandırılması için tasarlanmış KAMILA algoritması ile farklı değişkenler kullanılarak gözlemlerin, hane ve fert bazında bilişim teknolojilerinin kullanımının sosyo-demografik ve ekonomik faktörler açısından analiz edilerek benzer gözlemlerin kümelenmesidir. Bu amaç doğrultusunda, TÜİK tarafından derlenmiş, 2019 ve 2021 senelerine ait Hanehalkı Bilişim Teknolojileri Kullanımı Araştırması Mikro Veri Setleri kullanılmıştır. İki farklı seneye ait veri setlerinin kullanılması ile Covid-19 pandemisinin etkilerinin, hane ve fert bazında ortaya çıkardığı değişikliklerin gözlenmesi hedeflenmiştir. KAMILA algoritmasının çalışma süresi her analiz için not edilmiş ve ilgili tablolarda bilgi olarak sunulmuştur. Çalışma süresinin gözlem sayısı, değişken sayısı ve hesaplanan küme sayısına göre değişiklik gösterdiği gözlemlenmiştir.

4.4. Hanehalkı Veri Setlerine Ait Bulgular

2019 ve 2021 senelerine ait Hanehalkı Bilişim Teknolojileri Kullanımı Araştırması Mikro Veri Setleri, KAMILA algoritması ile sınıflandırılmış ve ulaşılan sonuçlar 2019 senesine ait bulgular, 2021 senesine ait bulgular ve 2019-2021 seneleri karşılaştırması olarak üç başlıkta toplanmıştır.

Veri setinde bulunan iki sayısal değişkenin, hane büyüklüğü ve hanenin toplam aylık geliri, tanım aralıkları birbirlerinden farklı olduğundan, bu iki değişkene standartlaştırma işlemi uygulanmıştır.

4.4.1. 2019 Senesine Ait Bulgular

2019 senesi hanehalkı veri seti ile hanede bulunan kişi sayısı, toplam aylık geliri (TL olarak), hanede bulunan bilişim ekipmanları, hanede kullanılan internet bağlantısı türleri ve istatistiki bölge birimleri arasındaki benzerlikler incelenmiş ve gözlemlerin sınıflandırılması amaçlanmıştır.

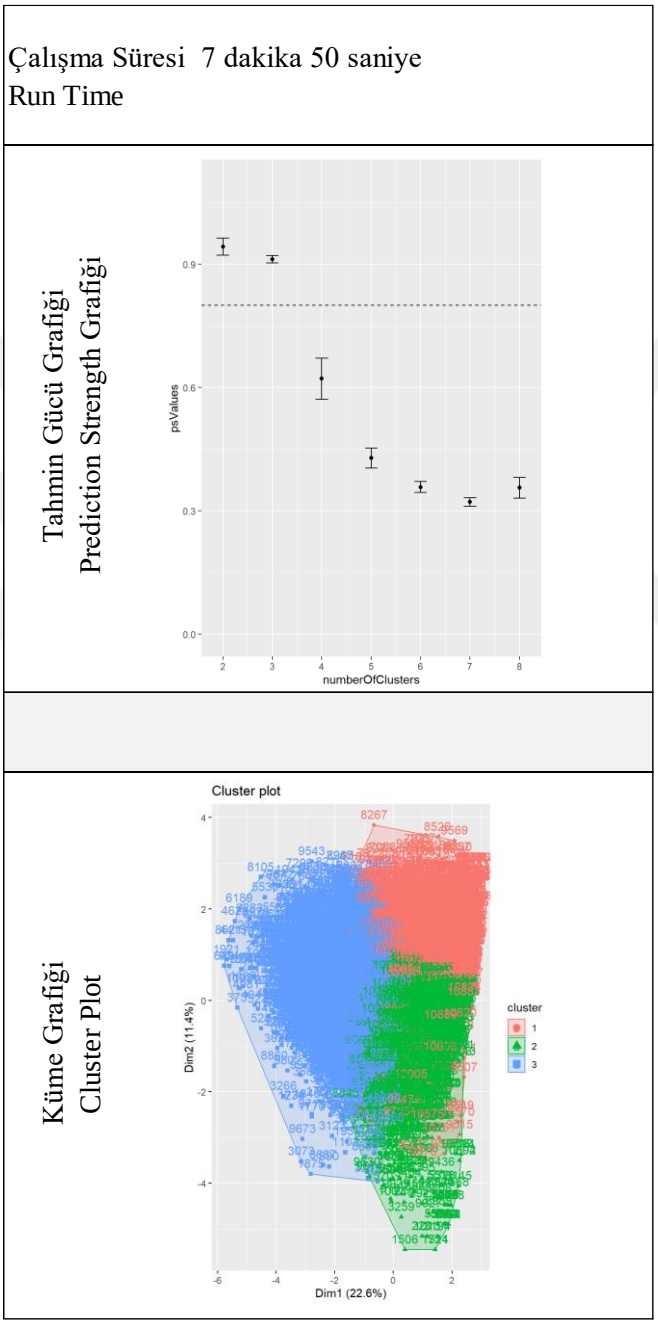
2019 senesine ait hanehalkı original veri setinde toplam 13 561 gözlem bulunmaktadır. Bu gözlemlerden kapsamiçi (12 956 gözlem), hane internet erişim durumu değişkeni için evet cevabını veren hanelere ait gözlemler analize dahil edilmiştir (11 163 gözlem). Verinin %98'i (10 894 gözlem) ile çalışılmıştır. Hane toplam aylık gelir değişkeni için alt ve üst değerler 500 TL – 15.000 TL, ortalama değeri 3.670 TL; hanede bulunan kişi sayısı değişkeni için alt ve üst değerler 1-9, medyan değeri 3 olarak hesaplanmıştır. Cep telefonuna sahip olmayan ve ISDN bağlantı kullanan hane sayısı oldukça düşük olduğundan analize dahil edilmemişlerdir. Değişkenlerin bilgileri Tablo 4.2'te verilmiştir.

Tablo 4.2: 2019 Senesi Hanehalkı Analizi Değişken Listesi ve Özellikleri

	Değişken No	Değişken Adı	Değişken Açıklama / Seçenekleri	Değişken Tipi	Seviyeler	Sıklık	Oran %
	1	HHB	Hane büyüklüğü / Hanede bulunan kişi sayısı	Sayısal			
	2	Hane_Gelir_Aylık	Hanenin toplam aylık geliri (TL olarak)	Sayısal			
Hanede Bulunan Bilişim Ekipmanları	3	BT_Bilg_Masaüstü	Masaüstü bilgisayar	Kategorik	1 - Evet	2104	0.19
					2 - Hayır	8790	0.81
	4	BT_Bilg_Taşınabilir	Dizüstü, netbook vb.	Kategorik	1 - Evet	4374	0.40
					2 - Hayır	6520	0.60
	5	BT_Bilg_Tablet	Tablet bilgisayar	Kategorik	1 - Evet	3263	0.30
					2 - Hayır	7631	0.70
	6	BT_Oyunkonsol	Oyun konsolu (Playstation, Wii, Xbox vb.)	Kategorik	1 - Evet	612	0.06
					2 - Hayır	10282	0.94
	7	BT_Int_bağlanan_TV	İnternete bağlanabilen TV (Smart TV)	Kategorik	1 - Evet	4410	0.40
					2 - Hayır	6484	0.60
Evde Kullanılan İnternet Bağlantısı Türleri	8	Int_Bağlantı_ Sabit_Geniş	Sabit genişbant bağlantı(ADSL, kablolu internet (Uydunet), fiber bağlantı vb.)	Kategorik	1 - Evet	5759	0.53
					2 - Hayır	5135	0.47
	9	Int_Bağlantı_ Mobil_Geniş	Mobil genişbant bağlantı (3G, 4.5G,taşınabilir bir cihaz ile) cep telefonu, tablet vb.) veya 3G, 4.5G,	Kategorik	1 - Evet	10703	0.98
					2 - Hayır	191	0.02
	10	Int_Baglantı_ Tel_Darbant	Cep telefonu üzerinden darbant bağlantı (WAP, GPRS)	Kategorik	1 - Evet	918	0.08
					2 - Hayır	9976	0.92
	11	IBBS_1	İstatistiki bölge birimleri sınıflama düzeyi (Düzey 1)	Kategorik	1 - TR1	1526	0.14
					2 - TR2	694	0.06
					3 - TR3	1169	0.11
					4 - TR4	1072	0.10
					5 - TR5	1153	0.11
					6 - TR6	1198	0.11
					7 - TR7	751	0.07
					8 - TR8	715	0.07
					9 - TR9	519	0.05
					10 - TRA	564	0.05
					11 - TRB	656	0.06
					12 - TRC	877	0.08

Analiz sonuçlarına ait tahmin gücü (prediction strength) grafiği ve küme grafiği (cluster plot) Şekil 4.2'te verilmiştir. Tahmin gücü grafiğine göre verinin 3 kümeye ayrılması uygun gözükmektedir.

Kamila algoritmasının çalışma süresi, 11 değişken, 10 894 gözlem ve 28 seviyeli veri setinde 7 dakika 50 saniye olmuştur.



Şekil 4.2: 2019 Senesi Hane Verisi Kümeleme Sonuçlarına Ait Analiz Görselleri

Tablo 4.3'de analiz sonucu elde edilen sınıflandırma sonucuna ait bilgiler verilmiştir. Her değişken için ait olduğu küme içerisinde en çok gözlenen seviye renklendirilmiştir. Mavi renk ile

renklendirilen seviyeler, kümeler arasında bir fark olmadığını belirtirken, kırmızı ile renklendirilen seviyeler, kümeler arasında farklılaşma olduğunu işaret etmektedir.

Tablo 4.3: 2019 Senesi Hanehalkı Kümeleme Sonuçları

	Değişken No	Değişken Adı	Değişken Tipi	Seviyeler	Küme 1	Küme 2	Küme 3	**Toplam**
	1	HHB	Sayısal					
	2	Hane_Gelir_Aylık	Sayısal					
Hanede Bulunan Bilişim Ekipmanları	3	BT_Bilg_Masaüstü	Kategorik	1 - Evet	445	277	1382	**2104**
				2 - Hayır	2766	2984	3040	**8790**
	4	BT_Bilg_Taşınabilir	Kategorik	1 - Evet	426	648	3300	**4374**
				2 - Hayır	2785	2613	1122	**6520**
	5	BT_Bilg_Tablet	Kategorik	1 - Evet	723	235	2305	**3263**
				2 - Hayır	2488	3026	2117	**7631**
	6	BT_Oyunkonsol	Kategorik	1 - Evet	24	23	565	**612**
				2 - Hayır	3187	3238	3857	**10282**
	7	BT_Int_bağlanan_TV	Kategorik	1 - Evet	794	734	2882	**4410**
				2 - Hayır	2417	2527	1540	**6484**
Evde Kullanılan İnternet Bağlantısı Türleri	8	Int_Bağlantı_ Sabit_Geniş	Kategorik	1 - Evet	774	902	4083	**5759**
				2 - Hayır	2437	2359	339	**5135**
	9	Int_Bağlantı_ Mobil_Geniş	Kategorik	1 - Evet	3176	3159	4368	**10703**
				2 - Hayır	35	102	54	**191**
	10	Int_Baglantı_ Tel_Darbant	Kategorik	1 - Evet	200	277	441	**918**
				2 - Hayır	3011	2984	3981	**9976**
	11	IBBS_1	Kategorik	1 - TR1	239	375	912	**1526**
				2 - TR2	39	313	342	**694**
				3 - TR3	215	435	519	**1169**
				4 - TR4	182	302	588	**1072**
				5 - TR5	219	318	616	**1153**
				6 - TR6	367	440	391	**1198**
				7 - TR7	244	249	258	**751**
				8 - TR8	164	274	277	**715**
				9 - TR9	166	232	121	**519**
				10 - TRA	355	105	104	**564**
				11 - TRB	367	146	143	**656**
				12 - TRC	654	72	151	**877**

Birinci kümedeki hane özellikleri incelendiğinde, hanede bulunan kişi sayısı 5, hane toplam aylık geliri ortalaması 2.359 TL, bilişim ekipmanlarına sahip olmayan, mobil internet bağlantısı

kullanan, Kuzeydoğu Anadolu, Ortadoğu Anadolu ve Güneydoğu Anadolu Bölgeleri'nde yaşamakta oldukları belirlenmiştir.

İkinci kümedeki hanelerin hanede bulunan kişi sayısı 2, hane toplam aylık geliri ortalaması 2.715 TL, bilişim ekipmanlarına sahip olmayan, mobil internet bağlantısı kullanan, Akdeniz ve Doğu Karadeniz Bölgeleri'nde yaşamakta oldukları belirlenmiştir.

Üçüncü kümede yer alan hanelerin hanede bulunan kişi sayısı 3, hane toplam aylık geliri ortalaması 5.357 TL, taşınabilir bilgisayar, tablet ve internete bağlanan TV ekipmanlarına sahip, sabit geniş bant bağlantı kullanan ve Marmara Bölgesi (İstanbul, Batı Marmara, Doğu Marmara), Ege, Batı Anadolu, Orta Anadolu ve Batı Karadeniz Bölgeleri'nde yaşamakta oldukları sonucuna ulaşılmıştır.

Tüm bulgular birlikte değerlendirildiğinde, birinci kümede bulunan hanelerin, hanede bulunan birey sayısı yüksek, hane ortalama gelir düzeyi düşük, bilişim ekipmanlarına sahip olmayan, mobil internet bağlantısı kullanan, Türkiye'nin doğusunda yer alan bölgelerde bulunduğu belirlenmiştir.

İkinci kümede yer alan hanelerin, hanede bulunan birey sayısı düşük, hane toplam aylık geliri ortalaması düşük, bilişim ekipmanlarına sahip olmayan, mobil internet bağlantısı kullanan, Türkiye'nin kuzey doğu ve güneyde yer alan bölgelerinde yaşamakta oldukları belirlenmiştir.

Üçüncü kümede yer alan hanelerin, hane toplam aylık geliri ortalaması diğer iki kümeden yüksektir. Bununla birlikte, hanede taşınabilir bilgisayar, tablet ve internete bağlanan TV ekipmanları bulunmakta ve sabit geniş internet bağlantısı kullanmaktadırlar.

Birinci ve ikinci kümenin özellikleri karşılaştırıldığında, hane ortalama gelirilerinin benzer değerlerde olduğu, bunun yanında hanede bulunan kişi sayısında belirgin bir fark olduğu görülmektedir. Üçüncü kümede yer alan haneler ise, diğer iki kümeden hem hane ortalama gelir düzeyi hem de hanede bulunan ekipmanların çeşitliliği ve kullanılan internet bağlantısı açısından farklılaştığı görülmektedir.

4.4.2. 2021 Senesine Ait Bulgular

2021 senesi hanehalkı veri seti ile hanede bulunan kişi sayısı, toplam aylık geliri (TL olarak), hanede bulunan bilişim ekipmanları, hanede kullanılan internet bağlantısı türleri ve istatistiki bölge birimleri arasındaki benzerlikler incelenmiştir.

2021 senesine ait hanehalkı original veri setinde toplam 13 662 gözlem bulunmaktadır. Hane internet erişim durumu değişkeni için evet cevabını veren hanelere ait gözlemler analize dahil edilmiştir. Analizde verinin %97'si (11 914 gözlem) ile çalışılmıştır. Hane toplam aylık gelir değişkeni için alt ve üst değerler 300 TL – 19.000 TL, ortalama değeri 4.516 TL; hanede bulunan kişi sayısı

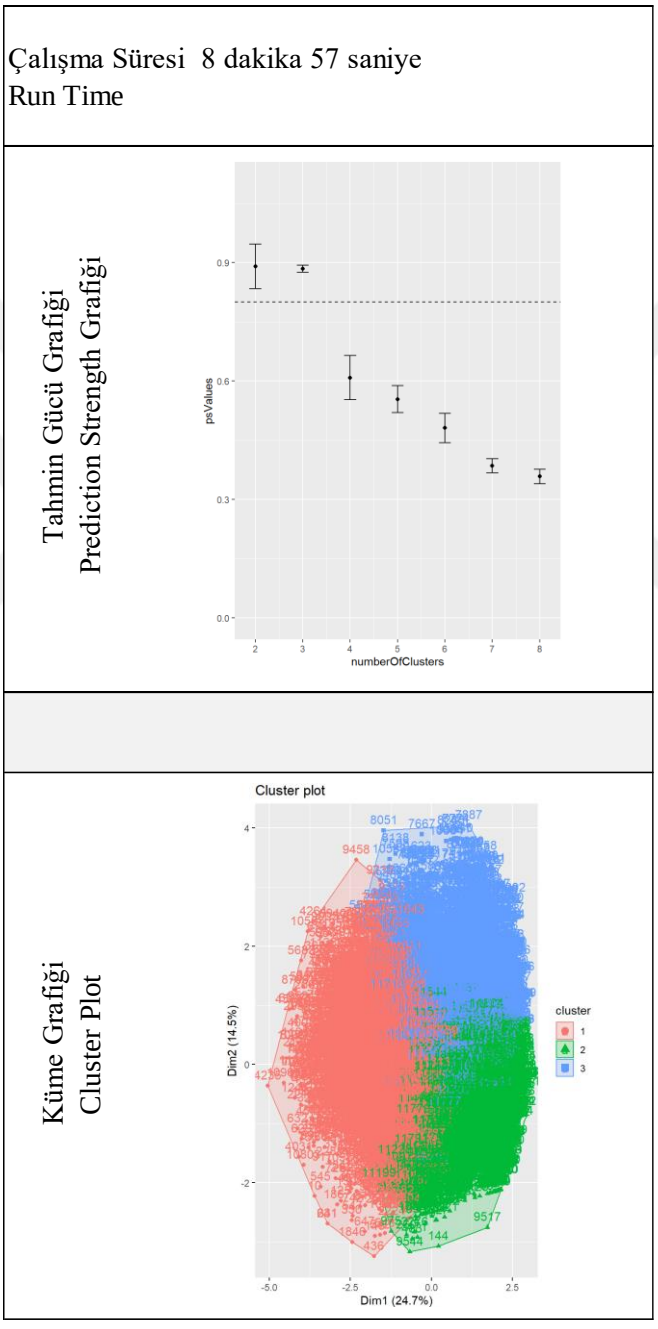
değişkeni için alt ve üst değerler 1-8, medyan değeri 3 olarak hesaplanmıştır. Cep telefonuna sahip olmayan hane olmadığından, BT_Tel_Cep değişkeni kümeleme analizine dahil edilmemiştir. Değişkenlerin bilgileri Tablo 4.4'te verilmektedir.

Tablo 4.4: 2021 Senesi Hanehalkı Analizi Değişken Listesi ve Özellikleri

	Değişken No	Değişken Adı	Değişken Açıklama / Seçenekleri	Değişken Tipi	Seviyeler	Sıklık	Oran %
	1	HHB	Hane büyüklüğü / Hanede bulunan kişi sayısı	Sayısal			
	2	Hane_Gelir_Aylık	Hanenin toplam aylık geliri (TL olarak)	Sayısal			
Hanede Bulunan Bilişim Ekipmanları	3	BT_Bilg_Masaüstü	Masaüstü bilgisayar	Kategorik	1 - Evet	2004	0.17
					2 - Hayır	9910	0.83
	4	BT_Bilg_Dizüstü	Dizüstü bilgisayar	Kategorik	1 - Evet	4674	0.39
					2 - Hayır	7240	0.61
	5	BT_Tablet	Tablet bilgisayar	Kategorik	1 - Evet	3264	0.27
					2 - Hayır	8650	0.73
		BT_Tel_Cep	Cep telefonu ya da akıllı telefon	Kategorik	1 - Evet	11914	1.00
					2 - Hayır	0	0.00
	6	BT_Diğer_Cihaz	Diğer cihazlar (akıllı TV, akıllı hoparlör, oyun konsolu, e-kitap okuyucu, akıllı saat vb.)	Kategorik	1 - Evet	7214	0.61
					2 - Hayır	4700	0.39
Evde Kullanılan İnternet Bağlantısı Türleri	7	Int_Bağlantı_ Sabit_Geniş	Sabit genişbant bağlantı(DSL, ADSL, VDSL, kablo, optik fiber, uydu, Wi-Fi)	Kategorik	1 - Evet	7727	0.65
					2 - Hayır	4187	0.35
	8	Int_Bağlantı_ Mobil_Geniş	Mobil genişbant bağlantı (3G, 4.5G,taşınabilir bir cihaz ile cep telefonu, tablet vb.) veya 3G, 4.5G modemi üzerinden	Kategorik	1 - Evet	11500	0.97
					2 - Hayır	414	0.03
	9	IBBS_1	İstatistiki bölge birimleri sınıflama düzeyi (Düzey 1)	Kategorik	1 - TR1	1575	0.13
					2 - TR2	764	0.06
					3 - TR3	1402	0.12
					4 - TR4	1178	0.10
					5 - TR5	1191	0.10
					6 - TR6	1293	0.11
					7 - TR7	752	0.06
					8 - TR8	775	0.07
					9 - TR9	697	0.06
					10 - TRA	551	0.05
					11 - TRB	672	0.06
					12 - TRC	1064	0.09

Analiz sonuçlarına ait tahmin gücü grafiği ve küme grafiği Şekil 4.3'te verilmiştir. Tahmin gücü grafiğine göre verinin 3 kümeye ayrılması uygun gözükmektedir.

Kamila algoritmasının çalışma süresi, 9 değişken, 11 914 gözlem ve 24 seviyeli veri setinde 8 dakika 57 saniye olmuştur.



Şekil 4.3: 2021 Senesi Hane Verisi Kümeleme Sonuçlarına Ait Analiz Görselleri

Tablo 4.5'de analiz sonucu elde edilen sınıflandırma sonucuna ait bilgiler verilmiştir. Her değişken için ait olduğu küme içerisinde en çok gözlenen seviye renklendirilmiştir.

Tablo 4.5: 2021 Senesi Hanehalkı Kümeleme Sonuçları

	Değişken No	Değişken Adı	Değişken Tipi	Seviyeler	Küme 1	Küme 2	Küme 3	**Toplam**
	1	HHB	Sayısal					
	2	Hane_Gelir_Aylık	Sayısal					
Hanede Bulunan Bilişim Ekipmanları	3	BT_Bilg_Masaüstü	Kategorik	1 - Evet	1331	231	442	**2004**
				2 - Hayır	3382	3600	2928	**9910**
	4	BT_Bilg_Dizüstü	Kategorik	1 - Evet	3629	627	418	**4674**
				2 - Hayır	1084	3204	2952	**7240**
	5	BT_Tablet	Kategorik	1 - Evet	2150	176	938	**3264**
				2 - Hayır	2563	3655	2432	**8650**
	6	BT_Diğer_Cihaz	Kategorik	1 - Evet	3609	1709	1896	**7214**
				2 - Hayır	1104	2122	1474	**4700**
Evde Kullanılan İnternet Bağlantısı Türleri	7	Int_Bağlantı_ Sabit_Geniş	Kategorik	1 - Evet	4532	1450	1745	**7727**
				2 - Hayır	181	2381	1625	**4187**
	8	Int_Bağlantı_ Mobil_Geniş	Kategorik	1 - Evet	4475	3732	3293	**11500**
				2 - Hayır	238	99	77	**414**
	9	IBBS_1	Kategorik	1 - TR1	952	389	234	**1575**
				2 - TR2	364	320	80	**764**
				3 - TR3	629	540	233	**1402**
				4 - TR4	602	422	154	**1178**
				5 - TR5	619	367	205	**1191**
				6 - TR6	371	479	443	**1293**
				7 - TR7	242	254	256	**752**
				8 - TR8	283	322	170	**775**
				9 - TR9	240	285	172	**697**
				10 - TRA	79	136	336	**551**
				11 - TRB	167	161	344	**672**
				12 - TRC	165	156	734	**1055**

Birinci kümede bulunan haneler incelendiğinde, hanede bulunan kişi sayısı 4, hane toplam aylık gelir ortalaması 6.671 TL, dizüstü bilgisayar ve diğer cihazlara sahip, sabit geniş ve mobil geniş bağlantı kullanan, Marmara Bölgesi (İstanbul, Batı Marmara, Doğu Marmara), Ege ve Batı Anadolu Bölgeleri'nde yaşamakta oldukları sonucuna ulaşılmıştır.

İkinci kümede bulunan haneler analiz edildiğinde, hanede bulunan kişi sayısı 2, hane toplam aylık gelir ortalaması 3.079 TL, herhangi bir bilişim ekipmanına sahip olmayan, mobil geniş bağlantı kullanan, Akdeniz ve Karadeniz (Batı Karadeniz ve Doğu Karadeniz) Bölgeleri'nde hane sahibi oldukları bulunmuştur.

Üçüncü kümede bulunan hanelerin küme özellikleri incelendiğinde, hanede bulunan kişi sayısı 5, hane toplam aylık gelir ortalaması 3.037 TL, diğer cihaz sahibi, sabit geniş ve mobil geniş bağlantı kullanan, Orta Anadolu, Kuzeydoğu Anadolu, Ortadoğu Anadolu ve Güneydoğu Anadolu Bölgeleri'nde yaşadıkları sonucuna ulaşılmıştır.

Tüm bulgular birlikte değerlendirildiğinde, birinci kümede yer alan hanelerin, hane toplam aylık geliri ortalaması diğer iki kümeden yüksektir. Bu haneler dizüstü bilgisayar ve diğer cihazlara sahip olmakla birlikte, sabit geniş ve mobil geniş internet bağlantısı kullanmaktadır. Birinci kümede bulunan haneler, Türkiye'nin batı bölgelerinde yer almaktadır.

İkinci kümede yer alan hanelerin, hanede bulunan birey sayısı düşük, hane toplam aylık geliri ortalaması düşük, bilişim ekipmanlarına sahip olmayan, mobil internet bağlantısı kullanan, Türkiye'nin kuzey ve güneyde yer alan bölgelerinde yaşamakta oldukları belirlenmiştir.

Üçüncü kümede yer alan hanelerin, hane toplam aylık geliri ortalaması düşüktür. Bununla birlikte, hanede diğer cihaz olarak isimlendirilen, akıllı TV, akıllı hoparlör, oyun konsulu, e-kitap okucuyu, akıllı saat ekipmanlarından bir veya bir kaçına sahiptirler. Bu haneler sabit geniş ve mobil geniş internet bağlantısı kullanmaktadır. Bu haneler, Türkiye'nin doğusunda ve ortasında yer alan bölgelerimizde yaşamaktadırlar.

İkinci ve üçüncü kümenin özellikleri karşılaştırıldığında, hane ortalama gelirilerinin benzer değerlerde olduğu, bunun yanında hanede bulunan kişi sayısında belirgin bir fark olduğu görülmektedir. Birinci kümede yer alan haneler ise, diğer iki küme bulunan hanelerden, hane toplam aylık geliri ortalamasına ve dizüstü bilgisayara sahip olma özellikleriyle ayrılmaktadır.

4.4.3 2019 – 2021 Senelerine Ait Hanehalkı Bulgularının Karşılaştırılması

İki sene arasında kümeler arası farklılıklar incelendiğinde, 2019 senesinde ülkeminiz doğusunda yer alan, hanehalkı birey sayısı en yüksek, hane aylık gelir ortalaması düşük olan hanelere, 2021 senesinde Orta Anadolu Bölgesi'nin dahil olduğu gözlemlenmiştir. 2019 senesinde ülkemizin kuzey doğu ve güneyinde yer alan, hanehalkı birey sayısı en düşük, hane aylık geliri düşük olan hanelere, 2021 senesinde Batı Karadeniz Bölgesi dahil olmuştur. 2021 ülkemizin batısında yer alan, hane aylık gelir ortalaması en yüksek olan hanelerin bir kümede toplandığı belirlenmiştir. Bu sonuca göre 2021 senesinde kümeler ülkemizin batısında bulunana haneler bir kümede (İstanbul, Batı

Marmara, Doğu Marmara, Ege ve Batı Anadolu), kuzey ve güneyinde bulunan haneler bir kümede (Batı Karadeniz, Doğu Karadeniz ve Akdeniz), orta ve doğusunda bulunan haneler (Kuzeydoğu Anadolu, Ortadoğu Anadolu, Güneydoğu Anadolu ve Orta Anadolu) bir kümede olacak şekilde gruplanmıştır. Bu şekilde bölgeler arasında daha homojen bir dağılım gözlenmiştir. Bu değişiklik doğrultusunda göze çarpan farklılık, orta ve doğu bölgelerde bulunan hanelerin 2021 senesinde sabit geniş internet bağlantısı kullanım oranının artmasıdır. Bir diğer farklılık ise, 2019 senesinde hanede bulunan kişi sayısı 3, hane toplam aylık gelir ortalaması 5.357 TL, birden çok bilişim ekipmanına sahip, Marmara Bölgesi, Ege, Batı Anadolu, Orta Anadolu ve Batı Karadeniz Bölgeleri'nde yaşamakta hanelerde, 2021 senesinde tablete sahip olmanın kümeler arası ayrım sebebi olmaktan çıkmasıdır.

4.5. Fert Veri Setlerine Ait Bulgular

2019 ve 2021 senelerine ait Hanehalkı Bilişim Teknolojileri Kullanımı Araştırması Mikro Veri Setleri, fert düzeyinde KAMILA algoritması ile sınıflandırılmış ve ulaşılan sonuçlar 2019 senesine ait bulgular, 2021 senesine ait bulgular ve 2019-2021 seneleri karşılaştırması olarak üç başlıkta toplanmıştır. Veri setinde bulunan tek sayısal değişken ferdin yaşını belirten yaş değişkenidir. Fert veri seti ile temelde iki sorunun cevabı araştırılmıştır. Bunlardan ilki internet üzerinden mal ve hizmet alışverişi yapan bireylerin yaş ve cinsiyet ve e-ticaret alışkanlıklarına göre gözlemlerin kümelenmesidir. İkinci olarak fertlerin, taşınabilir cihazlar ile internetteki faaliyetlerinin, yaş, internet kullanım sıklığı, eğitim durumları ve meslek gruplarına göre sınıflandırılmasıdır.

4.5.1. 2019 Senesine Ait Bulgular – Analiz 1

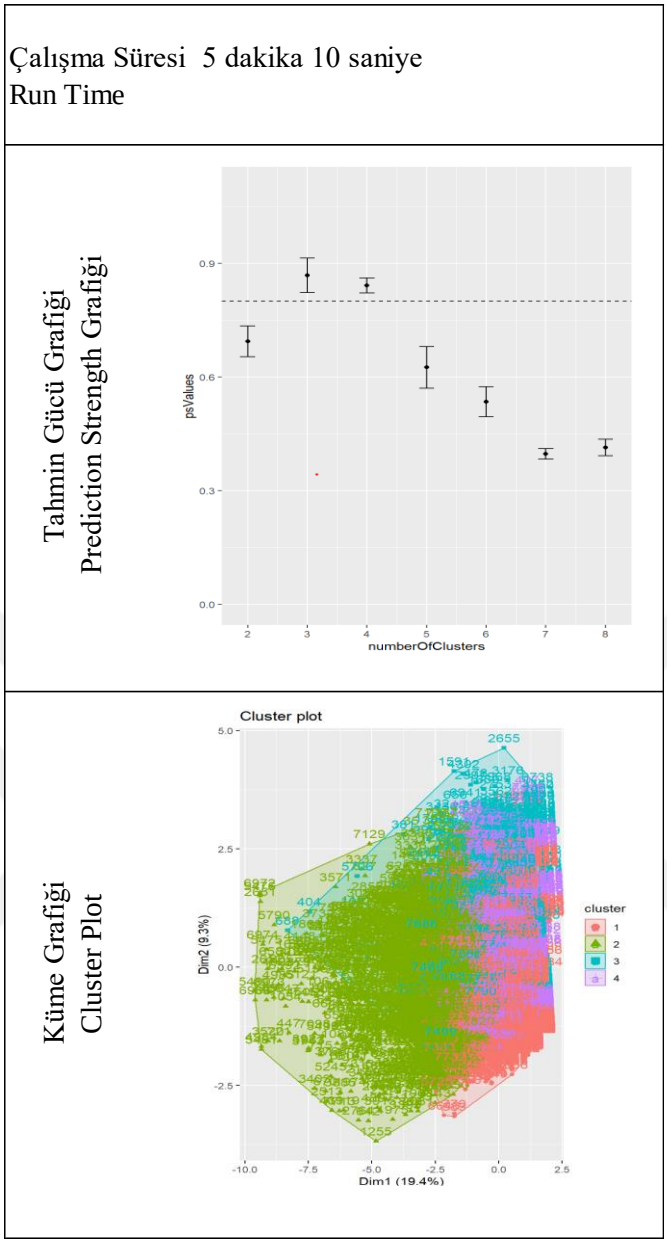
2019 senesine ait fert bilgileri original veri setinde toplam 45 060 gözlem bulunmaktadır. Bu gözlemlerden cevaplılık durumu evet olan (28 675) gözlemlerden, son 12 ay içerisinde internet üzerinden mal ve hizmet alan fertlere (7 906) ait gözlemler analize dahil edilmiştir. 7 906 gözlem üzerinden yapılan hesaplamada yaş değişkeni için, ortalama 33.06 ve standard sapma 10.89 olarak hesaplanmıştır. İlk soruya konu olan veri setine ait değişkenlerin bilgileri Tablo 4.6'da verilmektedir.

Tablo 4.6: 2019 Senesine Ait Fert Verisi Analiz 1 Değişkenler

	Değişken No	Değişken Adı	Değişken Açıklama / Seçenekleri	Değişken Tipi	Seviyeler	Sıklık	Oran %
	1	Yaş	Ferdin yaşı	Sayısal			
	2	Cinsiyet	Ferdin cinsiyeti	Kategorik	1 - Erkek	4233	0.54
					2 - Kadın	3673	0.46
Son 12 Ay İçerisinde İnternet Üzerinden Alınan Mal veya Hizmet Türleri	3	Eticaret_Tur_Giyim	Giyim ve spor malzemeleri	Kategorik	1 - Evet	5337	0.68
					2 - Hayır	2569	0.32
	4	Eticaret_Tur_Evesyası	Ev eşyası (mobilya, oyuncak, beyaz eşya vb., tüketici elektroniği hariç)	Kategorik	1 - Evet	2144	0.27
					2 - Hayır	5762	0.73
	5	Eticaret_Tur_KitapDergi	e- kitap dahil	Kategorik	1 - Evet	1607	0.20
					2 - Hayır	6299	0.80
	6	Eticaret_Tur_Bilg_Donanım	(modem, yazıcı ve diğer donanımlar)	Kategorik	1 - Evet	906	0.11
					2 - Hayır	7000	0.89
	7	Eticaret_Tur_Elektronik_Arac	Elektronik araçlar(cep telefonu, kamera, radyo, TV, DVD oynatıcı, video vb.)	Kategorik	1 - Evet	1563	0.20
					2 - Hayır	6343	0.80
	8	Eticaret_Tur_İlaç	İlaç	Kategorik	1 - Evet	325	0.04
					2 - Hayır	7581	0.96
	9	Eticaret_Tur_Gıda	Gıda maddeleri ile günlük gereksinimler (çiçek, kozmetik, tütün ve içecekler dahil)	Kategorik	1 - Evet	2000	0.25
					2 - Hayır	5906	0.75
	10	Eticaret_Tur_Telekom_hizmet	Telekomünikasyon hizmetleri (TV, internet abonelik hizmetleri (ADSL vb.), sabit veya cep telefonu abonelikleri, ön ödemeli telefon	Kategorik	1 - Evet	1154	0.15
					2 - Hayır	6752	0.85
	11	Eticaret_Tur_Seyahat	Seyahat ile ilgili diğer işlemler (ulaşım için bilet ayırma, araç kiralama vb.)	Kategorik	1 - Evet	2395	0.30
					2 - Hayır	5511	0.70
	12	Eticaret_Tur_Konaklama	Tatil konaklaması (otel vb. rezervasyonlar)	Kategorik	1 - Evet	1121	0.14
					2 - Hayır	6785	0.86
	13	Eticaret_Tur_BiletAlım	Sportif ve kültürel faaliyetler için bilet satın alımı (sinema, tiyatro, konser, maç vb.)	Kategorik	1 - Evet	1275	0.16
					2 - Hayır	6631	0.84
	14	Eticaret_Tur_FilmMuzik	Film ve müzik	Kategorik	1 - Evet	621	0.08
					2 - Hayır	7285	0.92
	15	Eticaret_Tur_eOgrenme	e-öğrenme araçları (görsel - işitsel materyaller, çevrimiçi öğrenme yazılımı, elektronik ders kitapları vb.)	Kategorik	1 - Evet	266	0.03
					2 - Hayır	7640	0.97
	16	Eticaret_Tur_OyunBilgi_Yzlm	Oyun yazılımı, diğer bilgisayar yazılımı ve yazılım güncellemeleri	Kategorik	1 - Evet	492	0.06
					2 - Hayır	7414	0.94
	17	Eticaret_Tur_Diger	Diğer	Kategorik	1 - Evet	253	0.03
					2 - Hayır	7653	0.97

Analiz sonuçlarına ait tahmin gücü grafiği Şekil 4.4'de verilmiştir. Buna göre verinin 4 kümeye ayrılması uygun gözükmektedir.

Kamila algoritmasının çalışma süresi, 17 değişken, 7 906 gözlem ve 32 seviyeli veri setinde 5 dakika 10 saniye olmuştur.



Şekil 4.4: 2019 Senesine Ait Fert Verisi Kümeleme Sonuçlarına Ait Görseller – Analiz 1

Tablo 4.7'deki kümeleme sonuçlarına göre, birinci kümede yer alan fertlerin yaş ortalaması 23, çoğunlukla kadınlardan oluşan, giyim alışverişi yapan bireylerden oluştuğu bulunmuştur. İkinci kümede yer alan bireylerin yaş ortalaması 33, çoğunlukla erkek bireylerden oluşan, giyim, gıda, seyahat, konaklama ve bilet alım alışkanlıklarına sahip bireylerden oluştuğu gözlemlenmiştir. Üçüncü kümede yer alan bireylerin yaş ortalaması 36, çoğunlukla erkek ve e-ticaret hizmetlerinden faydalanmayan bireylerin çoğunlukta olduğu söylenebilmektedir. Dördüncü kümenin yaş ortalaması 52, çoğunlukla erkek bireylerden oluşan ve giyim alışverişi yapan bireylerden oluştuğunu söylemek mümkündür.

Bu sonuçlara göre,üçüncü kümede ve dördüncü kümede yer alan fertlerin yaş ortalaması ve giyim alışverişi yapıp yapmama durumlarına göre farklılaşmakta oldukları, diğer mal ve hizmet türleri satın alma alışkanlıklarında benzerlik gösterdikleri anlaşılmaktadır.

Tablo 4.7: 2019 Senesine Ait Fert Verisi Analiz 1 Kümeleme Sonuçları

	Değişken No	Değişken Adı	Değişken Tipi	Seviyeler	Küme 1	Küme 2	Küme 3	Küme 4	**Toplam**
	1	Yaş	Sayısal						
	2	Cinsiyet	Kategorik	1 - Erkek	1285	936	726	1286	**4233**
				2 - Kadın	1619	495	355	1204	**3673**
Son 12 Ay İçerisinde İnternet Üzerinden Alınan Mal veya Hizmet Türleri	3	Eticaret_Tur_Giyim	Kategorik	1 - Evet	2118	1166	527	1526	**5337**
				2 - Hayır	786	265	554	964	**2569**
	4	Eticaret_Tur_Evesyası	Kategorik	1 - Evet	334	836	263	711	**2144**
				2 - Hayır	2570	595	818	1779	**5762**
	5	Eticaret_Tur_KitapDergi	Kategorik	1 - Evet	406	734	161	306	**1607**
				2 - Hayır	2498	697	920	2184	**6299**
	6	Eticaret_Tur_Bilg_Donanım	Kategorik	1 - Evet	136	516	116	138	**906**
				2 - Hayır	2768	915	965	2352	**7000**
	7	Eticaret_Tur_Elektronik_Arac	Kategorik	1 - Evet	326	584	229	424	**1563**
				2 - Hayır	2578	847	852	2066	**6343**
	8	Eticaret_Tur_İlaç	Kategorik	1 - Evet	42	158	49	76	**325**
				2 - Hayır	2862	1273	1032	2414	**7581**
	9	Eticaret_Tur_Gıda	Kategorik	1 - Evet	520	772	210	498	**2000**
				2 - Hayır	2384	659	871	1992	**5906**
	10	Eticaret_Tur_Telekom_hizmet	Kategorik	1 - Evet	217	611	118	208	**1154**
				2 - Hayır	2687	820	963	2282	**6752**
	11	Eticaret_Tur_Seyahat	Kategorik	1 - Evet	427	1192	377	399	**2395**
				2 - Hayır	2477	239	704	2091	**5511**
	12	Eticaret_Tur_Konaklama	Kategorik	1 - Evet	64	798	126	133	**1121**
				2 - Hayır	2840	633	955	2357	**6785**
	13	Eticaret_Tur_BiletAlım	Kategorik	1 - Evet	198	912	96	69	**1275**
				2 - Hayır	2706	519	985	2421	**6631**
	14	Eticaret_Tur_FilmMuzik	Kategorik	1 - Evet	91	495	20	15	**621**
				2 - Hayır	2813	936	1061	2475	**7285**
	15	Eticaret_Tur_eOgrenme	Kategorik	1 - Evet	32	201	15	18	**266**
				2 - Hayır	2872	1230	1066	2472	**7640**
	16	Eticaret_Tur_OyunBilgi_Yzlm	Kategorik	1 - Evet	106	338	18	30	**492**
				2 - Hayır	2798	1093	1063	2460	**7414**
	17	Eticaret_Tur_Diger	Kategorik	1 - Evet	82	17	53	101	**253**
				2 - Hayır	2822	1414	1028	2389	**7653**

4.5.2. 2019 Senesine Ait Bulgular – Analiz 2

Fert veri setindeki gözlemler kullanılarak ulaşılmak istenen ikinci soru fertlerin yaş, eğitim durumu, sahip oldukları meslek, son üç ay içinde ortalama hangi sıklıkla internet kullandıkları, son üç ay içinde ev ve işyeri dışında internete bağlanmak için taşınabilir cihazlardan hangilerini kullandıkları ve son üç ay içinde kişisel amaçla internet kullanılarak yaptıkları faaliyetlere göre sınıflandırılmasıdır.

2019 senesine ait fert bilgileri original veri setinde toplam 45 060 gözlem bulunmaktadır. Bu gözlemlerden cevaplılık durumu evet olan (28 675) gözlemlerden, internet ile yaptığı faaliyetler hakkında olumlu veya olumsuz bilgi veren (20 316) gözlem analize dahil edilmiştir. 20 316 gözlem üzerinden yapılan hesaplamada yaş değişkeni için ortalama 37.19 ve standart sapma 13.21 olarak hesaplanmıştır.

Analiz öncesi meslek sınıflaması için veride düzenleme yapılmış ve 4 haneli kodun ilk rakamı bir seviyeyi göstermek üzere alt başlıklar birleştirilmiştir. Örnek olarak 1323 – inşaat müdürleri için meslek kodu 1 olarak belirlenmiştir. Buna göre 1'den 9'a kadar gruplar oluşturulmuş, meslek kodu 10 ise 22-28 Mart tarihleri arasında çalışmayan veya geri dönebileceği bir işyeri olmayan fertleri kapsayacak şekilde düzenlenmiştir.

Tablo 4.8'de meslek kodlarına ait açıklamalara yer verilmiş ve ilgili açıklamalar 2019 senesine ait Hanehalkı Bilişim Teknolojileri Kullanım İstatistikleri Anket Soru Formu'ndan alınmıştır. Bu form kurum tarafından veri setleri ile birlikte araştırmacıya sağlanmıştır. Bu konuda 2019 ve 2021 senelerinde yapılan anketlerdeki uygulama farklıdır. '*Çalıştığınız işyerindeki veya işinizdeki görev ve sorumluluklarınıza en uygun seçenek'* sorusu için 2019 senesine ait soru formunda seviyelerin karşılıkları olan görev ve sorumluluklar belirtilmişken, 2021 senesine ait soru formunda 4 haneli uluslararası meslek sınıflamasına ait kod, 436 seçenek içerisinden belirlenip not edilmiştir. İlgili soru için Tablo 4.9'da belirtilen seviyeler 2019 senesine ait seviyelerdir ve 2021 senesi için geçerli değildir.

Tablo 4.8: 2019 Senesinde Fertlerin Görev ve Sorumluluklarına Ait Kodlar ve Karşılık Gelen Açıklamalar

ISCO08_Meslek
0 - Silahlı kuvvetlerle ilgili meslekler
1 - Yöneticiler
2 - Bilim ve mühendislik alanlarındaki profesyonel meslek mensupları veya Sağlık profesyonelleri veya Eğitim ile ilgili profesyonel meslek mensupları veya İş ve yönetim ile ilgili profesyonel meslek mensupları veya Bilgi ve iletişim teknolojileri ile ilgili profesyonel meslek mensupları veya Hukuk, sosyal bilimler ve kültür ile ilgili profesyonel meslek mensupları
3 - Bilim ve mühendislik ile ilgili yardımcı profesyonel meslek mensupları veya Yardımcı sağlık profesyonelleri veya İş ve idare ile ilgili yardımcı profesyonel meslek mensupları veya Hukuk, sosyal, kültür ve benzeri alanlar ile ilgili yardımcı profesyonel meslek mensupları veya Bilgi ve iletişim teknisyenleri
4 - Büro hizmetlerinde çalışan elemanlar
5- Hizmet ve satış elemanları
6- Nitelikli tarım, ormancılık ve su ürünleri çalışanları
7 - Sanatkarlar ve ilgili işlerde çalışanlar
8 - Tesis ve makine operatörleri ve montajcılar
9 - Nitelik gerektirmeyen işlerde çalışanlar
10 - 25 Mart - 01 Nisan 2019 tarihleri arasında mal ya da nakdi (para) gelir elde etmek amacıyla bir işte çalışmayanlar

2019 senesine internet kullanım sıklığı ve fertlerin eğitim durumunu belirten seviyelerin karşılıkları Tablo 4.9'da listelenmektedir.

Tablo 4.9: Analiz 2'de Dahil Edilen İki Değişkene Ait Seviyeler ve Karşılıkları

Kullanım_Sıklık_İnternet
6 - hemen her gün
9 - haftada bir defadan az (iki üç haftada bir)
13 - haftada en az bir defa
Okul_Biten
2 - ilkokul veya bir okul bitirmedi
3 - Genel ortaokul/ Mesleki veya teknik ortaokul/ İlköğretim
4 - Genel lise/ Mesleki veya teknik lise
5- 2,3 veya 4 yıllık yüksekokul/ fakülte veya Yüksek lisans veya Doktora

2019 senesi için Analiz 2'ye dahil edilen 28 değişkenin bilgileri Tablo 4.10'da verilmektedir. Bu değişkenlerden değişken 2'e ait 3 seviye, değişken 3'e ait 4 seviye ve değişken 5'e ait 11 seviye bulunmaktadır.

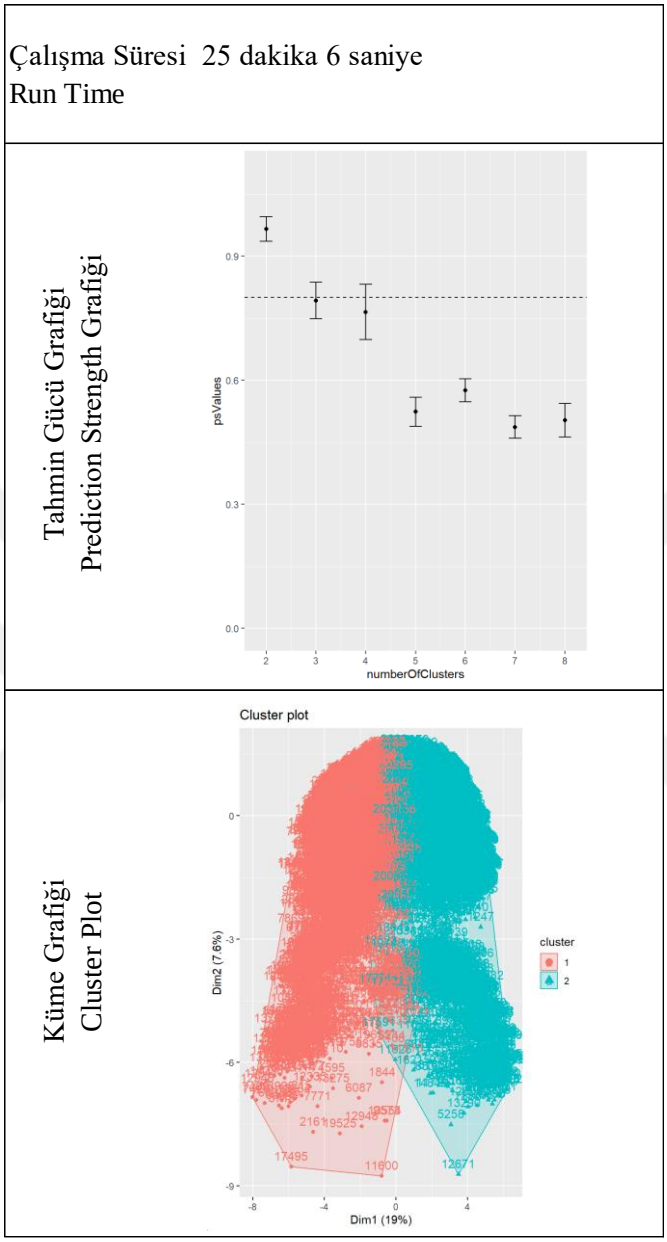
Tablo 4.10: 2019 Senesine Ait Fert Verisi Analiz 2 Değişken Bilgileri

	Değişken No	Değişken Adı	Değişken Açıklama / Seçenekleri	Değişken Tipi	Seviyeler	Sıklık	Oran %
	1	Yaş	Ferdin yaşı	Sayısal			
	2	Kullanım_Sıklık_İnternet	Son üç ay içinde ortalama internet kullanım sıklığı	Kategorik	6	18209	0.90
					9	400	0.02
					13	1707	0.08
	3	Okul_Biten	Tamamlanan en son okul / en yüksek eğitim seviyesi	Kategorik	2	5776	0.28
					3	4380	0.22
					4	5187	0.26
					5	4973	0.24
	4	ISCO08_Meslek	Çalıştığınız işyerindeki/ işinizdeki görev ve sorumluluklarınıza en uygun seçenek	Kategorik	0	242	0.01
					1	384	0.02
					2	1581	0.08
					3	260	0.01
					4	1146	0.06
					5	2042	0.10
					6	499	0.02
					7	711	0.03
					8	518	0.03
					9	2832	0.14
					10	10101	0.50
Son üç ay içinde ev ve işyeri dışında internete bağlanmak için kullanılan taşınabilir cihazlar	5	Mobil_Int_Cep	Cep telefonu	Kategorik	1 - evet	19532	0.96
					2 - hayır	784	0.04
	6	Mobil_Int_Tasınabilir	Taşınabilir bilgisayar (dizüstü, netbook vb.)	Kategorik	1 - evet	4538	0.22
					2 - hayır	15778	0.78
	7	Mobil_Int_Tablet	Tablet	Kategorik	1 - evet	1888	0.09
					2 - hayır	18428	0.91
	8	Mobil_Int_Diger_Cihaz	Diğer cihazlar (taşınabilir oyun konsolu, e-kitap okuyucu, akıllı saat vb.)	Kategorik	1 - evet	274	0.01
					2 - hayır	20042	0.99
	9	Mobil_Int_Hic	Ev ve iş yeri dışında herhangi bir mobil cihaz ile internete bağlanmadım	Kategorik	1 - evet	599	0.03
					2 - hayır	19717	0.97

Tablo 4.10: 2019 Senesine Ait Fert Verisi Analiz 2 Değişken Bilgileri - Devam

	Değişken No	Değişken Adı	Değişken Açıklama / Seçenekleri	Değişken Tipi	Seviyeler	Sıklık	Oran %
Son üç ay içinde kişisel amaçla internet üzerinden faaliyetler	10	Int_Faal_ePosta	e-posta gönderme/ alma	Kategorik	1 - evet	8849	0.44
					2 - hayır	11467	0.56
	11	Int_Faal_Telefon	İnternet üzerinden sesli veya görüntülü arama yapmak (Skype, Messenger, WhatsApp, BİP,Facetime,Viber vb. kullanarak)	Kategorik	1 - evet	16750	0.82
					2 - hayır	3566	0.18
	12	Int_Faal_Sosyal_GrKatılım	Sosyal medya (Facebook, Twitter, Instagram vb.) üzerinde profil oluşturma, mesaj gönderme veya fotoğraf vb. içerik paylaşma	Kategorik	1 - evet	16242	0.80
					2 - hayır	4074	0.20
	13	Int_Faal_Mesaj	Mesajlaşma (WhatsApp, Messenger, Skype, BİP, Viber vb.)	Kategorik	1 - evet	18975	0.93
					2 - hayır	1341	0.07
	14	Int_Faal_Online_Haber	Çevrimiçi haber sitelerini / gazeteleri / haber dergilerini okumak	Kategorik	1 - evet	13887	0.68
					2 - hayır	6429	0.32
	15	Int_Faal_Saglık_Bilgi_arm	Sağlıkla ilgili bilgi arama (yaralanmalar, hastalıklar, beslenme, sağlığın iyileştirilmesi gibi)	Kategorik	1 - evet	13953	0.69
					2 - hayır	6363	0.31
	16	Int_Faal_MalHiz_Bilgi	Mal ve hizmetler hakkında bilgi arama	Kategorik	1 - evet	12906	0.64
					2 - hayır	7410	0.36
	17	Int_Faal_Website_Paylasım	Web siteleri aracılığıyla (bloglar, facebook, twitter vb. sosyal ağlar) toplumsal veya siyasal konular ile ilgili görüşleri paylaşma	Kategorik	1 - evet	4551	0.22
					2 - hayır	15765	0.78
	18	Int_Faal_Oyl_Ktlm	Toplumsal veya siyasal bir konuda online bir oylamaya katılma	Kategorik	1 - evet	1697	0.08
					2 - hayır	18619	0.92
	19	Int_Faal_IsArama	İş arama ya da iş başvurusu yapma	Kategorik	1 - evet	1897	0.09
					2 - hayır	18419	0.91
	20	Int_Faal_Web_Icerik_yukleme	Kendi oluşturduğunuz metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleme	Kategorik	1 - evet	8395	0.41
					2 - hayır	11921	0.59
	21	Int_Faal_Muzik_Dinlm	Müzik dinlemek (Spotify, web radyosu vb.)	Kategorik	1 - evet	14120	0.70
					2 - hayır	6196	0.30
	22	Int_Faal_MalHizmet_Satıs	Mal veya hizmet satışı (gittigidiyor, sahibinden, letgo vb.)	Kategorik	1 - evet	4357	0.21
					2 - hayır	15959	0.79
	23	Int_Faal_Banka_Islem	İnternet bankacılığı	Kategorik	1 - evet	9142	0.45
					2 - hayır	11174	0.55
	24	Int_Faal_Bulut_Depo_Kullanım	Resim, müzik, video veya dosya gibi kişisel dokümanları internetteki bir alanda depolama (Google Drive, iCloud, Dropbox,	Kategorik	1 - evet	4816	0.24
					2 - hayır	15500	0.76
Son üç ay içinde internet üzerinden katılım sağlanan eğitim faaliyetleri	25	Int_Faal_Egt_Kurs	Çevrimiçi (Online) bir kurs alma	Kategorik	1 - evet	620	0.03
					2 - hayır	19696	0.97
	26	Int_Faal_Ogrn_Mtryl	Çevrimiçi (Online) öğrenme materyallerini (Görsel-işitsel materyaller, çevrimiçi öğrenme yazılımı, elektronik ders kitapları)	Kategorik	1 - evet	797	0.04
					2 - hayır	19519	0.96
	27	Int_Faal_Egt_Web	Web sitesi/ portal üzerinden eğitmen ve öğrencilerle iletişime geçme	Kategorik	1 - evet	603	0.03
					2 - hayır	19713	0.97

Kamila algoritmasının çalışma süresi 27 değişken, 20 316 gözlem ve 64 seviyeli veri setinde 25 dakika, 6 saniye olmuştur.



Şekil 4.5: 2019 Senesine Ait Fert Verisi Kümeleme Sonuçlarına Ait Görseller – Analiz 2

Şekil 4.5'e göre küme sayısı 2 olduğunda tahmin gücü 0.80'in üzerinde olmaktadır.

Birinci kümede yer alan bireylerin yaş ortalaması 33, interneti her gün kullandıkları, eğitim düzeyi yüksek, 0, 1, 2, 3, 4 ve 5. meslek gruplarında çalışan, kişisel amaçla internet faaliyetlerini etkin kullanan bireylerden oluştuğu değerlendirilmiştir. İkinci kümede yer alan bireylerin yaş ortalaması 41, internet her gün veya daha az kullanan, yüksek eğitim seviyesine sahip olmayan, 6, 7, 8, ve 9. meslek

gruplarında çalışan veya çalışacak bir işi olmayan, interneti sesli veya görüntülü arama yapma, sosyal medyaya katılım sağlama ve mesajlaşma faaliyetleri için kullanan bireylerden oluştuğu belirlenmiştir.

Tablo 4.11: 2019 Senesine Ait Fert Bilgileri Analiz 2 Kümeleme Sonuçları

	Değişken No	Değişken Adı	Değişken Tipi	Seviyeler	Küme 1	Küme 2	**Toplam**
	1	Yaş	Sayısal				
	2	Kullanım_Sıklık_İnternet	Kategorik	6	10377	7832	**18209**
				9	4	396	**400**
				13	67	1640	**1707**
	3	Okul_Biten	Kategorik	2	692	5084	**5776**
				3	1864	2516	**4380**
				4	3428	1759	**5187**
				5	4464	509	**4973**
	4	ISCO08_Meslek	Kategorik	0	226	16	**242**
				1	336	48	**384**
				2	1531	50	**1581**
				3	242	18	**260**
				4	972	174	**1146**
				5	1281	761	**2042**
				6	97	402	**499**
				7	372	339	**711**
				8	310	208	**518**
				9	1134	1698	**2832**
				10	3947	6154	**10101**
Son üç ay içinde ev ve işyeri dışında internete bağlanmak için kullanılan taşınabilir cihazlar	5	Mobil_Int_Cep	Kategorik	1 - evet	10348	9184	**19532**
				2 - hayır	100	684	**784**
	6	Mobil_Int_Tasınabilir	Kategorik	1 - evet	3980	558	**4538**
				2 - hayır	6468	9310	**15778**
	7	Mobil_Int_Tablet	Kategorik	1 - evet	1629	259	**1888**
				2 - hayır	8819	9609	**18428**
	8	Mobil_Int_Diger_Cihaz	Kategorik	1 - evet	253	21	**274**
				2 - hayır	10195	9847	**20042**
	9	Mobil_Int_Hic	Kategorik	1 - evet	63	536	**599**
				2 - hayır	10385	9332	**19717**

Tablo 4.11: 2019 Senesine Ait Fert Verisi Analiz 2 Kümeleme Sonuçları – Devam

	Değişken No	Değişken Adı	Değişken Tipi	Seviyeler	Küme 1	Küme 2	**Toplam**
Son üç ay içinde kişisel amaçla internet üzerinden faaliyetler	10	Int_Faal_ePosta	Kategorik	1 - evet	8091	758	**8849**
				2 - hayır	2357	9110	**11467**
	11	Int_Faal_Telefon	Kategorik	1 - evet	9663	7087	**16750**
				2 - hayır	785	2781	**3566**
	12	Int_Faal_Sosyal_GrKatılım	Kategorik	1 - evet	9641	6601	**16242**
				2 - hayır	807	3267	**4074**
	13	Int_Faal_Mesaj	Kategorik	1 - evet	10376	8599	**18975**
				2 - hayır	72	1269	**1341**
	14	Int_Faal_Online_Haber	Kategorik	1 - evet	9467	4420	**13887**
				2 - hayır	981	5448	**6429**
	15	Int_Faal_Saglık_Bilgi_arm	Kategorik	1 - evet	9312	4641	**13953**
				2 - hayır	1136	5227	**6363**
	16	Int_Faal_MalHiz_Bilgi	Kategorik	1 - evet	9414	3492	**12906**
				2 - hayır	1034	6376	**7410**
	17	Int_Faal_Website_Paylasım	Kategorik	1 - evet	3504	1047	**4551**
				2 - hayır	6944	8821	**15765**
	18	Int_Faal_Oyl_Ktlm	Kategorik	1 - evet	1531	166	**1697**
				2 - hayır	8917	9702	**18619**
	19	Int_Faal_IsArama	Kategorik	1 - evet	1609	288	**1897**
				2 - hayır	8839	9580	**18419**
	20	Int_Faal_Web_Icerik_yukleme	Kategorik	1 - evet	6106	2289	**8395**
				2 - hayır	4342	7579	**11921**
	21	Int_Faal_Muzik_Dinlm	Kategorik	1 - evet	9289	4831	**14120**
				2 - hayır	1159	5037	**6196**
	22	Int_Faal_MalHizmet_Satıs	Kategorik	1 - evet	3742	615	**4357**
				2 - hayır	6706	9253	**15959**
	23	Int_Faal_Banka_Islem	Kategorik	1 - evet	7657	1485	**9142**
				2 - hayır	2791	8383	**11174**
	24	Int_Faal_Bulut_Depo_Kullanım	Kategorik	1 - evet	4256	560	**4816**
				2 - hayır	6192	9308	**15500**
Son üç ay içinde internet üzerinden katılım sağlanan eğitim faaliyetleri	25	Int_Faal_Egt_Kurs	Kategorik	1 - evet	598	22	**620**
				2 - hayır	9850	9846	**19696**
	26	Int_Faal_Ogrn_Mtryl	Kategorik	1 - evet	733	64	**797**
				2 - hayır	9715	9804	**19519**
	27	Int_Faal_Egt_Web	Kategorik	1 - evet	564	39	**603**
				2 - hayır	9884	9829	**19713**

4.5.3. 2021 Senesine Ait Bulgular – Analiz 1

2021 senesi, fert bilgileri ile temelde iki sorunun cevabı aranmaktadır. Bunlardan ilki son üç ay içerisinde internet üzerinden mal veya hizmet alan fertlerin, yaş, cinsiyet, son üç ay içinde internet kullanılan cihazlar ve satın aldıkları mal veya hizmet türlerine göre benzerliklerin ortaya çıkarılmasıdır.

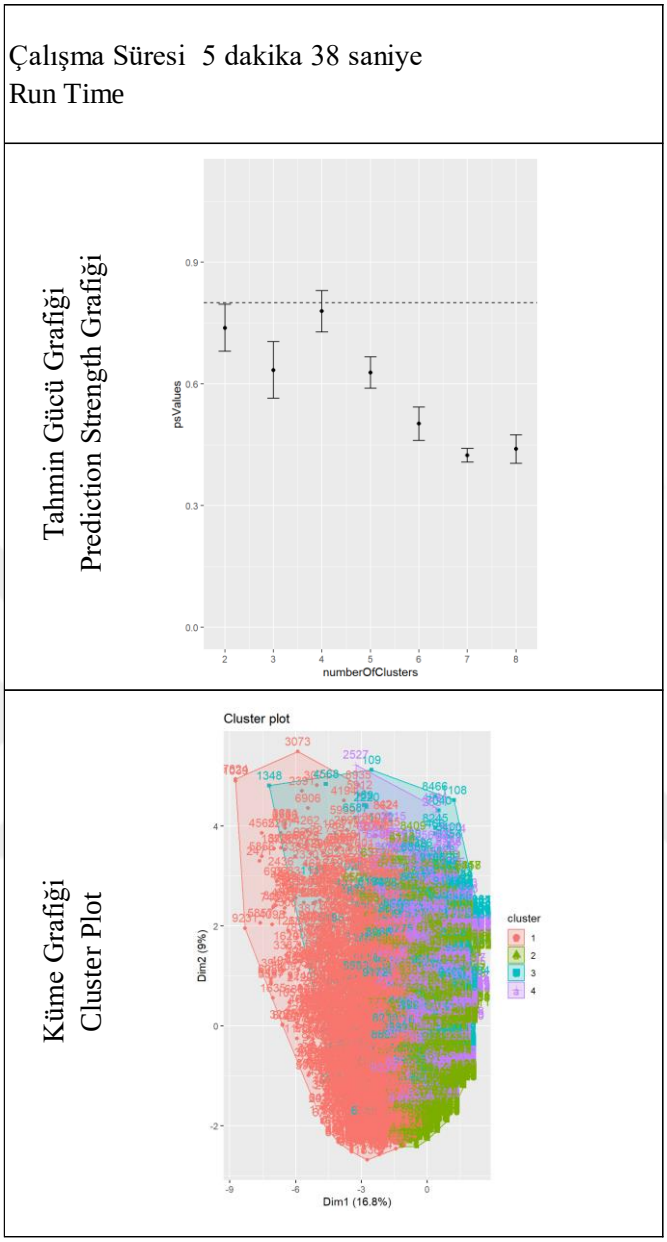
2021 senesine ait fert bilgileri original veri setinde toplam 30 530 gözlem bulunmaktadır. Bu gözlemlerden son üç ay içinde web sitesi veya mobil uygulama üzerinden mal ve hizmet alan fertlere (9 438) ait gözlemler analize dahil edilmiştir. 23 değişkene ait 9 438 gözlem üzerinden yapılan hesaplamada yaş değişkeni için ortalama 32.83 ve standart sapma 11.24 olarak hesaplanmıştır.

2021 Fert veri setine ait analize dahil edilen değişkenlerin bilgileri Tablo 4.12'de verilmektedir.

Tablo 4.12: 2021 Senesine Ait Fert Verisi Analiz 1 Değişken Bilgileri

`	Değişken No	Değişken Adı	Değişken Açıklama / Seçenekleri	Değişken Tipi	Seviyeler	Sıklık	Oran %
	1	Yaş	Ferdin yaşı	Sayısal			
	2	Cinsiyet	Ferdin cinsiyeti	Kategorik	1 - Erkek	4789	0.51
					2 - Kadın	4649	0.49
Son 3 Ay İçerisinde İnternet Üzerinden Alınan Mal veya Hizmet Türleri	3	Eticaret_Tur_Giyim	Giyim ve spor malzemeleri	Kategorik	1 - Evet	6720	0.71
					2 - Hayır	2718	0.29
	4	Eticaret_Tur_Spor_Mlz	Spor malzemeleri (spor giyim hariç)	Kategorik	1 - Evet	1260	0.13
					2 - Hayır	8178	0.87
	5	Eticaret_Tur_Cocuk_Bk_Oyn	Çocuk oyuncakları veya çocuk bakım ürünleri (çocuk bezi, biberon, bebek arabası vb.)	Kategorik	1 - Evet	1647	0.17
					2 - Hayır	7791	0.83
	6	Eticaret_Tur_Evesyası	Mobilya, ev aksesuarları (halı, perde vs.), bahçe malzemeleri (bahçe bitkileri, alet ve edavatlar vb.)	Kategorik	1 - Evet	2011	0.21
					2 - Hayır	7427	0.79
	7	Eticaret_Tur_CD_Plk	Cd, plak gibi müzik aletleri	Kategorik	1 - Evet	221	0.02
					2 - Hayır	9217	0.98
	8	Eticaret_Tur_DVD, Blu_Ray	DVD, Blu-ray vb. Film izleme ürünleri	Kategorik	1 - Evet	161	0.02
					2 - Hayır	9277	0.98
	9	Eticaret_Tur_KitapDergi	Basılı kitap, dergi, gazete	Kategorik	1 - Evet	2567	0.27
					2 - Hayır	6871	0.73
	10	Eticaret_Tur_Elktr_Arc_Aks	Bilgisayar, tablet, cep telefonu veya aksesuarları	Kategorik	1 - Evet	1646	0.17
					2 - Hayır	7792	0.83
	11	Eticaret_Tur_Elektronik_Arac	Elektronik araçlar veya beyaz eşyaları	Kategorik	1 - Evet	1345	0.14
					2 - Hayır	8093	0.86
	12	Eticaret_Tur_İlaç	İlaç veya gıda takviyeleri	Kategorik	1 - Evet	1178	0.12
					2 - Hayır	8260	0.88
	13	Eticaret_Tur_Yemek_Siparis	Lokantalardan, fast food zincirlerinden, catering şirketlerinden yapılan teslimatlar	Kategorik	1 - Evet	3517	0.37
					2 - Hayır	5921	0.63
	14	Eticaret_Tur_Hzr_Tz_Gıda	İnternetten veya doğrudan öğün kiti sağlayıcılarından alınan, yemek ve içecek hazırlamak için önceden porsiyonlanmış veya	Kategorik	1 - Evet	2823	0.30
					2 - Hayır	6615	0.70
	15	Eticaret_Tur_Kozmetik	Kozmetik, güzellik ve sağlık malzemeleri	Kategorik	1 - Evet	2624	0.28
					2 - Hayır	6814	0.72
	16	Eticaret_Tur_Tmz_Ksl_Bkm	Temizlik ürünleri, kişisel bakım malzemeleri (deterjan, temizlik bezleri, diş fırçası, hijyenik ürünler vbç)	Kategorik	1 - Evet	2445	0.26
					2 - Hayır	6993	0.74
	17	Eticaret_Tur_Motr_Arc_Ydk	Bisiklet, motosiklet (moped), araba veya diğer araçlar ile bu araçların yedek parçaları	Kategorik	1 - Evet	487	0.05
					2 - Hayır	8951	0.95
	18	Eticaret_Tur_Diger	Diğer	Kategorik	1 - Evet	226	0.02
					2 - Hayır	9212	0.98

Kamila algoritmasının çalışma süresi 18 değişken, 9 438 gözlem ve 34 seviyeli veri setinde 5 dakika, 38 saniye olmuştur.



Şekil 4.6: 2021 Senesine Ait Fert Verisi Kümeleme Sonuçlarına Ait Görseller – Analiz 1

Şekil 4.6' ya göre küme sayısı 4 olduğunda tahmin gücü eşik değeri 0.80'e en yakın değeri almaktadır. Bu sebeple küme sayısı 4 iken analiz sonuçları incelenmiştir.

Birinci kümede yer alan fertlerin yaş ortalaması 31, çoğunlukla kadın bireylerden oluşan, giyim, kitap veya dergi, hazır gıda ve gıda malzemeleri, kozmetik ve kişisel bakım ürünleri satın alan bireylerden oluştuğu belirlenmiştir. İkinci kümede yaş ortalaması 23, çoğunlukla kadın ve sadece giyim alışverişi yapan bireyler bulunmaktadır. Üçüncü küme yaş ortalaması 53, çoğunlukla erkek ve sadece

giyim alışverişi yapan bireylerden oluşmaktadır. Dördüncü kümede ise yaş ortalaması 37, erkek ve sadece giyim alışverişi yapan bireyler bulunmaktadır. Bu bilgi ışığında üçüncü ve dördüncü kümelerdeki bireyleri farklılaştıran faktörün yaş değişkeni olduğu söylenebilir.

Tablo 4.13: 2021 Senesine Ait Fert Verisi Analiz 1 Kümeleme Sonuçları

`	Değişken No	Değişken Adı	Değişken Tipi	Seviyeler	Küme 1	Küme 2	Küme 3	Küme 4	**Toplam**
	1	Yaş	Sayısal						
	2	Cinsiyet	Kategorik	1 - Erkek	901	1577	789	1522	**4789**
				2 - Kadın	1164	1745	470	1270	**4649**
Son 3 Ay İçerisinde İnternet Üzerinden Alınan Mal veya Hizmet Türleri	3	Eticaret_Tur_Giyim	Kategorik	1 - Evet	1834	2342	747	1797	**6720**
				2 - Hayır	231	980	512	995	**2718**
	4	Eticaret_Tur_Spor_Mlz	Kategorik	1 - Evet	662	246	106	246	**1260**
				2 - Hayır	1403	3076	1153	2546	**8178**
	5	Eticaret_Tur_Cocuk_Bk_Oyn	Kategorik	1 - Evet	827	231	82	507	**1647**
				2 - Hayır	1238	3091	1177	2285	**7791**
	6	Eticaret_Tur_Evesyası	Kategorik	1 - Evet	973	335	258	445	**2011**
				2 - Hayır	1092	2987	1001	2347	**7427**
	7	Eticaret_Tur_CD_Plk	Kategorik	1 - Evet	147	38	17	19	**221**
				2 - Hayır	1918	3284	1242	2773	**9217**
	8	Eticaret_Tur_DVD, Blu_Ray	Kategorik	1 - Evet	102	26	11	22	**161**
				2 - Hayır	1963	3296	1248	2770	**9277**
	9	Eticaret_Tur_KitapDergi	Kategorik	1 - Evet	1043	774	252	498	**2567**
				2 - Hayır	1022	2548	1007	2294	**6871**
	10	Eticaret_Tur_Elktr_Arc_Aks	Kategorik	1 - Evet	779	392	182	293	**1646**
				2 - Hayır	1286	2930	1077	2499	**7792**
	11	Eticaret_Tur_Elektronik_Arac	Kategorik	1 - Evet	589	243	179	334	**1345**
				2 - Hayır	1476	3079	1080	2458	**8093**
	12	Eticaret_Tur_İlaç	Kategorik	1 - Evet	738	86	176	178	**1178**
				2 - Hayır	1327	3236	1083	2614	**8260**
	13	Eticaret_Tur_Yemek_Siparis	Kategorik	1 - Evet	1667	885	329	636	**3517**
				2 - Hayır	398	2437	930	2156	**5921**
	14	Eticaret_Tur_Hzr_Tz_Gıda	Kategorik	1 - Evet	1584	435	325	479	**2823**
				2 - Hayır	481	2887	934	2313	**6615**
	15	Eticaret_Tur_Kozmetik	Kategorik	1 - Evet	1484	583	232	325	**2624**
				2 - Hayır	581	2739	1027	2467	**6814**
	16	Eticaret_Tur_Tmz_Ksl_Bkm	Kategorik	1 - Evet	1600	263	266	316	**2445**
				2 - Hayır	465	3059	993	2476	**6993**
	17	Eticaret_Tur_Motr_Arc_Ydk	Kategorik	1 - Evet	212	111	55	109	**487**
				2 - Hayır	1853	3311	1204	2683	**9051**
	18	Eticaret_Tur_Diger	Kategorik	1 - Evet	24	69	50	83	**226**
				2 - Hayır	2041	3253	1209	2709	**9212**

4.5.4. 2021 Senesine Ait Bulgular – Analiz 2

2021 senesi fert veri seti ile yanıtı araştırılmak istenen ikinci soru, fertlerin yaş, eğitim durumu, sahip oldukları meslek, son üç ay içinde ortalama hangi sıklıkla internet kullandıkları, son üç ay içinde internete bağlanmak için kullanılan araçlar ve son üç ay içinde kişisel amaçla internet kullanılarak yapılan faaliyetlere göre benzerliklerine ulaşmaktır.

2021 senesine ait fert bilgileri original veri setinde toplam 30 530 gözlem bulunmaktadır. Bu gözlemlerden internet ile yaptığı faaliyetler hakkında olumlu veya olumsuz bilgi veren (24 328) gözlem analize dahil edilmiştir. Bu gözlemler aynı zamanda son 3 ay içerisinde internet kullanımı yapan bireyleri içermektedir. Son internet kullanımı 3 aydan fazla olan bireyler, internet üzerinden yapmış oldukları faaliyeter hakkında herhangi bir bilgi vermemiştir. 24 328 gözlem üzerinden yapılan hesaplamada yaş değişkeni için ortalama 37.91 ve standart sapma 13.86 olarak elde edilmiştir.

ISCO 08 meslek sınıflaması için Uluslararası Standart Meslek Sınıflaması (ISCO 08), TÜİK Sınıflama Sunucusu'ndan elde edilmiştir.[12]

Analiz öncesi meslek sınıflaması için veride düzenleme yapılmış ve 4 haneli kodun ilk rakamı bir seviyeyi göstermek üzere alt başlıklar birleştirilmiştir. Örnek olarak 1323 – inşaat müdürleri için meslek kodu 1 olarak alınmıştır. Buna göre 1'den 9'a kadar gruplar belirlenmiş, meslek kodu 10 ise 22-28 Mart tarihleri arasında çalışmayan veya geri dönebileceği bir işyeri olmayan fertleri kapsayacak şekilde düzenlenmiştir. Analizde kullanılan değişkenlerin bilgileri Tablo 4.14'de verilmektedir.

[12] https://biruni.tuik.gov.tr/DIESS/SiniflamaSatirListeAction.do?surumId=210&seviye=4&detay=H&tu%20rId=41&turAdi=%209.%20Meslek%20S%C4%B1n%C4%B1flamalar%C4%B1

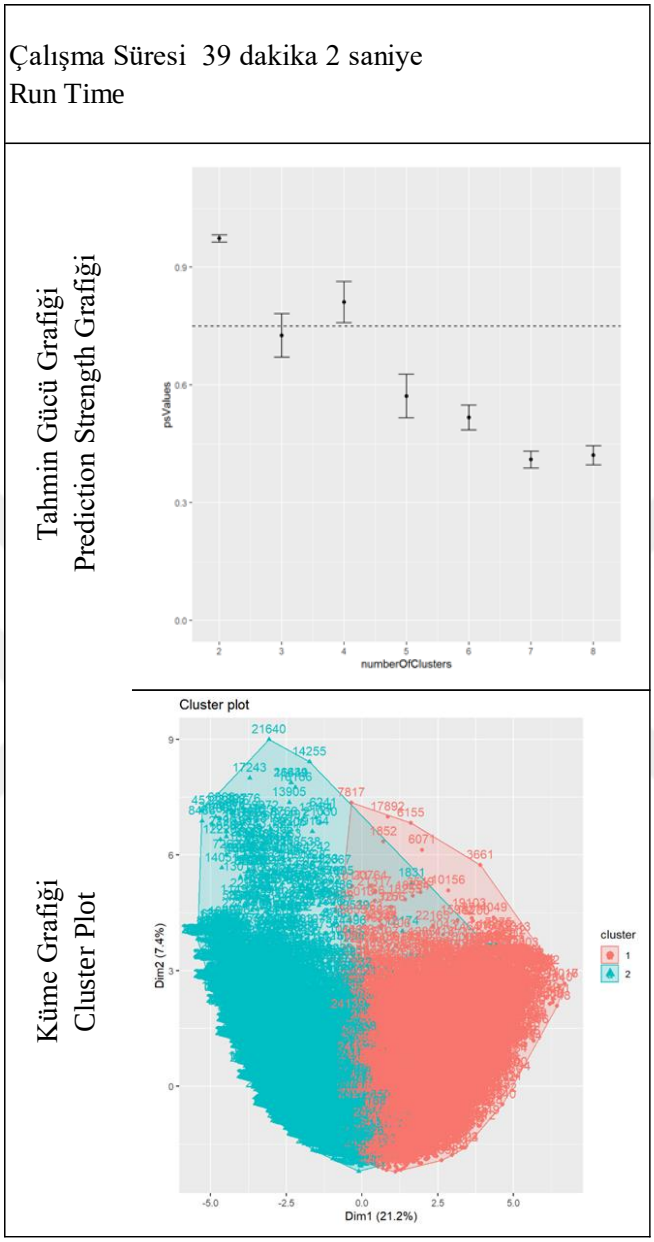
Tablo 4.14: 2021 Senesine Ait Fert Verisi Analiz 2 Değişken Bilgileri

	Değişken No	Değişken Adı	Değişken Açıklama / Seçenekleri	Değişken Tipi	Seviyeler	Sıklık	Oran %
	1	Yaş	Ferdin yaşı	Sayısal			
	2	Kullanım_Sıklık_İnternet	Son üç ay içinde ortalama internet kullanım sıklığı	Kategorik	6	22750	0.94
					9	288	0.01
					13	1290	0.05
	3	Okul_Biten	Tamamlanan en son okul / en yüksek eğitim seviyesi	Kategorik	2	6920	0.28
					3	4887	0.20
					4	6693	0.28
					5	5828	0.24
	4	ISCO08_Meslek	Çalıştığınız işyerindeki/ işinizdeki görev ve sorumluluklarınıza en uygun seçenek	Kategorik	0		
					1	793	0.03
					2	2021	0.08
					3	897	0.04
					4	813	0.03
					5	1986	0.08
					6	693	0.03
					7	1472	0.06
					8	1228	0.05
					9	1602	0.07
					10	12823	0.53
Son üç ay içinde internet kullanılan cihazlar	5	Fert_Int_Cep	Cep telefonu	Kategorik	1 - evet	24080	0.99
					2 - hayır	248	0.01
	6	Fert_Int_Dizustu	Taşınabilir bilgisayar (dizüstü, netbook vb.)	Kategorik	1 - evet	6857	0.28
					2 - hayır	17471	0.72
	7	Fert_Int_Tablet	Tablet	Kategorik	1 - evet	3097	0.13
					2 - hayır	21231	0.87
	8	Fert_Int_Masaustu	Masaüstü Bilgisayar	Kategorik	1 - evet	3317	0.14
					2 - hayır	21011	0.86
	9	Fert_Int_Diger_Cihaz	Diğer Cihazlar	Kategorik	1 - evet	4780	0.20
					2 - hayır	19548	0.80

Tablo 4.14: 2021 Senesine Ait Fert Verisi Analiz 2 Değişken Bilgileri Devam

	Değişken No	Değişken Adı	Değişken Açıklama / Seçenekleri	Değişken Tipi	Seviyeler	Sıklık	Oran %
Son üç ay içinde kişisel amaçla internet üzerinden faaliyetler	10	Int_Faal_ePosta	e-posta gönderme/ alma	Kategorik	1 - evet	10394	0.43
					2 - hayır	13934	0.57
	11	Int_Faal_Telefon	İnternet üzerinden sesli veya görüntülü arama yapma	Kategorik	1 - evet	21904	0.90
					2 - hayır	2424	0.10
	12	Int_Faal_Sosyal_GrKatılım	Sosyal medyada içerik paylaşma	Kategorik	1 - evet	17850	0.73
					2 - hayır	6478	0.27
	13	Int_Faal_Mesaj	Mesajlaşma	Kategorik	1 - evet	22553	0.93
					2 - hayır	1775	0.07
	14	Int_Faal_Online_Haber	Çevrimiçi haber sitelerini / gazeteleri / haber dergilerini okumak	Kategorik	1 - evet	15991	0.66
					2 - hayır	8337	0.34
	15	Int_Faal_Saglık_Bilgi_arm	Sağlıkla ilgili bilgi arama	Kategorik	1 - evet	16773	0.69
					2 - hayır	7555	0.31
	16	Int_Faal_MalHiz_Bilgi	Mal ve hizmetler hakkında bilgi arama	Kategorik	1 - evet	14103	0.58
					2 - hayır	10225	0.42
	17	Int_Faal_Website_Paylasım	Web siteleri aracılığıyla veya sosyal medya aracılığıyla toplumsal veya siyasal konular ile ilgili görüşleri paylaşma	Kategorik	1 - evet	2943	0.12
					2 - hayır	21385	0.88
	18	Int_Faal_Oyl_Ktlm	Toplumsal veya siyasal bir konuda online tartışma / oylamaya katılma	Kategorik	1 - evet	844	0.03
					2 - hayır	23484	0.97
	19	Int_Faal_IsArama	İş arama ya da iş başvurusu yapma	Kategorik	1 - evet	2520	0.10
					2 - hayır	21808	0.90
	20	Int_Faal_MalHizmet_Satıs	Mal veya hizmet satışı	Kategorik	1 - evet	2602	0.11
					2 - hayır	21722	0.89
	21	Int_Faal_Banka_Islem	İnternet bankacılığı (web sitesi veya mobil bankacılık uygulamaları)	Kategorik	1 - evet	13473	0.55
					2 - hayır	10891	0.45
Son üç ay içinde internet üzerinden katılım sağlanan eğitim faaliyetleri	22	Int_Faal_Egt_Kurs	Çevrimiçi (Online) bir kurs alma	Kategorik	1 - evet	2534	0.10
					2 - hayır	21794	0.90
	23	Int_Faal_Ogrn_Mtryl	Çevrimiçi (Online) öğrenme materyallerini (Görsel-işitsel materyaller, çevrimiçi öğrenme yazılımı, elektronik ders kitapları) kullanma	Kategorik	1 - evet	3211	0.13
					2 - hayır	21117	0.87

Kamila algoritmasının çalışma süresi 23 değişken, 24 328gözlem ve 55 seviyeli veri setinde 39 dakika, 2 saniye olmuştur.



Şekil 4.7: 2021 Senesine Ait Fert Verisi Kümeleme Sonuçlarına Ait Görseller – Analiz 2

Tablo 4.15: 2021 Senesine Ait Fert Verisi Analiz 2 Kümeleme Sonuçları

	Değişken No	Değişken Adı	Değişken Açıklama / Seçenekleri	Değişken Tipi	Seviyeler	Küme 1	Küme 2	**Toplam**
	1	Yaş	Ferdin yaşı	Sayısal				
	2	Kullanım_Sıklık_İnternet	Son üç ay içinde ortalama internet kullanım sıklığı	Kategorik	6	11795	10955	**22750**
					9	4	284	**288**
					13	42	1248	**1290**
	3	Okul_Biten	Tamamlanan en son okul / en yüksek eğitim seviyesi	Kategorik	2	437	6483	**6920**
					3	1823	3064	**4887**
					4	4416	2277	**6693**
					5	5165	663	**5828**
	4	ISCO08_Meslek	Çalıştığınız işyerindeki/ işinizdeki görev ve sorumluluklarınıza en uygun seçenek	Kategorik	0			
					1	651	142	**793**
					2	1939	82	**2021**
					3	797	100	**897**
					4	722	91	**813**
					5	1264	722	**1986**
					6	83	610	**693**
					7	639	833	**1472**
					8	542	686	**1228**
					9	433	1169	**1602**
					10	4771	8052	**12823**
Son üç ay içinde internet kullanılan cihazlar	5	Fert_Int_Cep	Cep telefonu	Kategorik	1 - evet	11776	12304	**24080**
					2 - hayır	65	183	**248**
	6	Fert_Int_Dizustu	Taşınabilir bilgisayar (dizüstü, netbook vb.)	Kategorik	1 - evet	6120	737	**6857**
					2 - hayır	5721	11750	**17471**
	7	Fert_Int_Tablet	Tablet	Kategorik	1 - evet	2533	564	**3097**
					2 - hayır	9308	11923	**21231**
	8	Fert_Int_Masaustu	Masaüstü Bilgisayar	Kategorik	1 - evet	2817	500	**3317**
					2 - hayır	9024	11987	**21011**
	9	Fert_Int_Diger_Cihaz	Diğer Cihazlar	Kategorik	1 - evet	3889	891	**4780**
					2 - hayır	7952	11596	**19548**

Tablo 4.15: 2021 Senesine Ait Fert Verisi Analiz 2 Kümeleme Sonuçları – Devam

	Değişken No	Değişken Adı	Değişken Açıklama / Seçenekleri	Değişken Tipi	Seviyeler	Küme 1	Küme 2	**Toplam**
Son üç ay içinde kişisel amaçla internet üzerinden faaliyetler	10	Int_Faal_ePosta	e-posta gönderme/ alma	Kategorik	1 - evet	9284	1110	**10394**
					2 - hayır	2557	11377	**13934**
	11	Int_Faal_Telefon	İnternet üzerinden sesli veya görüntülü arama yapma	Kategorik	1 - evet	11434	10470	**21904**
					2 - hayır	407	2017	**2424**
	12	Int_Faal_Sosyal_GrKatılım	Sosyal medyada içerik paylaşma	Kategorik	1 - evet	10303	7457	**17760**
					2 - hayır	1448	5030	**6478**
	13	Int_Faal_Mesaj	Mesajlaşma	Kategorik	1 - evet	11744	10809	**22553**
					2 - hayır	97	1678	**1775**
	14	Int_Faal_Online_Haber	Çevrimiçi haber sitelerini / gazeteleri / haber dergilerini okumak	Kategorik	1 - evet	10318	5673	**15991**
					2 - hayır	1523	6814	**8337**
	15	Int_Faal_Saglık_Bilgi_arm	Sağlıkla ilgili bilgi arama	Kategorik	1 - evet	10669	6104	**16773**
					2 - hayır	1172	6383	**7555**
	16	Int_Faal_MalHiz_Bilgi	Mal ve hizmetler hakkında bilgi arama	Kategorik	1 - evet	10284	3819	**14103**
					2 - hayır	1557	8668	**10225**
	17	Int_Faal_Website_Paylasım	Web siteleri aracılığıyla veya sosyal medya aracılığıyla toplumsal veya siyasal konular ile ilgili görüşleri paylaşma	Kategorik	1 - evet	2210	733	**2943**
					2 - hayır	9631	11754	**21385**
	18	Int_Faal_Oyl_Ktlm	Toplumsal veya siyasal bir konuda online tartışma / oylamaya katılma	Kategorik	1 - evet	742	102	**844**
					2 - hayır	11099	12385	**23484**
	19	Int_Faal_IsArama	İş arama ya da iş başvurusu yapma	Kategorik	1 - evet	1963	557	**2520**
					2 - hayır	9878	11930	**21808**
	20	Int_Faal_MalHizmet_Satıs	Mal veya hizmet satışı	Kategorik	1 - evet	2282	324	**2606**
					2 - hayır	9559	12163	**21722**
	21	Int_Faal_Banka_Islem	İnternet bankacılığı (web sitesi veya mobil bankacılık uygulamaları)	Kategorik	1 - evet	9852	3585	**13437**
					2 - hayır	1989	8902	**10891**
Son üç ay içinde internet üzerinden katılım sağlanan eğitim faaliyetleri	22	Int_Faal_Egt_Kurs	Çevrimiçi (Online) bir kurs alma	Kategorik	1 - evet	2379	155	**2534**
					2 - hayır	9462	12332	**21794**
	23	Int_Faal_Ogrn_Mtryl	Çevrimiçi (Online) öğrenme materyallerini (Görsel-işitsel materyaller, çevrimiçi öğrenme yazılımı, elektronik ders kitapları) kullanma	Kategorik	1 - evet	2947	264	**3211**
					2 - hayır	8894	12223	**21117**

Şekil 4.7.'e göre küme sayısı 2 olduğunda tahmin gücü eşik değeri 0.80'in üzerinde olmaktadır.

Birinci kümede yer alan fertlerin yaş ortalaması 33, interneti her gün kullanan, yüksek eğitim seviyesine sahip, 1,2,3,4, ve 5. Meslek gruplarında çalışan, dizüstü bilgisayara sahip bireylerin çoğunlukta olduğu, internet faaliyetlerinde aktif bireylerden oluştuğu görülmektedir. İkinci kümede yer alan bireylerin yaş ortalaması 43, yüksek eğitim seviyesine sahip olmayan, 6,7,8 ve 9. Meslek gruplarında çalışan veya çalışacak bir iş sahibi olmayan, interneti sesli veya görüntülü arama yapma,

sosyal medyaya katılım sağlama ve mesajlaşma faaliyetleri için kullanan bireylerden oluştuğu belirlenmiştir.

4.5.5. 2019 ve 2021 Senelerine Ait Fert Bulgularının Karşılaştırması

4.5.5.1. Analiz 1'in Karşılaştırması

Analiz 1'de cevabı aranan soru fertlerin yaş, cinsiyet ve belirlenen süre içerisinde internet üzerinden alınan mal veya hizmet türlerine göre benzerliklerinin ortaya konmasıdır. 2019 senesine ait veride son 12 ay içerisinde internet üzerinden alınan mal ve hizmetler sorgulanmışken, 2021 senesine ait veride son 3 ay içerisinde internet üzerinden alınan mal ve hizmetler sorgulanmıştır ve karşılaştırma bu bilgi ile yapılmıştır.

Analiz 1'de 2019 senesine ait 7 906 gözlem, 2021 senesine ait 9 438 gözlem incelenmiştir. Cinsiyet değişkeni için 2019 senesinde erkek bireylerin oranı %54, kadın bireylerin oranı %46, 2021 senesinde erkek bireylerin oranı %51, kadın bireylerin oranı %49 olarak bulunmuştur. 2019 ve 2021 senelerine ait küme grafiklerine göre küme sayısı 4 iken kümeler arası ayrışma en iyi sonucu vermiştir.

2019 senesine ait kümeleme analizinde, kadın bireylerin bir küme içerisinde sınıflanırken, 2021 senesine ait kümeleme analizinde kadın bireylerin iki küme içerisinde sınıflanmıştır. 2019 senesinde yaş ortalaması 23 olan, kadın bireylerin e-ticaret faaliyetlerinden giyim alışverişi yaptıkları belirlenmiştir. 2021 senesine ait analizde, yaş ortalaması 31 olan kadın bireylerin e-ticaret faaliyeti olarak kitap/ dergi alışverişi, yemek siparişi, gıda alışverişi, kozmetik ve kişisel bakım ürünleri aldıkları gözlenmiştir. 2021 senesinde yaş ortalaması 23 olan kadın bireylerin sadece giyim alışverişi için e-ticaret faaliyetinde bulunduğu görülmüştür.

2019 senesinde ait kümeleme analizinde, yaş ortalaması 33 olan erkek bireylerin giyim alışverişi, gıda, seyahat, konaklama ve bilet alım faaliyetlerinde bulundukları belirlenmiştir. 2021 senesinde yaş ortalaması 37 olan erkek bireylerin sadece giyim alışverişi için e-ticaret faaliyetinde bulunduğu gözlenmektedir.

İnternet üzerinden alınan mal ve hizmet türleri karşılaştırıldığında, 2019 senesine ait analizde giyim alışverişi kümeler arası ayırıcı bir değişken iken, 2021 senesine ait analizde ayırıcı olmaktan çıkmış ve tüm kümelerdeki bireyler tarafından yapılan bir faaliyet haline gelmiştir. 2019 senesine ait sınıflandırmada gıda ürünleri alışverişi yapan fertlerin çoğunlukla erkek ve yaş ortalaması 36 iken, 2021 senesinde gıda ürünleri alan fertlerin çoğunlukla kadın ve yaş ortalamasının 31 olduğu belirlenmiştir.

4.5.5.2. Analiz 2'nin Karşılaştırması

Analiz 2'de fertlerin yaş, internet kullanım sıklık, eğitim düzeyi, meslek, son üç ay içerisinde internete bağlanmak için kullanılan cihazlar, son üç ay içerisinde kişisel amaçla internet üzerinden yapılan faaliyetler ve eğitim faaliyetleri alanlarında verdikleri cevaplar analiz edilmiş ve benzerlikleri belirlenmeye çalışılmıştır.

Analiz 2'de 2019 senesine ait 20 316 gözlem, 2021 senesine ait 24 328 gözlem incelenmiştir. Yaş değişkeninin 2019 ve 2021 senelerine ait kümeleme dağılımlarının benzer olduğu görülmüştür.

İki sene arasındaki belirgin fark, 2021 senesindeki birinci kümeye ait profilinde gerçekleşmiştir. 2019 senesinde birinci kümede yer alan yaş ortalaması 33, interneti her gün kullanan, eğitim düzeyi yüksek, kişisel amaçla internet faaliyetlerini etkin kullanan bireyler, 2021 senesinde bunlara ek olarak dizüstü bilgisayar kullanmaya başlamışlardır.

2019 ve 2021 seneleri için, değişken bazında dağılım incelendiğinde elde edilen sonuçlar şöyle olmaktadır:

İnternet kullanım sıklığı hemen her gün seçeneğine olumlu cevap verenler için 2019 senesinde oran %90 iken bu oran 2021 senesinde %94, haftada bir defadan az (iki üç haftada bir) seçeneğine olumlu cevap verenler için 2019 senesinde oran %2 iken bu oran 2021 senesinde %1, haftada en az bir defa seçeneğine olumlu cevap verenler için 2019 senesinde oran %8 iken bu oran 2021 senesinde %5 olmuştur.

Bitirilen okul değişkeni için fertlerin verdikleri cevapların oranı her iki sene için de benzerlik göstermektedir.

Son üç ay içerisinde internete bağlanmak için kullanılan cihazlar için karşılaştırma yapıldığında;

- Cep telefonu kullanım oranı 2019 senesi için %96 iken, 2021 senesinde bu oran %99 (+%3),
- Tablet kullanım oranı 2019 senesi için %9 iken, 2021 senesinde bu oran %12 (+%3),
- Taşınabilir dizüstü bilgisayar kullanım oranı 2019 senesi için %22 iken, 2021 senesinde bu oran %28 (+%6),
- Diğer cihazların kullanım oranı 2019 senesi için %1 iken, 2021 senesinde bu oran %20 (+%19) olmuştur.

Son üç ay içerisinde kişisel amaçla internet üzerinden yapılan faaliyetler ve eğitim faaliyetleri alanlarında karşılaştırma yapıldığında oranları en çok değişen seçenekler,

- İnternet üzerinden sesli veya görüntülü arama yapmak için telefon kullanım oranı 2019 senesi için %82 iken, 2021 senesinde bu oran %90 (+%8),
- Sosyal medyada içerik paylaşma (sosyal_gr_katılım) için oranı 2019 senesi için %80 iken, 2021 senesinde bu oran %73 (- %7),
- Mal ve hizmetler hakkında bilgi arama için oran 2019 senesi için %64 iken, 2021 senesinde bu oran %58 (- %6),
- Web siteleri aracılığıyla veya sosyal medya aracılığıyla toplumsal veya siyasal konular ile ilgili görüşleri paylaşma için oran 2019 senesi için %22 iken, 2021 senesinde bu oran %12 (- %10),
- Toplumsal veya siyasal bir konuda online tartışma / oylamaya katılma için oran 2019 senesi için %8 iken, 2021 senesinde bu oran %3 (- %5),
- Mal veya hizmet satışı için oran 2019 senesi için %21 iken, 2021 senesinde bu oran %11 (- %10),
- İnternet bankacılığı (web sitesi veya mobil bankacılık uygulamaları) için oran 2019 senesi için %45 iken, 2021 senesinde bu oran %55 (+ %10),
- Çevrimiçi (Online) bir kurs alma için oran 2019 senesi için %3 iken, 2021 senesinde bu oran %10 (+ %7),
- Çevrimiçi (Online) öğrenme materyallerini (Görsel-işitsel materyaller, çevrimiçi öğrenme yazılımı, elektronik ders kitapları) kullanma için oran 2019 senesi için %4 iken, 2021 senesinde bu oran %13 (+ %9) olmuştur.

2019 ve 2021 seneleri, Analiz 2'nin bulguları özetlendiğinde, incelenen değişkenler için, Covid 19 pandemisi öncesi ve sonrasında internet kullanılan cihazların kullanım oranları artmış, internet faaliyetleri içerisinde sesli veya görüntülü arama, internet bankacılığı kullanımı, çevrimiçi kurs ve öğrenme materyallerini edinme oranı artmış; sosyal medyada içerik paylaşma, mal ve hizmetler hakkında bilgi alma, web siteleri aracılığıyla veya sosyal medya aracılığıyla toplumsal veya siyasal konular ile ilgili görüşleri paylaşma, toplumsal veya siyasal bir konuda online tartışma / oylamaya katılma, mal veya hizmet satışı faaliyetlerinde oransal olarak azalma belirlenmiştir.

2021 senesi, analiz 2'ye ait bulguların, Covid-19 döneminde bireylerin e-ticaret sıklığını etkileyen faktörlerin tespit edilmesini ve tüketici profillerinin ilgili dönemdeki alışveriş alışkanlıklarının belirlenmesini hedefleyen Alkan vd. (2022)'nin çalışmalarında ulaşılan sonuçlar ile genç yaşta, yüksek eğitim seviyesine sahip, dizüstü bilgisayar kullanan bireylerin internet bankacılığı kullanımının belirlenmesi kapsamında benzerlik gösterdiği ve paralel sonuçlara ulaşıldığı görülmüştür.

SONUÇ

Çok değişkenli istatistiksel yüntemlerden biri olan kümeleme analizinin amacı, gözlemleri benzerliklerine göre sınıflandırmaktır. Sınıflama işlemi yapılırken hedef, gözlem birimlerin küme içi benzerliklerinin maksimum, kümeler arası benzerliklerinin minimum olacak şekilde ayrılabilmesidir. Kümeleme analizindeki üç temel aşamadan ilki, birimlerin benzerlik veya yakınlıklarını belirlemek amacıyla uygun ölçünün şeçilmesidir. İkinci aşamada kullanılacak kümeleme yaklaşımına (hiyerarşik ve hiyerarşik olmayan) ve kümeleme tekniğine karar verilmekte ve üçüncü aşamada ise küme üyeliklerini elde edilmektedir. Kümeleme Analizi, günümüzde veri depolama ve işleme kapasitelerinin artması ve açık kaynak programlama dillerinin yaygınlaşmasıyla sıkça kullanılan bir analiz yöntemi olmaya devam etmektedir.

Kümeleme analizi tanımlayıcı bir yöntemdir ve analiz sonucunda birden çok çözüme ulaşılabilir. Küme sayısı ya da bir gözlemin herhangi bir kümede yer alacağı seçilen uzaklık matrisi ve kümeleme yöntemine göre değişiklik gösterebilmektedir. Bu yönüyle kümeleme analizi tek çözümlü bir yöntem değildir. Kümeleme analizi değişkenler arası birim farklılıklarına duyarlıdır ve bu nedenle sayısal değişkenlerin standartlaştırılması uygun olmaktadır.

Günümüzde artan veri işleme kapasiteleri ve açık kodlu programa dillerinin yaygınlaşması ile büyük hacimde veri setlerinin analiz edilebilirliği artmış ve yeni algoritmalar geliştirilmiştir. Bunlardan bir tanesi olan KAMILA algoritması, Foss vd. (2016) tarafından geliştirilen, yarı parametrik bir kümeleme algoritmasıdır. Bu algoritma sürekli rassal bir değişkenin olasılık yoğunluk fonksiyonu ve rassal değişkenin nasıl dağıldığı konusunda bilgi edinilmesini sağlamaktadır. Bu yolla olasılık yoğunluk fonksiyonundan anakütle hakkında bilgi veren ortalama, varyans gibi istatistiksel özelliklerle birlikte, rassal değişkenin belli bir aralıktaki değerleri hangi olasılıkla alacağı ile ilgili bilgi sağlayabilmektedir. Karma türdeki veri setleri, diğer bir deyişle sayısal ve kategorik değişkenlerin bir arada bulunduğu veri setlerini kümelemek için geliştirilen bu yöntem ile sayısal ve kategorik değişkenlerin katkısının eşit olarak ağırlıklandırılmasına önem verilmektedir.

Bu çalışmada bilişim teknolojilerindeki gelişmelerin bireysel ve toplumsal olarak etkilerinin analiz edilmesi amaçlanmıştır. Bu amaçla zengin bir veri kaynağı olan, Türkiye İstatistik Kurumu'nun yayınladığı, 2019 ve 2021 senelerine ait Hanehalkı Bilişim Teknolojileri Kullanımı Araştırması Mikro Veri Setleri kullanılmıştır. Araştırma hane ve fert bazında olmak üzere iki ana başlıkta toplanmıştır. Fert bazında yapılan çalışmada ise iki farklı soruya yanıt aranmıştır.

Hane bazında yapılan analizde, hane birey sayısı, hane aylık geliri (TL), hanede bulunan bilişim ekipmanları, kullanılan internet bağlantı türleri ve hanenin bulunduğu istatistiki bölgeler

değişkenleri ile gözlemlerin kümelenmesi sağlanmış ve küme sonuçları analiz edilmiştir. Analiz sonucunda her iki sene de, gözlem birimleri üç grupta kümelenmiştir. 2019 senesinde birinci kümede bulunan hanelerin hanede bulunan kişi sayısının en yüksek, toplam aylık hane gelirinin en az olduğu ve bu hanelerin doğu illerimizi kapsadığı belirlenmiştir. İkinci kümede yer alan hanelerde, hanede bulunan kişi sayısı en düşük, toplam aylık hane geliri düşük ve bu hanelerin Akdeniz ve Doğu Karadeniz Bölgeleri'nde yer aldığı sonucuna ulaşılmıştır. Üçüncü kümede yer alan hanelerde hanede bulunan kişi sayısı, diğer iki kümenin ortasında, toplam aylık hane geliri ise en yüksek olarak bulunmuştur. Bu haneler ülkemizin batısında bulunmaktadır. 2021 senesindeki bulgular, 2019 senesine paralel ve doğrular niteliktedir. İki seneye ait analizde öne çıkan farklılık ise, 2021 senesinde Orta Anadolu Bölgesi'nin batıda bulunan illerin profilinden uzaklaşıp, doğuda bulunan illerin profiline yaklaşmasıdır. Buna benzer olarak, Batı Karadeniz Bölgesi 2019 senesinde ülkemizin batısında bulunan illerin profiline sahipken, 2021 senesinde Akdeniz ve Doğu Karadeniz Bölgeleri'nden oluşan bölge profiline yaklaşmıştır. 2021 senesinde doğuda bulunan bölgeleri kapsayan üçüncü kümede çoğunlukla sabit geniş bağlantı kullanılmaya başlanması önemli bir fark olarak göze çarpmaktadır.

Fert bazında iki farklı analiz yapılmıştır. Bunlardan ilkinde bireylerin, yaş ve cinsiyetin özelliklerinin, e-ticaret alışkanlıklarına göre kümelenmesidir. İnternet üzerinden alınan mal ve hizmet türleri karşılaştırıldığında, 2019 senesinde yaş ortalaması 23 olan, kadın bireylerin e-ticaret faaliyetlerinden giyim alışverişi yaptıkları belirlenmiştir. 2021 senesinde yaş ortalaması 23 olan kadın bireylerin benzer şekilde sadece giyim alışverişi için e- ticaret faaliyetinde bulunduğu görülmüştür. 2021 senesine ait analizde, yaş ortalaması 31 olan kadın bireylerin e-ticaret faaliyeti olarak kitap/ dergi alışverişi, yemek siparişi, gıda alışverişi, kozmetik ve kişisel bakım ürünleri aldıkları gözlenmiştir. 2019 senesinde ait kümeleme analizinde, yaş ortalaması 33 olan erkek bireylerin giyim alışverişi, gıda, seyahat, konaklama ve bilet alım faaliyetlerinde bulundukları belirlenmiştir. 2021 senesinde yaş ortalaması 37 olan erkek bireylerin sadece giyim alışverişi için e-ticaret faaliyetinde bulunduğu gözlenmektedir. İnternet üzerinden alınan mal ve hizmet türleri karşılaştırıldığında, 2019 senesine ait analizde giyim alışverişi kümeler arası ayırıcı bir değişken iken, 2021 senesine ait analizde ayırıcı olmaktan çıkmış ve tüm kümelerdeki bireyler tarafından yapılan bir faaliyet haline gelmiştir. 2019 senesine ait sınıflandırmada gıda ürünleri alışverişi yapan fertlerin çoğunlukla erkek ve yaş ortalaması 33 iken, 2021 senesinde gıda ürünleri alan fertlerin çoğunlukla kadın ve yaş ortalamasının 31 olduğu belirlenmiştir.

Fert bazında yapılan ikinci analizde, fertlerin yaş, internet kullanım sıklığı, eğitim durumu, mesleği, taşınabilir cihazlar üzerinden yapılan internet faaliyetlerinin benzerliklerine göre kümelenme sonuçları incelenmiştir. Analiz bulgularına göre incelenen değişkenler için Covid 19 pandemisi öncesi ve sonrasında, internet kullanım sıklığı ve internet kullanılan cihazların kullanım oranları artmıştır.

İnternet faaliyetleri içerisinde sesli veya görüntülü arama, internet bankacılığı kullanımı, çevrimiçi kurs ve öğrenme materyallerini edinme oranı artmış; sosyal medyada içerik paylaşma, mal ve hizmetler hakkında bilgi alma, web siteleri aracılığıyla veya sosyal medya aracılığıyla toplumsal veya siyasal konular ile ilgili görüşleri paylaşma, toplumsal veya siyasal bir konuda online tartışma / oylamaya katılma, mal veya hizmet satışı faaliyetlerinde oransal olarak azalma belirlenmiştir. Her iki analiz senesi için de geçerli olmak üzere, fertlerin eğitim durumu ve mesleği, taşınabilir cihazlar üzerinden yapılan internet faaliyetlerinin çeşitliliği üzerinde önemli bir etkiye sahiptir. Eğitim seviyesi arttıkça, internette yapılan faaliyetlerin sayısı da artmaktadır.

Bu çalışma, Türkiye İstatistik Kurumu'nca derlenen Hanehalkı Bilişim Teknolojileri Kullanım İstatistikleri verileri kullanılarak yapılan çalışmalardan, karma tipte değişkenler içeren büyük veri setleri için tasarlanmış KAMILA algoritmasının ilk defa kullanılması yönüyle ayrışmaktadır. Çalışmanın devamı niteliğinde, 2021 senesi sonrasında yapılan anket çalışmasından elde edilecek veriler, bu çalışmada elde edilen sonuçlar ile karşılaştırılmak suretiyle analiz çıktılarının sürekliliği araştırılabilir.

KAYNAKLAR

Kitaplar

Alpar, R. (2021), *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*, Ankara: Detay Yayıncılık.

Altaş, D. (2013), İstatistiksel Analiz (1. Baskı), İstanbul: Beta Basım.

Bilder, C.R., Loughin, T.M. (2015), *Analysis of Categorical Data with R*, Florida: CRC Press Taylor and Francis Group.

Bulut,H. (2018), *R Uygulamaları İle Çok Değişkenli İstatistiksel Yöntemler*, Ankara: Nobel Yayın.

Çağlayan Akay, E., Kangallı Uyar, S.G. (2017), *R Uygulamalı Nonparametrik Ekonometri*, İstanbul: Der Yayınları.

Chatfield, C., Collins, A.J. (1992), Introduction to Multivariate Analysis, Cambridge: Chapman & Hall.

Çilingirtürk, A.M. (2011), *İstatistiksel Karar Almada Veri Analizi*, Ankara: Seçkin Yayıncılık.

Der, G., Everitt, B.S. (2009), *A Handbook of Statistical Analyses Using SAS*, United States: CRC Press Taylor & Francis Group.

Ergüt, Ö., Altaş, D., Yıldırım, İ.E.(Ed.) (2020), "Kümeleme Analizi ve Öğrencilerin Umutsuzluk Düzeylerinin İncelenmesi" *Uygulamalı Çok Değişkenli İstatistik Teknikleri*, Ankara: Seçkin: 75-90.

Hair, J.F.Jr., Black, W.C., Babin, B.J., Anderson, R.E. (2010), *Multivariate Data Analysis (7th Edition*), United States of America: Pearson.

Kvam, P.H, Vidakovic, B. (2007), *Nonparametric Statistics with Applications to Science and Engineering*, New Jersey: John Wiley & Sons.

Nakip, M., Yaraş, E. (2017), *Pazarlama Araştırma Teknikleri ve SPSS Uygulamaları*, Ankara: Seçkin Akademik ve Mesleki Yayınlar.

Oktay, E. (2017), *Kontenjans Tablolarından Elde Edilen İlişki Ölçüleri Öğretim Üyesi Değerleme Çalışması*, Erzurum: Erzurum Kültür Eğitim Kitap ve Kırtasiye.

Özdamar, K. (2004), *Paket Programlar ile İstatistiksel Veri Analizi-2 (Yenilenmiş 5.Baskı)*, Eskişehir: Kaan Kitapevi.

Scott, D.W. (1992), *Multivariate Density Estimation*, New York: John Wiley & Sons.

Sharma, S. (1996), *Applied Multivariate Techniques*, United States of America: John Wiley & Sons.

Tatlıdil, H. (2002), *Uygulamalı Çok Değişkenli İstatistiksel Analiz*, Ankara: Akademi Matbaası.

e-Kitap

Abonyi, J., Feil, B. (2007), *Cluster Analysis for Data Mining and System Identification*. e-ISBN 978-3-7643-7988-9

Everitt, B.S., Hand, D.J. (1981), *Finite Mixture Distributions*. doi: 10.1007/978-94-009-5897-5

Everitt, B., Hothorn, T. (2011), *An Introduction to Applied Multivariate Analysis with R*. doi:10. 1007/978-1-4419-9650-3

Süreli Yayınlar

Ahmad, A., Khan, S.S. (2019), "Survey of State-of-the-Art Mixed Data Clustering Algorithms", IEEE, Vol 7, 31883- 31902 doi: 10.1109/ACCESS.2019.2903568

Anıl, B., Köksal, E. (2016), "Türkiye'de İnterneti Kimler, Ne İçin Kullanıyor?", Marmara Üniversitesi İktisadi ve İdari Bilmler Dergisi, Cilt 38, Sayı 1, 1-13. doi: 10.14780/iibd.61602

Arıcıgil Çilan, Ç., Taş, N., Özdemir, M. (2013), "Gizli Sınıf Analizi ile Türkiye'de Kişisel İnternet Kullanım Profilinin Belirlenmesi", Dumlupınar Üniversitesi Sosyal Bilimler Dergisi EYİ 2013 Özel Sayısı

Arıcıgil Çilan, Ç., Kuzu, S. (2013), "Kişisel E-Ticaret Uygulamalarının Kategorik Veri Analizi Yöntemleri ile Değerlendirilmesi", The Journal of Operation Research, Statistics, Econometrics and Management Information Systems, Vol 1, Issue 1, 27-32.
This paper has been presented at 14th International Symposium on Econometrics Operations Research and Statistics

Bilgiç, E. (2019), "Karma Tipteki Verileri Kamila, K-Ortalamalar, K-Ortaylar ve K-Prototipler Algorıtmalariyla Kümeleme: Karşılaştırmalı Bir Uygulama", S.C.Ü. İktisadi ve İdari Bilimler Dergisi, Cilt 20, Sayı 2, 1-24
http://esjournal.cumhuriyet.edu.tr/tr/download/article-file/867054, (27 10 2022)

Cha, Sung-Hyuk (2007), "Comprehensive Survey on Distance/ Similarity Measures between Probability Density Functions", International Journal of Mathematical Models and Methods in Applied Science, Vol.1, Issue 3, 300.

Coşkun, M., Bülbül, H.İ. (2019), "Hanehalkı İnternet Hizmeti Sahipliğini Ektileyen Faktörlerin Karar Ağaçları ile İncelenmesi", Türk Bilim Araştırma Vakfı, Cilt 12, Sayı 2, 1-17.

Demirel, O. (2022), "Türkiye'de Bilgisayar Sahipliğini Etkileyen Faktörler: Logit ve Bivariate Probit Yaklaşımları", Alanya Akademik Bakış Dergisi, Cilt 6, Sayı 2, 2275-2291.
doi: 10.29023/alanyaakademik.1038258

Ecemiş, O., Coşkun, A. (2022), "Türkiye'de Bilişim Teknolojileri Kullanımının ÇKKV Yöntemleriyle İncelenmesi: 2014-2021 Dönemi", Avrupa Bilim Ve Teknoloji Dergisi Özel Sayı 37, 81-89.
doi: 10.31590/ejosat.1134753

Fidan, H. (2017), "Türkiye Bölgesel Sayısal Bölinme Düzeylerinin Belirlenmesinde Gini Yaklaşımı", Business and Economics Research Journal, Vol 8, Number 1, 49-62.
doi: 10.20409/berj.2017126244

Foss, A., Markatou, M. (2018), "kamila: Clustering Mixed Type Data in R and Hadoop", Journal of Statistical Software, 6:7.
doi: 10.18637/jss.v083.i13

Foss, A., Markatou, M., Ray, B., Heching, A. (2016), "A Semiparametric Method for Clustering Mixed Data", Mach Learn, 105:419-458.
doi: 10.1007/s10994-016-5575-7

Görgün Baran, A., Erdem, M.T. (2017),"Bilgi Toplumunda Dijital Bölünme: Bilişim ve İletişim Teknolojileri Kullanım Yetenekleri Üzerinden Bir Tartışma", Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt 22, Kayfor 15, Özel Sayısı, 1505-1518.

Jiao, L., Denœux, T., Liu, Z-g., Pan, Q. (2022), "EGMM: An Evidential Version of the Gaussian MixtureModel for Clustering", Applied Soft Computing, 129 (2022) 109619.
doi: 10.1016/ j.asoc/2022/109619

Marangoz, M., Özkoç, H.H., Aydın, A.E. (2019), "Tüketicilerin İnternet Üzerinden Alışveriş Davranışlarının Açıklanmasına Yönelik Bir Çalışma", Tüketici ve Tüketim Araştırmaları Dergisi, 11(1), 1-22.

Mbuga, F., Tortora, C. (2022), "Spectral Clustering of Mixed-Type Data", Stats 2022, 5, 1-11
doi: 10.3390/stats5010001

Selim, S., Balyaner, İ. (2017), "Türkiye'de Hanehalkının Sahip Olduğu Bilişim Teknolojileri Ürünleri Sayısını Belirleyen Faktörlerin Araştırılması: Bir Sayma Veri Modeli", Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, Cilt 9, Sayı 22: 428-454.
doi: 10.20875/makusobed.296800

Tang, Y., Browne, R.P., McNikholas, P.D. (2015), "Model Based Clustering of High Dimensional Binary Data", Computational Statistics and Data Analysis, Vol.87:84-101.
doi:10.1016/j.csda.2014.12.009

Tibshirani, R., Walther, G. (2005), "Cluster Validation by Prediction Strength", Journal of Computational and Graphical Statistics, Vol. 14, Number 3:511-528.
doi: 10.1198/106186005X59243

İnternet Kaynakları

Alkan, Ö., Ünver, Ş. (2022), "E-Commerce Use of Generation Z in Turkey", 8th International Mardin Artuklu Scientific Researches Conference held in Mardin in June 2022, 787-796.

Alkan, Ö., Ünver, Ş., Bayhan, Y.C. (2022), "Determination of Factors Affecting the Frequency of E-Commerce in the Covid-19 Period in Turkey", 10th International Conference on Social Sciences and Humanities held in Sivas on July 18-19,2022, 192-204.

cluster: "Findings Groups in Data": Cluster Analysis Extended Rousseeuw et al. Package 'cluster April 17, 2021'.pdf, https://cran.r-project.org/web/packages/cluster/index.html

dbscan: Density- Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms. Package 'dbscan October 27,2022'.pdf, https://cran.r-project.org/web/packages/dbscan/index.html

factoextra: Extract and Visualize the Results of Multivariate Data Analyses. Package 'factoextra October 13, 2022'.pdf, https://cran.r-project.org/web/packages/factoextra/index.html

ggplot2: Create Elegant Data Visualisation Using the Grammar of Graphics. Package 'ggplot2 June 25, 2021'.pdf, https://cran.r-project.org/web/packages/ggplot2/index.html

Hayır Kanat, M, "Coğrafya Alan İncelemeleri Bölge Kavramı ve NUTS Bölgeleri", https://avesis.yildiz.edu.tr/search?scope=All&q=mhayir, s.7

kamila: Methods for Clustering Mixed Type Data. Package 'kamila March 13, 2020'.pdf, https://cran.r-project.org/web/packages/kamila/index.html

Kara, E., Eşref, S., Çağıltay, K., "Türkiye'de Aktif İnternat Kullanım Eğilimleri: 2004-2014 Dönemi" http://inet-tr.org.tr/inetconf21/bildiri/36.pdf, (28 Ekim 2022).

mclust: Gaussian Mixture for Model-Based Clustering, Classification, and Density Estimation. Package 'mclust October 31, 2022'.pdf, https://cran.r-project.org/web/packages/mclust/index.html

McParland, D., Gormley, I.C., “Model Based Clustering for Mixed Data:clustMD”, https://arxiv.org/abs/1511.01720

RDocumentation flexmixedruns: Fitting mixed Gaussian/ multinominal mixture with flexmix https://www.rdocumentation.org/packages/fpc/versions/2.2-9/topics/flexmixedruns adresinden alındı

r-project. The R Project for Statistical Computing https://www.r-project.org/ adresinden alındı

r-project. What is R? https://www.r-project.org/about.html adresinden alındı

r-project. Contributers https://www.r-project.org/contributors.html adresinden alındı

Sezer, F., İşgör, İ.Y., Erdener, M.A. (2019), “İnternet ve Bilgisayar Kullanımı Üzerine Bir İnceleme”, ULEAD 2019 Annual Congress: ICRE, s:179- 185. https://www.researchgate.net/profile/Hemza-Boumaraf/publication/340438429_Students'_Spatial_Perception_for_3d_Printing_In_Architctural_Education/links/6000110ba6fdccdcb8518412/Students-Spatial-Perception-for-3d-Printing-In-Architectural-Education.pdf#page=200

TÜİK Türkiye İstatistik Kurumu Sınıflama Sunucusu . Uluslararası Standard Meslek Sınıflaması – ISCO 08. https://biruni.tuik.gov.tr/DIESS/SiniflamaSatirListeAction.do?surumId=210&seviye=4&detay=H&tu%20rId=41&turAdi=%209.%20Meslek%20S%C4%B1n%C4%B1flamalar%C4%B1 adresinden alındı.

Wang, J. (2021), “Analysis of Shared Bicycle Usage Based on K-Means and GMM Clustering Algorithm”, 2nd Seminar on Artificial Intelligence, Networking and Information Technology, s:95. doi: 10.1109/AINIT54228.2021.00028 https://c85689232ea394a8dc08a512c1f46793a2397178.vetisonline.com/document/9725075?arnumber=9725075

Diğer Kaynaklar

Arı, Ç. (2013), Maximum Likelihood Estimation of Robust Constrained Gaussian Mixture Models, Ph.D. Thesis, Ankara, Bilkent University The Department of Electrical and Electronics Engineering and the Graduate School of Engineering and Science.

Camkıran, C. (2017), Farklı Kümeleme Tekniklerinin Karşılaştırılması Üzerine Bir Uygulama, Yüksek Lisans Tezi, İstanbul, Marmara Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı İstatistik Bili Dalı.

Erçelik, E. (2019), On Estimation of Probability Density Function, Ph.D. Thesis, Istanbul, Istanbul Technical University Graduate School of Science Engineering and Technology.

Türkiye İstatistik Kurumu, Hanehalkı Bilişim Teknolojileri Kullanım İstatistikleri Soru Formu, 2019.