
GAZİ ÜNİVERSİTESİ
Gazi
1926

KÜMELEME ANALİZİNDE KULLANILAN BAZI BENZERLİK İNDEKSLERİNİN KARŞILAŞTIRILMASI

Hazan Kübra HACIOĞLU

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANABİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TEMMUZ 2016

Hazan Kübra HACIOĞLU tarafından hazırlanan "KÜMELEME ANALİZİNDE KULLANILAN BAZI BENZERLİK İNDEKSLERİNİN KARŞILAŞTIRILMASI" adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi İstatistik Anabilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Prof. Dr. Semra ERBAŞ

İstatistik, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum ...……

Başkan: Doç. Dr. Serpil AKTAŞ ALTUNAY

İstatistik, Hacettepe Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum ……...

Üye: Doç. Dr. Hülya OLMUŞ

İstatistik, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum ……...

Tez Savunma Tarihi: 01/07/2016

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

……..…….

Prof. Dr. Metin GÜRÜ

Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Hazan Kübra HACIOĞLU

01/07/2016

KÜMELEME ANALİZİNDE KULLANILAN BAZI BENZERLİK İNDEKSLERİNİN KARŞILAŞTIRILMASI

(Yüksek Lisans Tezi)

Hazan Kübra HACIOĞLU

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Temmuz 2016

ÖZET

Araştırmacılar veri seti hakkında çıkarsama yapabilmek için, birçok çalışmada homojen ve uygun sayıda gruba ihtiyaç duyarlar. Kümeleme analizi, veri setinin altında yatan doğal grupları ortaya koyan ve birçok alanda yaygın olarak kullanılan çok değişkenli istatistiksel bir yöntemdir. Kümeleme analizinde, anlamlı ve geçerli sonuçlara ulaşmada uygun küme sayısının belirlenmesi birçok araştırmacının sıklıkla karşılaştığı önemli sorunlardan biridir. Kümeleme kalitesinin değerlendirilmesinde ve uygun küme sayısının belirlenmesinde küme geçerlilik indeksleri kullanılmaktadır. Ancak bazı karmaşık yapılar içeren verilerde, küme üyeliklerindeki kararsızlıklar nedeniyle küme geçerlilik indeksleri birbirleriyle çelişen sonuçlar verebilmektedir. Bu çalışmada, en uygun küme sayısının belirlenmesinde kullanılan küme geçerlilik indeksleri tanıtılarak, R ortamında elde edilen yapay veri setleri ile karşılaştırılmıştır. Ayrıca İstatistiki Bölge Birimleri Sınıflandırması (İBBS) Düzey 2 bölgelerinin kadın işgücü ve eğitim istatistikleri kullanılarak bir uygulama çalışması sunulmuştur. Analiz sonuçlarına göre Silhouette indeksinin küme geçerliliği değerlendirilmesinde kullanılan geçerlilik indekslerinden daha başarılı sonuçlar verdiği saptanmıştır.

Bilim Kodu : 20512

Anahtar Kelimeler : Kümeleme Analizi, küme geçerlilik indeksi, en uygun küme sayısı, içsel geçerlilik indeksleri, Kadın işgücü

Sayfa Adedi : 83

Danışman : Prof. Dr. Semra ERBAŞ

COMPARISON OF SIMILARITY INDICES IN CLUSTER ANALYSIS

(M. Sc. Thesis)

Hazan Kübra HACIOĞLU

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

July 2016

ABSTRACT

Researchers are in need of homogeneous and optimal number of groups in order to make inference about data set. Cluster analysis is a commonly used multivariate statistical method in many fields which reveal natural groups underlying data sets. Determining the optimal number of cluster is an important problem while obtaining efficient and valid results in the cluster analysis. Cluster validity indices are used in the evaluation of the quality of the clustering and determining optimal number of cluster. But, when the complex data are being analysed, cluster analysis results can give conflicting results. In this study, the performances of validity indices which is used to determine the optimal number of clusters are introduced and compared to each other via artificial data set obtained in R programming. In addition, this experimental study has studied Classification of Statistical Regional Units Level-2 regions in Turkey through women labour and training statistics. According to the analysis results, it was obtained that Silhouette index is more successful than the cluster validity indices which are used in clustering validation.

Science Code : 20512

Key Words : Cluster analysis, cluster validity index, optimal number of cluster, internal validation measures, women labour

Page Number : 83

Supervisor : Prof. Dr. Semra ERBAŞ

TEŞEKKÜR

Çalışmalarımın her aşamasında bilgi, öneri ve katkılarıyla beni yönlendiren değerli hocam Prof. Dr. Semra ERBAŞ'a, eğitim hayatım boyunca maddi ve manevi destekleriyle beni yalnız bırakmayan, her kararımda sorgulamadan arkamda duran sevgili annem Şaziye HACIOĞLU, babam Mahmut Başar HACIOĞLU ve kardeşim Fatih HACIOĞLU'ya sonsuz teşekkürler.

İÇİNDEKİLER

Sayfa

ÖZET iv

ABSTRACT v

TEŞEKKÜR vi

İÇİNDEKİLER vii

ÇİZELGELERİN LİSTESİ ix

ŞEKİLLERİN LİSTESİ xi

SİMGELER VE KISALTMALAR xii

1. GİRİŞ 1

2. KÜMELEME ANALİZİ 5

2.1. Kümeleme Analizinde Değişken Seçimi ve Uzaklık Ölçüleri 6

2.2. Kümeleme Analizi Teknikleri 9

2.2.1. Hiyerarşik Kümeleme Yöntemleri 9

2.2.2. Hiyerarşik Olmayan Kümeleme Yöntemleri 12

3. KÜME GEÇERLİLİK İNDEKSLERİ 15

3.1. Calinski- Harabasz İndeksi 16

3.2. Krzanowski-Lai İndeksi 17

3.3. Davies-Bouldin İndeksi 17

3.4. Silhouette İndeksi 18

3.5. Dunn İndeksi 19

3.6. Gamma İndeksi 20

3.7. Tau İndeksi 20

3.8. McClain İndeksi 21

Sayfa

3.9. Point-Biserial İndeksi 22
3.10. Gplus İndeksi 22
3.11. Ratkowsky İndeksi 22
3.12. Kübik Kümeleme Kriteri (CCC) 23
3.13. C indeksi 24
3.14. S indeksi 24
3.15. SDbw İndeksi 25
3.16. Duda İndeksi 26
3.17. PseudoT^2 İndeksi 27
3.18. Gap İstatistiği 27
3.19. Ball İndeksi 28
3.20. Hubert İstatistiği 29
3.21. D indeksi 29
4. KÜME GEÇERLİLİK İNDEKSLERİNİN KARŞILAŞTIRILMASI 33
4.1. Yapay Veri Setlerine Ait Sonuçların Değerlendirilmesi 33
4.2. Düzey 2 Bölgelerinin Kadın İşgücü İstatistikleri Bakımından Kümelenmesi. 63
5. SONUÇ VE ÖNERİLER 75
KAYNAKLAR 79
ÖZGEÇMİŞ 83

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge 3.1. İçsel küme geçerlilik indeksleri 30

Çizelge 4.1. Yapay Veri setlerinin özellikleri 33

Çizelge 4.2. 3 kümeli farklı yoğunluklu veri setinin tek bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri 35

Çizelge 4.3. 3 kümeli farklı yoğunluklu veri setinin tam bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri 36

Çizelge 4.4. 3 kümeli farklı yoğunluklu veri setinin ward yöntemine göre uygun küme sayıları ve indeks değerleri 37

Çizelge 4.5. 3 kümeli farklı yoğunluklu veri setinin k-ortalama yöntemine göre uygun küme sayıları ve indeks değerleri 38

Çizelge 4.6. 3 kümeli farklı yoğunluklu veri setinin en uygun küme sayıları 39

Çizelge 4.7. 4 kümeli iyi ayrılmış veri setinin tek bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri 41

Çizelge 4.8. 4 kümeli iyi ayrılmış veri setinin tam bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri 42

Çizelge 4.9. 4 kümeli iyi ayrılmış veri setinin ward yöntemine göre uygun küme sayıları ve indeks değerleri 43

Çizelge 4.10. 4 kümeli iyi ayrılmış veri setinin k-ortalama yöntemine göre uygun küme sayıları ve indeks değerleri 44

Çizelge 4.11. 4 kümeli iyi ayrılmış veri setinin en uygun küme sayıları 45

Çizelge 4.12. Alt kümeli veri setinin tek bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri 47

Çizelge 4.13. Alt kümeli veri setinin tam bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri 48

Çizelge 4.14. Alt kümeli veri setinin ward yöntemine göre uygun küme sayıları ve indeks değerleri 49

Çizelge 4.15. Alt kümeli veri setinin k-ortalama yöntemine göre uygun küme sayıları ve indeks değerleri 50

Çizelge 4.16. Alt kümeli veri setinin en uygun küme sayıları 51

Çizelge **Sayfa**

Çizelge 4.17. Cassini veri setinin tek bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri 53

Çizelge 4.18. Cassini veri setinin tam bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri 54

Çizelge 4.19. Cassini veri setinin ward yöntemine göre uygun küme sayıları ve indeks değerleri 55

Çizelge 4.20. Cassini veri setinin k-ortalama yöntemine göre uygun küme sayıları ve indeks değerleri 56

Çizelge 4.21. Cassini veri setinin en uygun küme sayıları..... 57

Çizelge 4.22. İris veri setinin tek bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri 58

Çizelge 4.23. İris veri setinin tam bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri 59

Çizelge 4.24. İris veri setinin Ward yöntemine göre uygun küme sayıları ve indeks değerleri 60

Çizelge 4.25. İris veri setinin k-ortalama yöntemine göre uygun küme sayıları ve indeks değerleri 61

Çizelge 4.26. İris veri setinin en uygun küme sayıları 62

Çizelge 4.27. Analizde kullanılan değişkenler 64

Çizelge 4.28. Küme geçerlilik indeks değerleri ve uygun küme sayıları..... 65

Çizelge 4.29. Uygun küme sayısının belirlenmesinde kullanılan indeks kritik değerleri..... 66

Çizelge 4.30. Uygun küme sayısı k = 2 olan iller..... 69

Çizelge 4.31. Uygun küme sayısı k = 3 olan iller..... 70

Çizelge 4.32. Uygun küme sayısı k = 6 olan iller..... 71

Çizelge 4.33. Uygun küme sayısı k = 10 olan iller..... 72

Çizelge 4.34. Küme geçerlilik indekslerinin doğru sınıflama oranları 73

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil 2.1. Kümeleme analizinin uygulanma aşamaları 6

Şekil 2.2. Hiyerarşik kümeleme yöntemlerinin ağaç diyagramı 9

Şekil 2.3. Tek bağlantı tekniği 10

Şekil 2.4. Tam bağlantı tekniği 11

Şekil 2.5. Ortalama bağlantı tekniği 11

Şekil 3.1. Küme içi ve kümeler arası uzaklıklar 15

Şekil 4.1. Yapay veri setlerinin serpme grafikleri 34

Şekil 4.2. Farklı yoğunluklu veri seti için uygun küme sayılarının Silhouette indeksine göre gösterimi 40

Şekil 4.3. Farklı yoğunluklu veri seti için uygun küme sayılarının Dunn indeksine göre gösterimi 40

Şekil 4.4. İyi ayrılmış veri seti için uygun küme sayılarının Silhouette indeksine göre gösterimi 46

Şekil 4.5. İyi ayrılmış veri seti için uygun küme sayılarının Dunn indeksine göre gösterimi 46

Şekil 4.6. Alt kümeli veri seti için uygun küme sayılarının Silhouette indeksine göre gösterimi 52

Şekil 4.7. Alt kümeli veri seti için uygun küme sayılarının Dunn indeksine göre gösterimi 52

Şekil 4.8. D indeks grafikleri 67

Şekil 4.9. Hubert istatistiği grafikleri 68

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklamalar
CH	Calinski-Harabasz indeks değeri
DB	Davies-Bouldin indeks değeri
EUROSTAT	Avrupa Birliği İstatistik Ofisi
İBBS	İstatistiki Bölge Birimi Sınıflandırması
KL	Krzanowski-Lai indeks değeri
TÜİK	Türkiye İstatistik Kurumu

1. GİRİŞ

Kümeleme analizi gözlem vektörleri ya da değişkenler uzayındaki noktalar arasındaki uzaklıklara veya benzerliklere dayalı olarak birimleri kümelere ayırmada ve uygun küme sayısının belirlenmesinde yaygın olarak kullanılan çok değişkenli istatistiksel analiz yöntemidir [15].

Kümeleme analizi ilk kez John Snow tarafından 1854 yılında Londra'daki bir kolera salgını sırasında kullanılmıştır [18]. John Snow, salgın sırasında kendisine rapor edilen vakaları üzerinde işaretlemek üzere özel bir harita oluşturmuştur. Harita üzerinde işaretlenen vakaları inceleyerek, hastalığın belli bir sokak üzerinde yoğunlaştığını gözlemlemiştir. Bunun üzerine hastalığa son verilmek üzere o bölgedeki kuyu pompaları kaldırılmıştır. Bu olay kümeleme analizinin en basit ve bilinen ilk uygulamasıdır. Veri kümeleme analizi kavram olarak ise ilk kez 1939 yılında Tryon tarafından kullanılmıştır. Bu kavram 1960 yıllarından sonra gelişmiş ve kullanımı yaygınlaşmıştır. 1963 yılında Robert Sokal ve Peter Sneath'ın yazdığı "Sayısal Sınıflandırma İlminin Temelleri" adlı kitap bu alanda önemli bir katkıdır [1]. O zamandan beri kümeleme analizi büyük boyuttaki verilerin doğal gruplarını bulmak için, istatistik, veri madenciliği, tıp, biyoloji, psikoloji, sosyoloji, bankacılık ve pazarlama gibi birçok alanda yaygın olarak kullanılmaktadır.

Kümeleme analizi içerisinde çok sayıda farklı kümeleme tekniğinin olması verilerin kümelenmesinde hangi kümeleme tekniğinin daha başarılı olduğunu, en uygun küme sayısının nasıl belirleneceği ve elde edilen küme geçerliliğinin nasıl değerlendirileceği sorularını gündeme getirmiş ve bu konular üzerinde çok sayıda çalışma yapılmıştır. Bu çalışmalardan bazılarına aşağıda kısaca yer verilmiştir.

Friedman ve Rubin, kümelemenin geçerliliğini değerlendirmek için Wilks'in λ değerini önermişlerdir. Bu kritere göre, λ değeri ne kadar küçük ise kümelemenin geçerliliği o kadar yüksektir [16].

Calinski ve Harabasz, kümelemenin geçerliliğini değerlendirebilmek için, kümeler içi ve kümeler arası kareler toplamına dayanan bir indeks önermişlerdir. CH indeksini maksimum yapan küme sayısı geçerli küme sayısı kabul edilmektedir [7].

Dunn, iyi ayrılmış kümeler ve optimum dağınık parçalanma adlı çalışmasında, kümeleme sonuçlarının doğruluğunun değerlendirilmesinde kullanılan ve indeks değerini maksimum yapan küme sayısının en başarılı kümeleme olarak seçildiği yeni bir indeks önermiştir [13]. Fakat bu indekste küme sayısı arttıkça indeks değerinin hesaplanması giderek zorlaşmaktadır.

Baker ve Hubert, iki kümeleme sonucunda küme üyeliği tutarlı olan birimlerin sayısı ile tutarlı olmayan birimlerin sayısı arasındaki farkın, toplam birim sayısına oranlanması ile elde edilen Γ indeksini farklı kümeleme sonuçlarının değerlendirilmesi için önermişlerdir. Bu kritere göre, Γ indeks değeri 1'e yaklaştıkça, iki farklı kümeleme sonucunun birbirine olan benzerliği artmaktadır [4].

Davies ve Bouldin, küme içindeki birimlerin küme merkezine olan uzaklıklarını minimum, kümeler arasındaki uzaklıkları maksimum yapmayı amaçlayan DB indeksini önermişlerdir [10].

Hubert ve Arabie, küme sayısının artması durumunda kümeleme performansına bağlı olmaksızın Rand indeks değerinin de artması sorununa karşın genelleştirilmiş hipergeometrik dağılım varsayımı altında düzeltilmiş Rand indeksini önermiştir [26].

Rousseeuw, birimlerin kendi kümesi içerisindeki birimlere olan uzaklığına ve diğer kümelerdeki birimlere olan uzaklığına dayanan Silhouette istatistiğini önermiştir. Bu istatistiğine göre, maksimum ortalama Silhouette değerine ulaşılan küme sayısı uygun küme sayısı seçilmektedir [41].

Krzanowski ve Lai tarafından k ve k+1 küme sayılarına göre, küme içi kareler toplamını göz önünde bulundurarak hesapladıkları *DIFF* değerlerinin birbirine oranlanması ile elde edilen KL indeksini önerilmiştir. Bu indekse göre, KL indeks değerini en büyük yapan k uygun küme sayısıdır [28].

Tibshirani ve diğerleri, uygun küme sayısını belirlemek için Gap istatistiğini önermişlerdir. Veriyi kümelemede kullanılan tüm yöntemlerde uygulanabilir olan Gap istatistiği verinin yapısına uygun bir referans sıfır dağılımı oluşturulur. Bu referans dağılımından gerçek veri ile aynı hacimli rasgele örneklem üretilir. Bu rasgele örnekleme orijinal veriye uygulanan kümeleme yöntemi uygulanarak orijinal verinin küme içi saçılımı ile rasgele örneklemden elde edilen kümeleme yapısındaki küme içi saçılımlar karşılaştırılır [46]. Bu kritere göre, Gap istatistik değerini maksimum yapan küme sayısı uygun küme sayısı olarak alınır.

Kümeleme analizinde en uygun küme sayısının belirlenmesi araştırmacıların karşılaştığı önemli sorunlardan biridir. Bu çalışmanın amacı uygun küme sayısının belirlenmesinde ve kümeleme kalitesinin değerlendirilmesinde kullanılan bazı küme geçerlilik indekslerinin, farklı kümeleme yöntemleri kullanılarak karşılaştırılmasıdır. Çalışmanın ikinci bölümünde kümeleme kavramı, kümeleme analizinde kullanılan bazı uzaklık ölçüleri ve kümeleme tekniklerine yer verilmiştir. Üçüncü bölümde, uygun küme sayısının belirlenmesinde kullanılan bazı küme geçerlilik indeksleri tanıtılmıştır. Dördüncü bölümde farklı kümeleme yöntemleri için kümeleme geçerlilik indekslerinin işleyişi ve R ortamında elde edilen yapay veri setleri ve gerçek bir veri seti kullanılarak, k-ortalama kümeleme yöntemine göre küme geçerlilik indeksleri ile belirlenen uygun küme sayıları karşılaştırılmıştır.

2. KÜMELEME ANALİZİ

İnsanoğlu var olduğundan bu yana etrafında bulunan nesneleri bazı özelliklerine göre sınıflara ayırma eğiliminde olmuştur. Ancak birimlerin sayısı arttıkça birimleri sınıflandırmak daha da zorlaşmış ve yeni teknikler bulmayı gerektirmiştir. Bu gereksinim sonucu kümeleme analizi kavramı ortaya çıkmıştır [14].

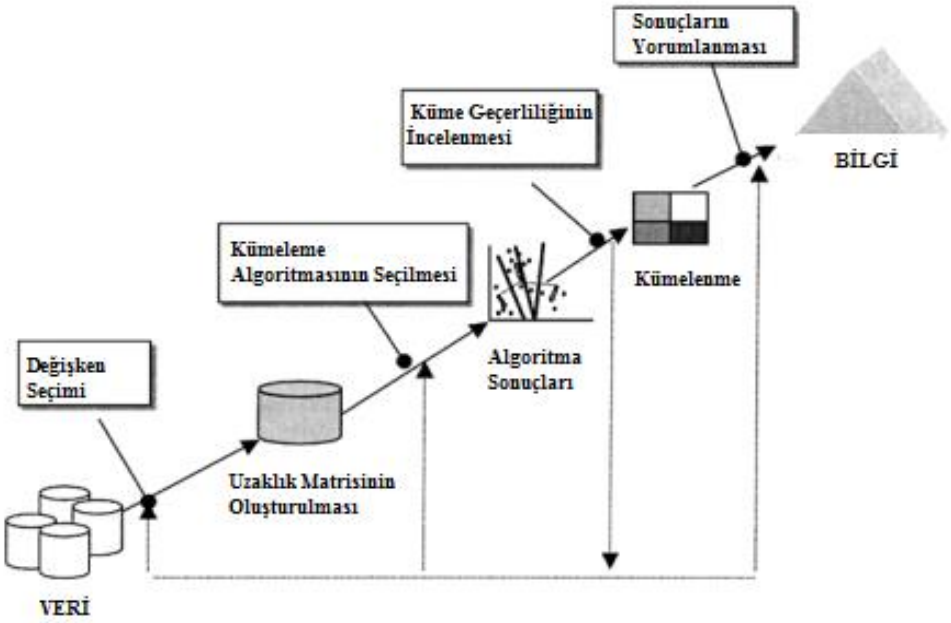
Kümeleme analizi, veri setinin altında yatan doğal grupları ortaya koyan ve birçok alanda yaygın olarak kullanılan çok değişkenli istatistiksel bir yöntemdir. Kümeleme analizi farklı yapıdaki verilerin küme yapısını ve küme sayısını araştırır. Kümeleme analizinin temel amacı, grupları belli olmayan birimleri benzerliklerine göre anlamlı gruplardan oluşan alt gruplara bölerek veriyi indirgemektir. Böylece elde edilen kümelerde küme içi homojenlik ve kümeler arası heterojenlik sağlanır [2].

Kümeleme analizinin uygulanabilmesi için verilerin normal dağılma varsayımı olsa da bu varsayım daha çok teoride kalmakta ve uygulamada kullanılmamaktadır. Uzaklık değerlerinin normal dağılıma uygunluğuna bakılmakta ve varsayımın sağlanması durumunda kovaryans matrisi için farklı bir varsayım gerekmemektedir [44].

Kümeleme analizinin uygulama aşamaları aşağıdaki gibidir;

- Değişkenlerin seçimi ve veri matrisinin oluşturulması
- Birimlerin birbiriyle olan benzerlik ya da uzaklıklarını gösterecek uygun bir benzerlik/uzaklık matrisinin oluşturulması
- Uygun bir kümeleme tekniği ile birimlerin uygun sayıda kümelere ayrılması
- Küme geçerliliği indeksleri ile kümelemenin kalitesinin değerlendirilmesi
- Kümelerin yorumlanmasıdır.

Kümeleme analizinin uygulanma aşamalarına ilişkin diyagram Şekil 2.1'de verilmiştir [23].



Şekil 2.1. Kümeleme Analizinin uygulanma aşamaları

Kümeleme analizinin uygulama aşamaları doğrultusunda, kümeleme algoritmalarının işleyişi ve küme geçerliliğinin kalitesinin belirlenmesine yer verilmiştir.

2.1. Kümeleme Analizinde Değişken Seçimi ve Uzaklık Ölçüleri

Kümeleme analizinin ilk adımında analize dâhil edilecek birimlere ait verilerden hangilerinin kümelerin oluşmasında etkili olup, hangilerinin etkili olmadığı belirlenir. Her kümeleme algoritması iki nokta arasındaki uzaklık veya benzerliği temel alır. Eğer iki nokta arasındaki uzaklık veya benzerlik ölçülmemiş ise geçerli bir kümeleme yapmak mümkün değildir. Kümeleme analizinde özellikle veri setini oluşturan değişkenlerin sürekli olması durumunda benzerliğin belirlenmesinde uzaklık ölçüleri kullanılır.

Kümeleme analizinde hangi teknik kullanılırsa kullanılsın amaç benzer birimleri bir araya getirerek küme içi benzerliği maksimum, kümeler arası benzerliği ise minimum yapmak olduğundan, kümeleme performansı seçilen uzaklık ölçüsü ile doğrudan ilişkilidir. Aşağıda birimler arasındaki benzerliğin belirlenmesinde kullanılan bazı uzaklık ölçüleri verilmektedir.

Öklid uzaklık ölçüsü

Öklid uzaklığı çok boyutlu uzayda iki birey ya da nesne arasındaki uzaklığı hesaplamada en yaygın kullanılan uzaklık ölçüsüdür. Her biri p tane sürekli değişken içeren (x_i,x_j) gözlem çiftleri arasındaki Öklid uzaklığı;

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^{\lambda}} \tag{2.1}$$

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^{2}} \qquad (\ \lambda=2 \text{ durumu için}) \tag{2.2}$$

ile ifade edilir [44].

Minkowski uzaklık ölçüsü

Uzaklıkların belirlenmesinde kullanılan diğer ölçülerden biri de Minkowski uzaklık ölçüsüdür. Bu uzaklık ölçüsü,

$$d(x_i, x_j) = \sum_{k=1}^{p}\left|x_{ik} - x_{jk}\right| \tag{2.3}$$

olarak tanımlanmıştır. Minkowski uzaklık ölçüsü q=1 için City-Block uzaklık ölçüsüne, q=2 için ise Öklid uzaklık ölçüsüne eşittir [1].

City-Block (Manhattan) uzaklık ölçüsü

City-Block uzaklık ölçüsü, değişkenler arasında korelasyon olmadığı durumlarda kullanılan bir uzaklık ölçüsüdür. City-Block uzaklık ölçüsü,

$$d(x_i, x_j) = \left[\sum_{k=1}^{p}\left|x_{ik} - x_{jk}\right|^{q}\right]^{1/q} \tag{2.4}$$

biçiminde ifade edilir.

Mahalanobis uzaklık ölçüsü

Değişkenler arası korelasyon olduğu durumlarda kullanılan bir uzaklık ölçüsüdür. *p* değişkenli bir analizde *i* ve *j* gözlemleri arasındaki Mahalanobis uzaklık ölçüsü,

$$d(x_i, x_j) = (x_i - x_j)' S^{-1} (x_i - x_j) \quad (2.5)$$

eşitliği ile ifade edilir.

Burada *S*, *pxp* tipindeki kovaryans matrisini göstermektedir. Mahalanobis uzaklığının avantajı, aykırı noktaları da hesaplaması ve analize katmasıdır [43].

Pearson Korelasyon Katsayısı

Birimler arasındaki uzaklık Pearson korelasyon katsayısı kullanılarak da hesaplanabilir. Pearson korelasyon katsayısı,

$$r(x_i, x_j) = \frac{\sum_{k=1}^{p}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{p}(x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{p}(x_{jk} - \bar{x}_j)^2}} \quad (2.6)$$

biçiminde tanımlanır.

Burada $\bar{x}_i$ i' inci gözlem üzerinden ölçülen tüm p değişken değerlerinin ortalaması olup

$$\bar{x}_i = \frac{1}{p} \sum_{k=1}^{p} x_{ik} \quad (2.7)$$

biçiminde hesaplanır. Korelasyon katsayısı kullanılarak iki gözlem vektörü arasındaki uzaklık:

$$d(x_i, x_j) = (1 - r(x_i, x_j)) \quad (2.8)$$

eşitliği ile bulunur.

2.2. Kümeleme Analizi Teknikleri

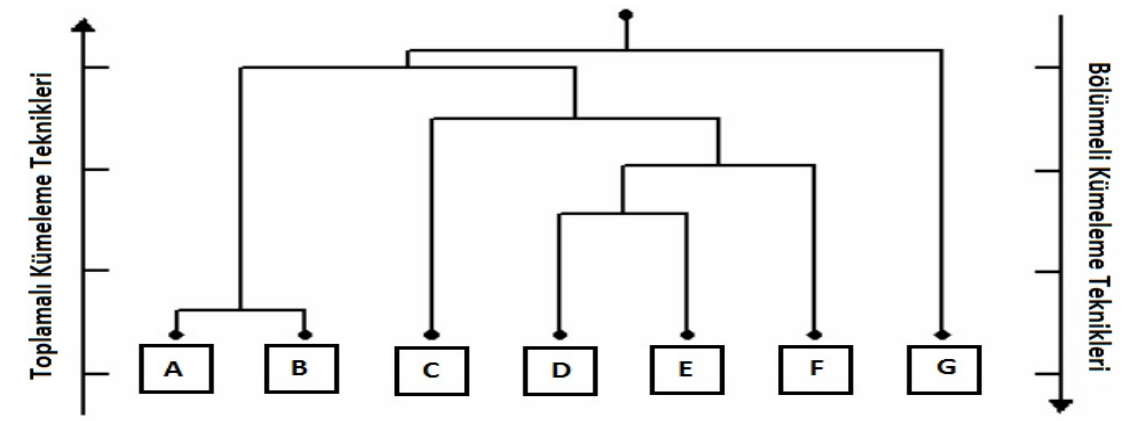
Birimlerin benzerliklerine göre kümelere atanmasında kullanılabilecek çeşitli yaklaşımlar vardır. Bu yaklaşımlardan en yaygın kullanılanları hiyerarşik ve hiyerarşik olmayan kümeleme yöntemleridir. Tüm yaklaşımlarda en önemli ölçüt, kümeler arası farklar ile kümeler içi benzerliklerin maksimum olmasını sağlamaktır.

2.2.1. Hiyerarşik kümeleme yöntemleri

Yöntem, aşamalı kümeleme yöntemi olarak da bilinir [24]. Hiyerarşik kümeleme yöntemleri izledikleri yola göre toplamalı (agglomerative) ve bölünmeli (divisive) aşamalı kümeleme teknikleri olarak iki alt sınıfta incelenmektedir. Hiyerarşik kümeleme algoritmaları, ardışık bir işlem süreci içerir.

Toplamalı hiyerarşik yöntemde her birim veya gözlem başlangıçta bir küme olarak kabul edilir. Daha sonra en yakın iki küme veya gözlem yeni bir kümede toplanarak birleştirilir. Böylece her adımda küme sayısı bir azaltılır.

Bölücü hiyerarşik yöntemde ise süreç toplamalı hiyerarşik yöntemin tam tersidir. Bu yöntemde tüm gözlemlerden oluşan büyük bir küme ile işe başlanır. Benzer olmayan gözlemler ayıklanarak daha küçük kümeler oluşturulur. Her gözlem tek başına küme oluşturana kadar işleme devam edilir. Bu yöntem uygun küme sayısının ne olduğu konusunda bilgi vermez [15]. Bu süreç dendogram veya ağaç diyagramı adı verilen şekille gösterilebilir. Toplamalı ve bölünmeli hiyerarşik kümeleme algoritmalarına ilişkine ağaç diyagramı Şekil 2.2'de verilmiştir.

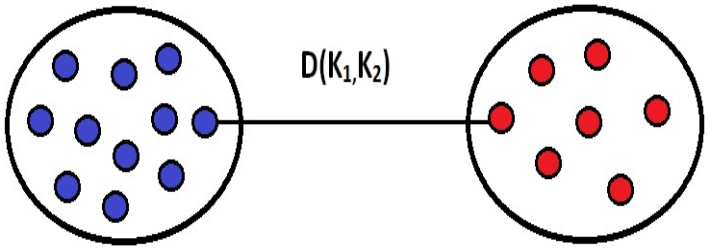


Şekil 2.2. Hiyerarşik kümeleme yöntemlerinin ağaç diyagramı ile gösterimi

Uygulamalarda çoğunlukla toplamalı hiyerarşik kümeleme yöntemi kullanılmaktadır. Toplamalı hiyerarşik yöntemler arasında en çok tek bağlantı (single linkage), tam bağlantı (complete linkage), grup ortalama yöntemi (average linkage) ve Ward yöntemi kullanılmaktadır.

Tek bağlantı tekniği

En yakın komşuluk olarak da bilinen tek bağlantı tekniği, uzaklıklar matrisini kullanarak birbirine en yakın birey ya da nesneleri birleştirmeye dayanmaktadır. Başlangıçta her birim bir küme olarak kabul edilir. Önce birbirine en yakın birim bir kümeye yerleştirilir. Daha sonra diğer en yakın uzaklık tespit edilerek ilk oluşturulan kümeye bu gözlem eklenir veya iki gözlemden oluşan yeni bir küme oluşturulur. Bu işleme tüm birimleri kapsayan bir küme elde edilinceye kadar devam edilir [15]. $D(K_1,K_2)$, K_1 ve K_2 kümeleri arası en yakın iki nesne çiftinin arasındaki uzaklığı ifade etmektedir. Bu tekniğe ilişkin grafiksel gösterim Şekil 2.3'te verilmiştir.



Şekil 2.3. Tek bağlantı tekniği

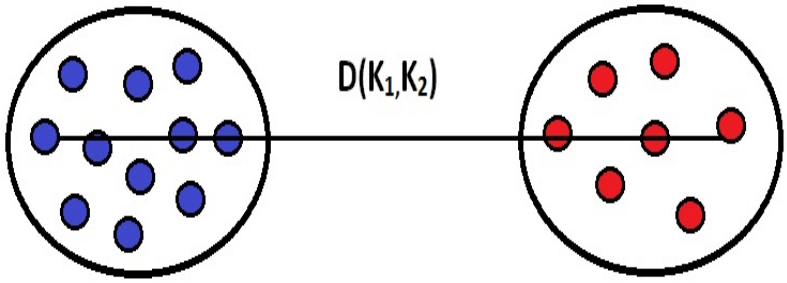
Bu yöntem sapan aykırı değerlerden oldukça fazla etkilenir. Bu sebepten birbirinden oldukça farklı özellikleri bulunan iki küme birleştirilebilir. Kümeler birbirinden oldukça ayrı iseler tek bağlantı yöntemi iyi sonuçlar verir.

Tam bağlantı tekniği

En uzak komşuluk olarak da bilinen bu teknikte tek bağlantıdan farklı olarak, birimler arasındaki maksimum uzaklıklar hesaplanır. Tam bağlantı tekniğinde, bir küme içindeki tüm birimler birbirlerine maksimum uzaklık veya minimum yakınlık ile bağlıdırlar. İşleme

önce birbirine en uzak birimler birleştirilerek başlanır. $D(K_1,K_2)$, K_1 ve K_2 kümeleri arası en uzak iki nesne çiftinin arasındaki uzaklığı ifade etmektedir.

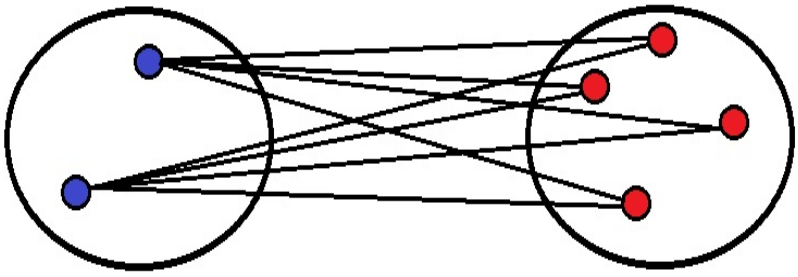
Şekil 2.4'te kümeler arasındaki uzaklıkların tam bağlantı tekniğine göre belirlenmesi grafiksel olarak gösterilmiştir.



Şekil 2.4. Tam bağlantı tekniği

Ortalama bağlantı tekniği

Ortalama bağlantı tekniğinin tek bağlantı ve tam bağlantı tekniklerinden tek farkı, uzaklıkların hesaplanmasındaki farklılıktır. Bu teknikte iki küme arasındaki uzaklık bir kümedeki birimlerin diğer kümedeki birimlere olan uzaklıklarının ortalaması ile elde edilir. Şekil 2.5'te kümeler arasındaki uzaklıkların ortalama bağlantı tekniğine göre belirlenmesi grafiksel olarak gösterilmiştir.



Şekil 2.5. Ortalama bağlantı tekniği

Minimum uzaklığa sahip küme çifti gruplandırılır ve işlem tekrar edilir. Bu teknik küçük ve yaklaşık olarak eşit varyansa sahip kümeler oluşturma eğilimindedir. Ortalama bağlantı tekniği tek bağlantı ve tam bağlantı teknikleri arasında sonuç vermesi nedeniyle alternatif bir yöntem olarak önerilmektedir [25].

Ward yöntemi

Literatürde Ward yöntemi olarak adlandırılan bu teknik en küçük varyanslı kümeleri birleştirdiği için "minimum varyans bağlantı tekniği" olarak da bilinir [17]. Ward yönteminde amaç, kümeler içindeki varyansı minimum kılmak olduğundan, küme içi kareler toplamı minimum olan (grup içi varyans minimum) iki kümeyi birleştirmeye çalışmaktadır. Kısacası teknik varyansları esas alarak birleştirme işlemini yapar ve birleştirme işlemine değişkenliği en az olan kümeler ile başlar [29].

2.2.2. Hiyerarşik olmayan kümeleme yöntemleri

Birim sayısının fazla olması durumunda, birimleri adım adım birbirine bağlayarak küme yapısını ortaya çıkarmak oldukça zaman almaktadır. Kümelenmesi gereken çok sayıda birimin bulunduğu uygulamalarda, ağaç diyagramının oluşturulması ağır bir iş yükü getirdiğinden hiyerarşik algoritmalar yerine hiyerarşik (aşamalı) olmayan kümeleme yöntemleri kullanılmaktadır [14]. Küme sayısı konusunda ön bilgi varsa veya araştırmacı anlamlı olacak küme sayısına karar vermiş ise hiyerarşik olmayan kümeleme yöntemleri tercih edilmektedir.

Hiyerarşik olmayan kümeleme yöntemlerinde küme sayısı *k* önceden tanımlanarak, *k* adet küme merkezi belirlenir ve *n* birimden oluşan veri seti *k* kümeye bölünür. Hiyerarşik olmayan kümeleme teknikleri küme merkezlerinin kümeyi temsil etmesi esasına dayanır. Hiyerarşik olmayan kümeleme yöntemlerinin en önemli dezavantajı küme sayılarının analizin başında belirlenmesi ve küme seçimlerinin keyfi olmasıdır [2]. Hiyerarşik olmayan kümeleme yöntemlerinden en çok kullanılanı k-ortalama yöntemidir.

k-Ortalama kümeleme tekniği

k-ortalama kümeleme yöntemi, hiyerarşik olmayan kümeleme yöntemlerinden en yaygın kullanılan kümeleme algoritmasıdır. *k*-ortalama yönteminde amaç diğer kümeleme yöntemlerinde olduğu gibi, kümeleme işlemi sonucunda elde edilen kümelerin, küme içi benzerliklerinin maksimum, kümeler arası benzerliklerinin ise minimum olmasını sağlamaktır. Küme benzerliği, kümenin ağırlık merkezi kabul edilen bir birim ile kümedeki diğer birimler arasındaki uzaklıkların ortalama değeri ile ölçülmektedir [11].

k − ortalama yöntemi, birimleri kümelerin önceden belirlenen sayısına göre gruplandırmakla işleme başlar. Böylece her biri tek gözlemden oluşan k tane küme ile isleme başlanır ve her bir yeni gözlem en yakın ortalamalı gruba eklenir. Gruba yeni bir gözlem eklendikten sonra küme ortalaması yeniden hesaplanır. Bu süreç tüm gözlemler gruplara atanıncaya kadar devam eder. Tüm gözlemler gruplara atandıktan sonra atandıkları küme ortalamasından daha yakın küme ortalaması varsa, gözlemlerin yerleri değiştirilmektedir. Süreç küme elemanlarının yerlerinin değişmez olmasına kadar tekrarlanır.

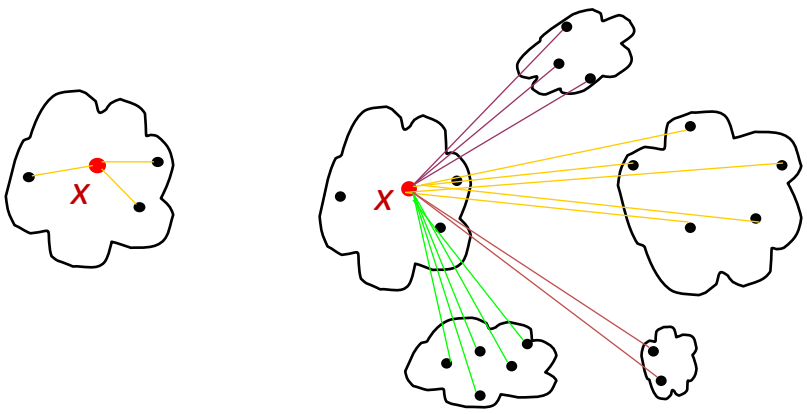
3. KÜME GEÇERLİLİK İNDEKSLERİ

Kümeleme analizinde anlamlı ve geçerli sonuçlara ulaşmada en uygun küme sayısının belirlenmesi birçok araştırmacının sıklıkla karşılaştığı önemli sorunlardan biridir. Özellikle k-ortalamalar gibi bazı kümeleme teknikleri analizin en başında küme sayısının belirlenmesini gerekli kılmaktadır [9]. Bir veri seti için farklı kümeleme algoritmalarının kullanılması sonucu farklı küme yapıları oluşabileceği için, oluşan küme yapılarının kalitesinin sorgulanması büyük önem taşımaktadır. Aynı veri kümesinden elde edilen farklı küme yapılarının anlamlı olup olmadığının sınanması kümeleme geçerliliği (cluster validity) olarak adlandırılmış ve analiz sonucu oluşan küme yapılarının kalitesi için çeşitli geçerlilik ölçütleri geliştirilmiştir. Böylece eldeki gözlemler ya da değişkenler için doğal küme yapısının ve küme sayısının ne olması gerektiği ortaya koyulabilmekte ve alınan kararlar küme geçerlilik kriterleri ile desteklenebilmektedir [6]. Ancak bazı karmaşık yapılar içeren verilerde, küme üyeliklerindeki kararsızlıklar nedeniyle, küme geçerlilik indeksleri uygun küme sayısının belirlenmesinde birbirleriyle çelişen sonuçlar verebilmektedir. Bu bölümde kümeleme kalitesinin belirlenmesinde ve en uygun küme sayısının belirlenmesinde kullanılan küme geçerlilik indeksleri tanıtılacaktır.

Kümeleme planlaması ve küme geçerlilik indeksleri temelde iki kritere dayanmaktadır.

- Yoğunluk: küme içerisindeki birimlerin birbirlerine yakınlıklarını ölçer.
- Ayrılma: iki kümenin birbirinden ne kadar iyi ayrıldıklarını gösterir. İki farklı küme arasındaki mesafeyi ölçer [27,29].

Küme içi ve kümeler arasındaki uzaklıkların grafiksel gösterimi Şekil 3.1'de verilmiştir.



Şekil 3.1. Küme içi ve kümeler arası uzaklıklar

Genel anlamda, uygun küme sayısının belirlenmesi ve kümelemenin geçerliliği için dışsal (external) kriterler ve içsel (internal) kriterler olarak adlandırılan iki farklı kümeleme geçerlilik ölçütü bulunmaktadır. İçsel kriterler, veri seti ile kümeleme yapısı arasındaki uyumun belirlenmesinde sadece veri setindeki doğal yapıyı ve nicel değerleri göz önünde bulundurarak kümeleme sonuçlarını değerlendirir. İçsel kriterlerin çoğu, kümeler içi kareler toplamını veya kümeler arası kareler toplamını temel alarak değerlendirmeyi yapmaktadır [44].

Kümeleme kalitesinin değerlendirilmesinde kullanılan dışsal kriterler, önceden yapılmış ve bilinen bir sınıflamayı belli bir kümeleme algoritmasının çalıştırılmasıyla elde edilen kümeleme sonuçlarıyla karşılaştıran ölçütlerdir [12]. Bu durumda kümelerde yer alan birimlerin hangi kümeye ait oldukları daha önceden bilinmekte ve uygulanan kümeleme algoritması için referans olarak kullanılmaktadır.

Kümeleme analizi çalışmalarında, uygun küme sayısının belirlenmesinde daha çok içsel kriterlerin kullanıldığı görülmektedir. Bunun nedeni, içsel geçerlilik kriterlerinin dışsal kriterlere göre doğal ve anlamlı küme yapıları oluşturmada daha başarılı sonuçlar vermesidir [3]. Bu çalışmada da içsel küme geçerlilik kriterleri kullanılmıştır.
Uygun küme sayısının belirlenmesinde yaygın olarak kullanılan bazı içsel küme geçerlilik kriterleri aşağıda verilmiştir.

3.1. Calinski-Harabasz İndeksi

Calinski ve Harabasz (1974), k kümeye sahip bir kümelemenin geçerliliğini değerlendirebilmek için,

$$CH(k) = \frac{BSS(k)/(k-1)}{WSS(k)/(n-k)} \tag{3.1}$$

indeksini önermişlerdir. Burada,

$$WSS(k) = \frac{1}{2}\sum_{i=1}^{k}\sum_{i,j\in C_i} d(i,j) \tag{3.2}$$

$$BSS(k) = \frac{1}{2}\sum_{i=1}^{k}\sum_{\substack{i\in C_i \\ j\notin C_i}} d(i,j) \tag{3.3}$$

olmak üzere, *WSS*(*k*) ve *BSS*(*k*) sırasıyla, kümeler içi ve kümeler arası kareler toplamıdır. Bu kritere göre, CH indeks değerini maksimum yapan küme sayısı, uygun küme sayısı olarak kabul edilir [7].

3.2. Krzanowski-Lai İndeksi

Krzanowski ve Lai (1988), küme içi kareler toplamı (*WSS*) değerinin azalışını kullanarak,

$$DIFF(k) = (k-1)^{\frac{2}{p}}WSS(k-1) - k^{\frac{2}{p}}WSS(k) \tag{3.4}$$

istatistiğini tanımlamışlardır ve bu istatistiğe bağlı olarak,

$$KL(k) = \left|\frac{DIFF(k)}{DIFF(k+1)}\right| \tag{3.5}$$

indeksini önermişlerdir. *k, KL* indeksi için uygun küme sayısı ve p, değişken sayısı olmak üzere, *KL* indeksinde uygun küme sayısının belirlenmesindeki temel düşünce, *WSS*(*k*) değerinin uygun küme sayısına ulaşana kadar hızlı bir şekilde azaldığı ve uygun küme sayısından sonra yavaş bir şekilde azaldığıdır [28]. Yani WSS(*k*) değerinin en hızlı azalış değerine ulaştığı küme sayısı uygun küme sayısı olarak tanımlanır. Krzanowski-Lai indeks değerini maksimum yapan k uygun küme sayısı olarak belirlenebilir [24].

3.3. Davies-Bouldin İndeksi

Davies-Bouldin (1979) tarafından önerilen indeks, küme içindeki gözlemlerin küme merkezine olan uzaklıklarını minimum yapmayı ve kümeler arası uzaklıkları maksimum yapmayı amaçlar [10].

i=1,2,…,k ve j=1,2,…,k olmak üzere i. ve diğer kümeler arasındaki maksimum karşılaştırma oranı her bir küme için R_i ile gösterilen küme indeksi,

$$R_i = \max_{i \neq j} \left(\frac{\delta_i + \delta_j}{d_{ij}} \right) \tag{3.6}$$

eşitliği ile hesaplanır.

Burada,

d_{ij} : i. ve j. küme merkezleri arasındaki uzaklığı

δ_i ve δ_j : i. ve j. kümedeki gözlemlerin kendi küme merkezlerine olan ortalama uzaklıklarını göstermektedir. Davies-Bouldin indeksi;

$$DB(k) = \frac{1}{k} \sum_{i=1}^{k} R_i \tag{3.7}$$

ile tanımlanır. Davies-Bouldin indeks değerini minimum yapan k uygun küme sayısı olarak belirlenebilir.

3.4. Silhouette İndeksi

Rousseeuw (1987), kümelenecek her bir birimin atandığı kümeye uygunluğunu tanımlamak amacıyla Silhouette indeksi önermiştir.

a(i); i. birimin kendi kümesindeki tüm noktalara olan ortalama uzaklıklarını

b(i); i. birimin diğer kümelerdeki tüm noktalara olan ortalama uzaklıkların minimumunu göstermek üzere, Silhouette indeksi;

$$S(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)} \tag{3.8}$$

şeklinde tanımlanır.

Hesaplanan i. birime ait $S(i)$ değer 1'e yaklaşırsa i. birimin atandığı kümeye daha uyduğu, $S(i)$ değeri 0'a yaklaşırsa veya negatif olursa i. birimin atandığı kümeye uygun olmadığı

sonucuna varılır [41]. Negatif değerler yalnızca bir birim en uygun kümesine atanamadığında ortaya çıkar. Yapılan tüm kümelemenin kalitesi (geçerliliği) için ise ortalama Silhouette değeri,

$$S = \frac{1}{n}\sum_{s_i \in S} S(i) \tag{3.9}$$

şeklinde tanımlanır. Bu kritere göre, maksimum ortalama Silhouette değerine ulaşılan küme sayısı uygun küme sayısı olarak alınır.

3.5. Dunn İndeksi

Dunn (1974), yüksek yoğunluklu iyi ayrımlı kümelerin belirlenmesinde oldukça etkin bir kriter önermiştir. İndeks değeri hesaplanırken, K_i ve K_j gösterilen iki küme arasındaki uzaklığı D(K_i,K_j), sırasıyla K_i ve K_j kümelerinde yer alan x ve y gözlem çiftleri arasındaki minimum uzaklık,

$$D\left(K_i, K_j\right) = \min_{x \in K_i, y \in K_j} d(x, y) \tag{3.10}$$

ile tanımlanır. K kümesinin çapı ise bu küme içerisinde yer alan gözlem çiftleri arasındaki maksimum uzaklık olarak,

$$Çap(K_i) = \max_{x,y \in K_i} d(x, y) \tag{3.11}$$

ile tanımlanır [13]. Bu tanımlamalar altında Dunn indeksi,

$$DN(k) = \min_{i=1,2,\ldots k}\left(\min_{j=1,2,\ldots k}\left(\frac{D\left(K_i, K_j\right)}{\max_{i=1,2,\ldots k} Çap(K_i)}\right)\right) \tag{3.12}$$

ile ifade edilir. Dunn indeksi değeri daha büyük olan kümeleme, geçerli kümeleme olarak kabul edilir [24].

3.6. Gamma İndeksi

Baker ve Hubert (1975) tarafından önerilen Gamma indeksi,

$$Gamma = \frac{s(+) - s(-)}{s(+) + s(-)} \tag{3.13}$$

ile tanımlanır. Burada s(+) tutarlı karşılaştırma sayısını, s(-) tutarsız karşılaştırma sayısını ifade eder. Karşılaştırmalar tüm kümeler içi ve kümeler arası farklılıklar arasında yapılır. Eğer bir karşılaştırmada küme içi farklılıklar kümeler arası farklılıklardan daha az ise karşılaştırma tutarlıdır[19]. Maksimum Gamma indeks değerine ulaşılan küme sayısı uygun küme sayısıdır [37].

3.7. Tau İndeksi

Rohlf (1974) tarafından önerilen Tau indeksi,

$$Tau = \frac{s(+) - s(-)}{[N_t(N_t - 1)/(2 - t)(\frac{N_t(N_t-1)}{2})]^{1/2}} \tag{3.14}$$

ile tanımlanır [40]. Burada t, karşılaştırma sayısını; N_t, gözlem çiftlerinin sayısını ve n, örnek hacmini göstermektedir.

$$N_t = \frac{n(n-1)}{2} \tag{3.15}$$

Tau indeks değerini maksimum yapan küme sayısı uygun küme sayısıdır [8].

3.8. McClain İndeksi

McClain ve Rao (1975) tarafından önerilen indeks,

$$McClain = \frac{\bar{S}_w}{\bar{S}_b} = \frac{S_w/N_w}{S_b/N_b} \tag{3.16}$$

ile ifade edilmektedir [34].

Burada N_w aynı kümeye ait gözlem çiftlerinin sayısı, N_b farklı kümelere ait gözlem çiftlerinin sayısı, S_w küme içi uzaklıklar toplamı, S_b kümeler arası uzaklıklar toplamı olmak üzere,

$$S_w = \sum_{k=1}^{K} \sum_{\substack{i,j \in C_k \\ i<j}} d(x_i, x_j) \tag{3.17}$$

$$S_b = \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \sum_{\substack{i \in C_k \\ j \in C_l}} d(x_i, x_j) \tag{3.18}$$

$$N_w = \sum_{k=1}^{K} \frac{n_k(n_k - 1)}{2} \tag{3.19}$$

$$N_b = N_t - N_w \tag{3.20}$$

eşitlikleri ile tanımlanır. McClain indeksine göre en küçük indeks değeri uygun küme sayısını belirlemek için kullanılır [8].

3.9. Point-Biserial İndeksi

Milligan (1980,1981) ve Kramer (1982) tarafından önerilen indeks,

$$Ptbiserial = \frac{[\bar{S}_b - \bar{S}_w][N_w N_b / N_t^2]^{1/2}}{s_d} \tag{3.21}$$

ile ifade edilir [35,36]. PtBiserial indeksine göre indeks maksimum değeri uygun küme sayısını belirlemek için kullanılır [37].

3.10. Gplus İndeksi

Rolh (1974) tarafından önerilen indeks,

$$Gplus = \frac{2s(-)}{N_t(N_t - 1)} \tag{3.22}$$

ile tanımlanmaktadır. Gplus indeksinin minimum değeri uygun küme sayısını belirlemede kullanılır [40].

3.11. Ratkowsky İndeksi

Ratkowsky ve Lance (1978) tarafından önerilen indeks,

$$Ratkowsky = \frac{\bar{S}}{k^{1/2}} \tag{3.23}$$

ile ifade edilir [39]. Burada $\bar{S}$, BSS_j ve TSS_j sırasıyla, kümeler arası kareler toplamı ve her değişkenin kareler toplamı olmak üzere,

$$\bar{S} = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\frac{BSS_j}{TSS_j}} \tag{3.24}$$

$$BSS_j = \sum_{k=1}^{j} n_k (c_{kj} - \bar{x}_j)^2 \tag{3.25}$$

$$TSS_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \tag{3.26}$$

eşitlikleri ile tanımlanır. Bu kritere göre maksimum Ratkowsky indeks değerine ulaşılan küme sayısı uygun küme sayısı olarak alınır.

3.12. Kübik Kümeleme Kriteri (CCC)

Sarle (1983) tarafından önerilen indeks,

$$CCC = In\left[\frac{1 - E(R^2)}{1 - R^2}\right] \frac{\sqrt{\frac{np^*}{2}}}{(0.001 + E(R^2))^{1.2}} \tag{3.27}$$

ile hesaplanır. Eşitlik 3.27'de yer alan

$$R^2 = 1 - \frac{\text{İz}(X^T\text{X} - \bar{X}^T Z^T Z \bar{X})}{\text{İz}(X^T\text{X})} \tag{3.28}$$

$$E(R^2) = 1 - \left[\frac{\sum_{j=1}^{p^*} \frac{1}{n+u_j} + \sum_{j=p^*+1}^{p} \frac{u_j^2}{n+u_j}}{\sum_{j=1}^{p} u_j^2}\right]\left[\frac{(n-k)^2}{n}\right]\left[1 + \frac{4}{n}\right] \tag{3.29}$$

olmak üzere, burada

- $X^T\text{X}$, (*pxp*)'lik örnek kareler toplamı matrisi,
- *Z*, (*nxk*)'lık küme gösterge matrisi olmak üzere $\bar{X}=(Z^TZ)^{-1}Z^TX$,
- $u_j = \frac{s_j}{c}$,
- $c = (\frac{v^*}{k})^{\frac{1}{p^*}}$,
- $v^* = \prod_{j=1}^{p^*} s_j$ ile ifade edilir.

CCC indeksine göre maksimum indeks değerine ulaşılan k uygun küme sayısını belirlemede kullanılır [42].

3.12. C İndeksi

Hubert ve Levin (1976) tarafından önerilen C indeks, S_{min} gözlem çiftleri arasındaki en küçük uzaklıklar toplamı, S_{max} gözlem çiftleri arasındaki en büyük uzaklıklar toplamı olmak üzere,

$$Cindeks = \frac{S_w - S_{min}}{S_{max} - S_{min}},\ S_{max} \neq S_{min} \tag{3.30}$$

ile ifade edilir. Bu kritere göre indeksin minimum değeri uygun küme sayısını belirlemede kullanılır [25].

3.13. SD İndeks

SD indeks, küme içi ortalama yoğunluk ve kümeler arası uzaklığa dayanan bir indekstir. SD indeks,

$$SD(k) = \alpha K\text{üme içi } yoğunluk(k) + Uzaklık(k) \tag{3.31}$$

eşitliği ile tanımlanır. Burada α, Uzaklık(kmax) ile ifade edilmektedir.
Küme içi ortalama yoğunluk,

$$K.\text{İ}.Yoğunluk(k) = \frac{\frac{1}{K}\sum_{k=1}^{K}\left\|v^{(k)}\right\|}{\|v\|} \tag{3.32}$$

ile hesaplanır. Burada,

- $v = (\text{Var}(V_1), \text{Var}(V_2), \ldots, \text{Var}(V_p))$ değişken varyans vektörü,
- $v^{(k)} = (\text{Var}(V_1^{(k)}), \ldots, \text{Var}(V_p^{(k)}))$ küme varyans vektörünü ifade etmektedir.

Kümeler arasındaki uzaklık,

$$Uzaklık(k) = \frac{D_{max}}{D_{min}} \sum_{k=1}^{K} \frac{1}{\sum_{\substack{k'=1 \\ k' \neq k}}^{K} \left\| C^{(k)} - C^{(k')} \right\|} \tag{3.33}$$

ile hesaplanır. Burada

- $D_{max} = \underset{k \neq k'}{max} \left\| C^{(k)} - C^{(k')} \right\|$ küme merkezleri arasındaki maksimum uzaklığı,
- $D_{min} = \underset{k \neq k'}{min} \left\| C^{(k)} - C^{(k')} \right\|$ küme merkezleri arasındaki minimum uzaklığı ifade eder.

SD indekse göre minimum indeks değeri uygun küme sayısını belirlemek için kullanılmaktadır [22].

3.14. SDbw İndeksi

SDbw küme geçerlilik indeksi kümeler arasındaki yoğunluk ve ayrıma dayanan bir indekstir. SDbw indeksi,

$$SDbw(k) = Küme\ içi\ yoğunluk(k) + Kümeler\ arası\ yoğunluk(k) \tag{3.34}$$

eşitliği ile tanımlanmaktadır. Kümeler arası yoğunluk,

$$K.A.Yoğunluk(k) = \frac{1}{k(k-1)} \sum_{i=1}^{k} [\sum_{j=1, i \neq j}^{k} \frac{Yoğunluk(u_{ij})}{max(yoğunluk(c_i), yoğunluk(c_j))}] \tag{3.35}$$

ile ifade edilir. Burada c_i ve c_j küme merkezleri, u_{ij} ise c_i ve c_j küme merkezlerinin orta noktasını ifade eder.

$$Yoğunluk(u_{ij}) = \sum_{l=1}^{n_{ij}} f(x_l, u_{ij}) \tag{3.36}$$

ile hesaplanır. Burada $f(x_l, u_{ij})$, $d(x, u_{ij}) > \sigma$ ise 0; aksi taktirde 1'dir.

$$\sigma = \frac{1}{K}\sqrt{\sum_{k=1}^{K} \|v^{(k)}\|} \tag{3.37}$$

SDbw indeksine göre minimum indeks değeri uygun küme sayısını belirlemek için kullanılmaktadır [23].

3.15. Duda İndeksi

Duda ve Hart (1973) tarafından önerilen Duda indeksi, *Je*(2) veriler iki küme halinde bölümlendiği küme içi kareler toplamı ve *Je*(1) sadece bir küme mevcut olduğunda hata kareleri olmak üzere,

$$Duda = \frac{Je(2)}{Je(1)} = \frac{W_k + W_l}{W_m} \tag{3.38}$$

ile ifade edilmektedir. Burada C_k ve C_l kümelerinin C_m kümesini oluşturduğu varsayılmaktadır.

Duda kritik değeri,

$$Duda\ kritik\ değeri = 1 - \frac{2}{\pi p} - z\sqrt{\frac{2\left(1 - \frac{8}{\pi^2 p}\right)}{n_m p}} \tag{3.39}$$

olmak üzere,

$$Duda \geq Duda\ kritik\ değeri \tag{3.40}$$

Eşitlik 3.40'ta gösterildiği gibi Duda kritik değerinden büyük olan ilk Duda indeks değerini sağlayan küme sayısı uygun küme sayısı olarak alınmaktadır [37].

3.16. PseudoT^2 İndeksi

Duda ve Hart (1973) tarafından önerilen index, genelde hiyerarşik kümeleme yöntemlerinde kullanılmaktadır [11].

PseudoT^2 indeksi,

$$Pseudot2 = \frac{V_{kl}}{\frac{W_k+W_l}{n_k+n_l-2}} \tag{3.41}$$

olarak tanımlanır. Burada n_k ve n_l, sırasıyla C_k ve C_l kümelerinin nesne sayılarıdır ve $V_{kl} = W_m - W_k - W_l$ ile ifade edilir. Pseudot2 kritik değeri,

$$Pseudot2\ kritik\ değeri = \left(\frac{1 - Duda\ kritik\ değeri}{Duda\ kritik\ değeri}\right) x(n_k + n_l - 2) \tag{3.42}$$

olmak üzere,

$$Pseudot2 \leq Pseudot2\ kritik\ değeri \tag{3.43}$$

Eşitlik 3.42'de gösterildiği gibi PreudoT^2 kritik değerinden küçük olan ilk PseudoT^2 indeks değerini sağlayan küme sayısı uygun küme sayısı olarak alınmaktadır [21].

3.17. Gap İndeksi

Tibshirani (2001) tarafından önerilen indeks,

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} logW_{kb} - logW_k \tag{3.44}$$

ile hesaplanır.

Burada B birim sayısı ve W_k küme içi dağılım matrisi,

$$W_k = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - c_k)(x_i - c_k)^T \tag{3.45}$$

olmak üzere, uygun küme sayısı;

$$\text{gap}(k) \geq \text{gap}(k+1) - s_{k+1} \tag{3.46}$$

koşulunu sağlayan en küçük k değeridir. Burada

$$s_k = sd_k \sqrt{1 + \frac{1}{B}} \tag{3.47}$$

$$sd_k = \sqrt{\frac{1}{B} \sum_{b=1}^{B} (logW_{kb} - \bar{l})^2} \tag{3.48}$$

$$\bar{l} = \frac{1}{B} \sum_{b=1}^{B} logW_{kb} \tag{3.49}$$

ile hesaplanır [46]. Gap indeks değerini maksimum yapan k sayısı uygun küme sayısı olarak belirlenir.

3.18. Ball İndeksi

Ball ve Hall (1965) önerilen küme geçerlilik indeksi, küme merkezlerinin ortalamasına dayanmaktadır. Ball indeksi,

$$Ball = \frac{W_k}{k} \tag{3.50}$$

ile ifade edilir [5]. Bu kritere göre hiyerarşi düzeyleri arasındaki en yüksek fark, uygun küme sayısını belirlemek için kullanılır.

3.19. Hubert İstatistiği

Hubert ve Arabie (1985) tarafından önerilen Hubert Γ istatistiği, herhangi iki matris arasındaki korelasyon katsayısına dayanır. İki matris simetrik olduğunda Hubert Γ istatistiği,

$$\Gamma(P,Q) = \frac{1}{N_t}\sum_{\substack{i=1 \\ i<j}}^{n-1} P_{ij}\, Q_{ij} \tag{3.51}$$

eşitliği ile ifade edilir [26]. Burada *P* veri setinin uzaklık matrisini, *Q* (*nxn*) boyutlu özellik matrisini ifade eder. k =1 ve k=n olduğunda indeks tanımlı değildir.
Normalleştirilmiş Hubert Γ istatistiği,

$$\bar{\Gamma} = \frac{\sum_{\substack{i=1 \\ i<j}}^{n-1}(P_{ij}-\mu_P)(Q_{ij}-\mu_Q)}{\sigma_P \sigma_Q} \tag{3.52}$$

ile ifade edilir. $\mu_P, \mu_Q, \sigma_P \; ve \; \sigma_Q$ P ve Q matrislerinin ortalama ve varyanslarını ifade eder. Hubert Γ istatistiği -1 ile 1 arasında değer almaktadır. Γ İstatistiğinin yüksek değerler alması kompakt kümelerin varlığına işaret eder. Hubert indeks değerleri grafiğinin diz oluşturduğu küme sayısı, veri altında yatan kümelerin sayısını gösterir. Hubert istatistiğinin ikinci farklılıklar grafiğinde oluşan diz uygun küme sayısını ayırt etmeye yardımcı olur. Bu grafiklerde önemli bir zirve uygun küme sayısını göstermektedir [23].

3.20. D İndeks

Lebart (2000) tarafından önerilen D indeks, küme içi durağanlığa dayanmaktadır. Küme içi durağanlık, bir kümedeki veriler arasındaki homojenlik derecesini ölçer [31]. Küme içi durağanlık,

$$w\left(P^k\right) = \frac{1}{k}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{x_i \in C_k} d(x_i, c_k) \tag{3.53}$$

eşitliği ile tanımlanır.

Küme içi durağanlık üzerine küme kazancı,

$$kazanç = w\left(P^{k-1}\right) - w\left(P^{k}\right) \tag{3.54}$$

ile ifade edilir. Kümeleme kazancı minimize edilmelidir [31]. D indeks grafiğinde küme sayısına karşın küme kazancı önemli bir azalmaya karşı keskin bir diz oluşturabilir. D indeksin ikinci farklılıklar grafiğinde oluşan diz veya kazanç değerlerindeki büyük sıçrama uygun küme sayısını ayırt etmeye yardımcı olur. Bu grafiklerde belirgin bir zirve uygun küme sayısını göstermektedir [8].

Çizelge 3.1'de içsel küme geçerlilik indekslerinin tanım ve en uygun küme seçimi kriterlerinin yer aldığı özet tablosu verilmektedir.

Çizelge 3.1. İçsel küme geçerlilik indeksleri

İNDEKS	SEMBOL	TANIM	OPTİMAL DEĞER
Krzanowski-Lai	KL	$KL = \left\lvert \frac{DIFF(k)}{DIFF(k+1)} \right\rvert$ $DIFF(k) = (k-1)^{\frac{2}{p}} WSS(k-1) - k^{\frac{2}{p}} WSS(k)$	Maksimum
Calinski-Harabasz	CH	$CH = \frac{BSS(k)/(k-1)}{WSS(k)/(n-k)}$	Maksimum
Silhouette	S	$S(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$ $S = \frac{1}{n} \sum_{s_i \in S} S(i)$	Maksimum
Dunn	DN	$DN = \min_{i=1,2,\dots k} \left(\min_{j=1,2,\dots k} \left(\frac{D\left(K_i, K_j\right)}{\max_{i=1,2,\dots k} Çap(K_i)} \right) \right)$	Maksimum
Gamma	Gamma	$Gamma = \frac{s(+) - s(-)}{s(+) + s(-)}$	Maksimum
Tau	Tau	$Tau = \frac{s(+) - s(-)}{[N_t(N_t-1)/(2-t)(\frac{N_t(N_t-1)}{2})]^{1/2}}$	Maksimum
Point-Biserial	Pt	$Pt = \frac{[\bar{S}_b - \bar{S}_w][N_w N_b / N_t^2]^{1/2}}{s_d}$	Maksimum

Çizelge 3.1. (devam) İçsel küme geçerlilik indeksleri

Ratkowsky	R	$R = \frac{\bar{S}}{k^{1/2}}$	Maksimum
Kübik Kümeleme Kriteri	CCC	$CCC = In\left[\frac{1-E(R^2)}{1-R^2}\right]\frac{\sqrt{\frac{np^*}{2}}}{(0.001+E(R^2))^{1.2}}$ $R^2 = 1 - \frac{\text{İz}(X^T\text{X} - \bar{X}^T Z^T Z\bar{X})}{\text{İz}(X^T\text{X})}$ $E(R^2) = 1 - \left[\frac{\sum_{j=1}^{p^*}\frac{1}{n+u_j} + \sum_{j=p^*+1}^{p}\frac{u_j^2}{n+u_j}}{\sum_{j=1}^{p}u_j^2}\right]\left[\frac{(n-k)^2}{n}\right]\left[1 + \frac{4}{n}\right]$	Maksimum
Davies-Bouldin	DB	$DB = \frac{1}{k}\sum_{i=1}^{k} R_i$ $R_i = \underset{i \neq j}{max}\left(\frac{\delta_i + \delta_j}{d_{ij}}\right)$	Minimum
McClain	McClain	$McClain = \frac{\bar{S}_w}{\bar{S}_b} = \frac{S_w/N_w}{S_b/N_b}$	Minimum
Gplus	Gplus	$Gplus = \frac{2s(-)}{N_t(N_t-1)}$	Minimum
SD indeks	SD	$SD = \alpha K.\text{İ}.yoğunluk(k) + uzaklık(k)$ $K.\text{İ}.yoğunluk(k) = \frac{\frac{1}{K}\sum_{k=1}^{K}\|v^{(k)}\|}{\|v\|}$ $Uzaklık(k) = \frac{D_{max}}{D_{min}}\sum_{k=1}^{K}\frac{1}{\sum_{\substack{k'=1 \\ k'\neq k}}^{K}\|C^{(k)} - C^{(k')}\|}$	Minimum
SDbw	SDbw	$\text{SDbw} = K.\text{İ}.yoğunluk(k) + K.A.yoğunluk(k)$ $K.\text{İ}.yoğunluk(k) = \frac{1}{k(k-1)}\sum_{i=1}^{k}[\sum_{j=1,i\neq j}^{k}\frac{yoğunluk(u_{ij})}{\max\left(yoğ(c_i), yoğ(c_j)\right)}]$ $yoğunluk(u_{ij}) = \sum_{l=1}^{n_{ij}} f(x_l, u_{ij})$ $\sigma = \frac{1}{K}\sqrt{\sum_{k=1}^{K}\|v^{(k)}\|}$	Minimum
Ball	Ball	$Ball = \frac{W_k}{k}$	Maksimum fark

Çizelge 3.1. (devam) İçsel küme geçerlilik indeksleri

Gap İstatistiği	Gap	$Gap = \frac{1}{B}\sum_{b=1}^{B} logW_{kb} - logW_k$	Kritik değer ≥ 0
Duda	Duda	$Duda = \frac{W_k + W_l}{W_m}$	Duda ≥ Duda kritik değeri
PseudoT^2	Pseudot2	$Pseudot2 = \frac{V_{kl}}{\frac{W_k+W_l}{n_k+n_l-2}}$	Pseudot2< Pseudot2 kritik değeri
Hubert İstatistiği	Γ	$\Gamma = \frac{1}{N_t}\sum_{\substack{i=1 \\ i<j}}^{n-1} P_{ij}\, Q_{ij}$	Grafik Yöntemi
D indeks	W	$W = \frac{1}{k}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{x_i \in C_k} d(x_i, c_k)$	Grafik Yöntemi

4. KÜME GEÇERLİLİK İNDEKSLERİNİN KARŞILAŞTIRILMASI

Çalışmanın bu bölümünde, uygun küme sayısının belirlenmesinde kullanılan içsel küme geçerlilik indekslerinin karşılaştırılmasına yer verilmiştir. Küme geçerlilik indekslerinin karşılaştırılmasında R ortamında elde edilen yapay veri setleri ve bir gerçek veri seti kullanılmıştır. Uygulamanın ilk bölümünde dört farklı kümelenmiş yapay veri seti ve Fisher'in İris veri seti, geleneksel hiyerarşik kümeleme yöntemlerinden tek bağlantı yöntemi, tam bağlantı yöntemi ve Ward yöntemi, hiyerarşik olmayan kümeleme yöntemlerinden ise *k*-ortalama yöntemine göre küme geçerlilik indeksleri ile belirlenen uygun küme sayıları karşılaştırılmıştır. Uygulamanın ikinci bölümünde ise Eurostat tarafından belirlenen 26 İstatistiki Bölge Birimi Sınıflandırması (İBBS) Düzey 2 bölgesinin 2014 yılına ait kadın eğitim ve işgücü istatistikleri kullanılarak, 21 içsel küme geçerlilik indeksinin k-ortalama kümeleme yöntemine göre belirlenen uygun küme sayılarına göre kümelenmesine yer verilmiştir. Analizler R programında gerçekleştirilmiştir. Küme geçerlilik indekslerinin hesaplanmasında NbClust paketi ve uzaklık ölçüsü olarak Öklid uzaklık ölçüsü kullanılmıştır.

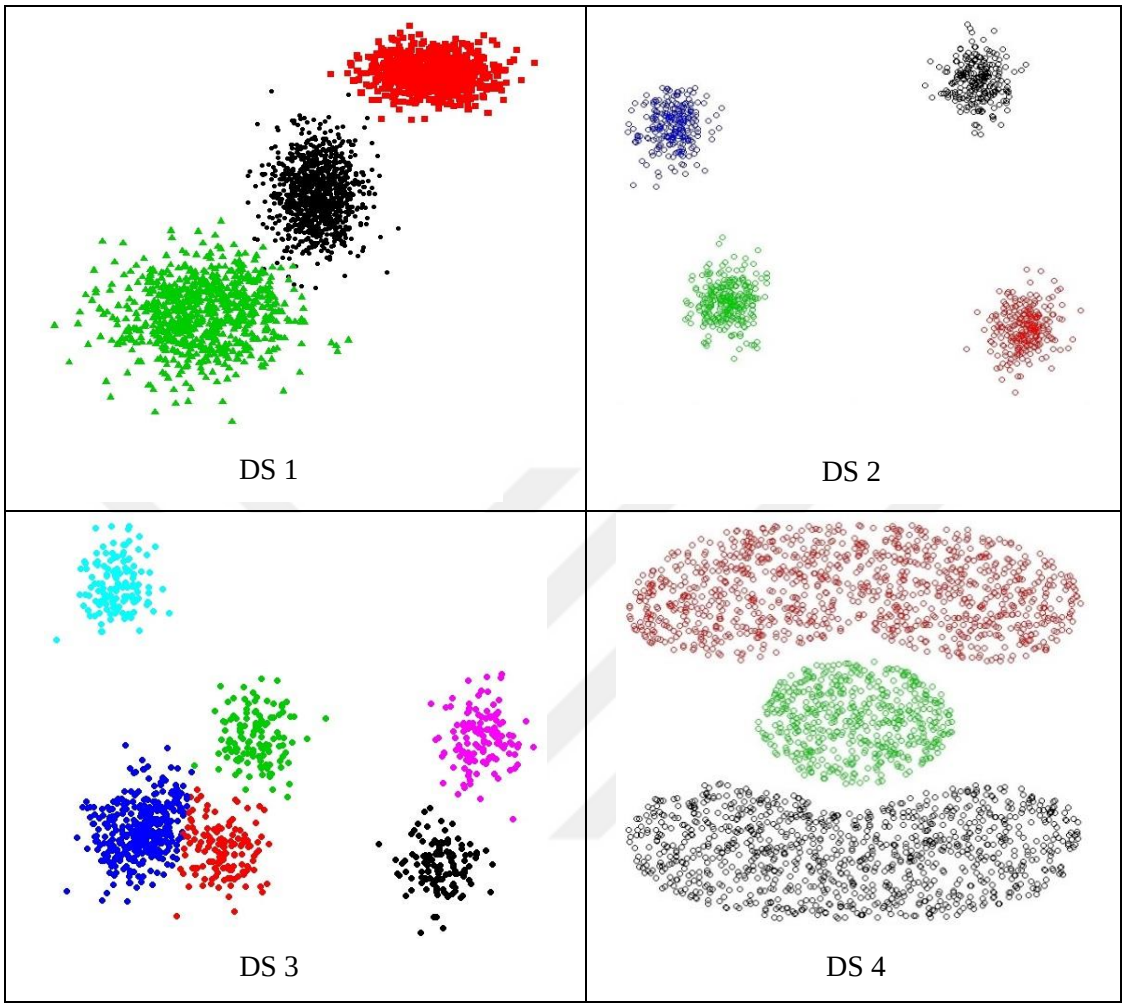
4.1. Yapay Verilere Ait Kümeleme Sonuçlarının Değerlendirilmesi

Uygulamada gerçekte 3, 4 ve 6 kümeli olarak üretilen farklı özelliklerdeki yapay veri setleri, R programının mlbench paketi bünyesinde bulunan 3 kümeli Cassini veri seti ve İris veri seti kullanılmıştır. Uygulamada kullanılan veri setlerinin özellikleri Çizelge 4.1'de verilmiştir.

Çizelge 4.1. Yapay veri setlerinin özelikleri

Veri Seti	*k*	n	Küme Yapısı
DS 1	3	3000	Normal dağılımlı farklı yoğunluklu sentetik veri seti
DS 2	4	1000	Normal dağılımlı iyi ayrılmış sentetik veri seti
DS 3	6	1200	Normal dağılımlı alt kümeli sentetik veri seti
DS 4	3	1500	Uniform dağılımlı Cassini veri seti
DS 5	3	150	İris veri seti

R ortamında elde edilen yapay veri setlerinin serpme grafikleri şekil 4.1'de verilmiştir.



Şekil 4.1. Yapay veri setlerinin serpme grafikleri

Uygun küme sayısının belirlenmesinde kullanılan 21 içsel küme geçerlilik indeksinin farklı kümeleme algoritmaları kullanılarak uygun küme sayılarını belirlemek ve indekslerin uygun küme sayısını belirlemedeki performanslarını karşılaştırmak amacıyla, tek bağlantı, tam bağlantı, Ward ve *k*-ortalama kümeleme yöntemlerine göre uygun küme sayıları belirlenmiştir. Bu yöntemlere göre elde edilen uygun küme sayıları ve indeks değerleri, farklı yoğunluklu (different density) 3 kümeli veri seti, iyi ayrılmış (wellseparated) 4 kümeli veri seti, alt kümeli (subcluster) veri seti, Cassini veri seti ve İris veri seti için sırasıyla verilmiştir. Yapay veri setleri ile yapılan karşılaştırmalarda, çizelgelerde koyu renkle işaretlenen indeks değerleri uygun küme sayısını veren indeks değerlerini göstermektedir.

Çizelge 4.2. 3 kümeli farklı yoğunluklu veri setinin tek bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri

Tek Bağlantı Kümeleme Yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	1,0032	0,9970	**1,0214**	1,0158	0,9871	1,0158	0,9997
CH	6,3073	**7,4484**	4,4398	3,3997	4,0615	3,4103	3,6398
DB	0,4156	**0,3663**	0,7963	1,7382	1,5136	1,6532	1,5035
Silhouette	0,2181	**0,4959**	-0,2010	-0,4792	-0,4917	-0,4941	-0,5289
Dunn	**0,0341**	0,0306	0,0276	0,0240	0,0240	0,0235	0,0204
Gamma	**0,0073**	0,0062	0,0001	0,0016	0,0010	0,0010	0,0039
Tau	23,9501	23,9995	24,0063	24,0091	24,0734	24,0749	**24,1231**
Ratkowsky	0,0290	0,0316	0,0281	0,0314	0,0338	0,0324	**0,0323**
Ptbiserial	0,0541	**0,0654**	0,0524	0,0265	0,0431	0,0377	0,0479
CCC	**89,4009**	40,1563	15,9255	-0,8068	-13,6914	-24,3382	-33,2879
C indeks	**0,2367**	0,2437	0,2438	0,2439	0,2488	0,2488	0,2486
Gplus	0,3038	0,2152	0,7854	0,3904	0,4627	0,4748	**0,1689**
McClain	**0,0003**	0,0007	0,0014	0,0038	0,0040	0,0048	0,0051
SDbw	0,9863	0,6536	0,4539	0,3846	0,4276	0,3997	**0,3082**
SD indeks	0,5228	**0,3739**	0,3990	0,6975	0,9366	1,1742	1,0840
Duda	**1,0073**	1,0062	0,9999	0,9984	1,0010	0,9990	0,9961
PseudoT^2	**-18,1185**	-15,3670	0,2833	4,0643	-2,4398	2,3702	9,6273
Gap	**-0,6303**	-0,9245	-0,9314	-0,9374	-0,9348	-1,8594	-2,3325
Ball	544486,4	**362243,3**	271605,6	217259,8	180566,3	154761,2	135145,7
Hubert	0	**0**	0	0	0	0	0
D indeks	16,8659	**16,8468**	16,8397	16,8327	16,8113	16,8080	16,7895

Çizelge 4.2 incelendiğinde tek bağlantı yöntemine göre, gerçekte 3 kümeli farklı yoğunluklu veri seti için Calinski-Harabasz (CH), Davies-Bouldin (DB), Silhouette, Ptbiserial, SD indeks, Ball, D indeks ve Hubert indeksleri diğerlerine göre daha başarılı bir performans göstererek uygun küme sayısını 3 olarak doğru tahmin etmişlerdir. Dunn, Gamma, CCC, C indeks, McClain, Duda, PseudoT^2 ve Gap indekslerine göre uygun küme sayısı 2 iken, Krzanowski-Lai (KL) indeksine göre uygun küme sayısı 4 olarak tahmin edilmiştir. Ancak dikkat edilecek olursa Tau, Ratkowsky, Gplus ve SDbw indekslerini uygun küme sayısını belirlemedeki performansları önemli derecede düşüktür. Bu indekslere göre uygun küme sayısı 8 olarak belirlenmiştir.

Çizelge 4.3. 3 kümeli farklı yoğunluklu veri setinin tam bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri

Tam bağlantı kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	0,6581	**3,6847**	0,5187	3,1503	0,4255	1,8746	2,7089
CH	3849,985	**9833,924**	5205,327	7542,470	6370,793	8986,464	1413,970
DB	0,6016	**0,5547**	0,5861	0,7118	0,6883	0,6878	0,6854
Silhouette	0,5624	**0,5653**	0,5056	0,4474	0,4013	0,4434	0,4519
Dunn	**0,0071**	0,0044	0,0049	0,0057	0,0064	0,0065	0,0053
Gamma	**2,0400**	0,8038	1,3392	0,6195	5,2484	2,5955	0,7455
Tau	60,6817	154,5851	173,2750	312,6260	328,8661	540,3338	**789,4778**
Ratkowsky	**0,5338**	0,5093	0,4434	0,4119	0,3764	0,3596	0,3377
Ptbiserial	0,6080	**0,6749**	0,6671	0,5645	0,5589	0,5421	0,5193
CCC	**135,7865**	115,9745	91,4068	94,4620	81,8999	88,3821	92,3652
C indeks	0,2498	0,2315	0,2495	0,2619	0,2905	0,2344	**0,2157**
Gplus	0,6411	1,5612	1,2762	1,3941	0,5493	0,7113	**0,5043**
McClain	**0,2590**	0,4329	0,4520	0,7129	0,7281	0,7303	0,7482
SDbw	1,5454	0,3151	0,3078	0,4030	0,3170	0,2491	**0,1892**
SD indeks	0,1778	**0,1387**	0,1839	0,2135	0,2661	0,3290	0,3185
Duda	0,3288	0,5540	0,4273	**0,6170**	0,1598	0,2778	0,5696
PseudoT^2	3794,4444	512,7930	1825,2786	**307,2923**	3595,1526	1461,2523	55,1695
Gap	**0,5099**	0,4736	0,1674	0,1959	-0,1599	0,1049	0,2642
Ball	214900,090	**56238,757**	37629,522	16685,128	13217,649	6895,479	4129,474
Hubert	0	**0**	0	0	0	0	0
D indeks	11,3162	**6,4771**	6,1186	4,9564	4,8171	3,6507	3,0896

Çizelge 4.3 incelendiğinde hiyerarşik kümeleme yöntemlerinden tam bağlantı yöntemine göre, gerçekte 3 kümeli farklı yoğunluklu veri seti için Krzanowski-Lai (KL), Calinski-Harabazs (CH), Davies-Bouldin (DB), Silhouette, Ptbiserial, SD indeks, Ball, Hubert ve D indeks diğer indekslere göre daha başarılı olmakla birlikte uygun küme sayısını 3 olarak doğru tahmin etmişlerdir. Dunn, Gamma, Ratkowsky, CCC, McClain ve gap indekslerine göre uygun küme sayısı 2 iken, Duda ve PseudoT^2 indekslerine göre uygun küme sayısı 5'tir. Ayrıca tek bağlantı kümeleme yönteminde olduğu gibi Tau, C indeks, Gplus ve SDbw indekslerini uygun küme sayısını belirlemedeki oldukça düşük bir performans göstererek, yaklaşık olarak bile tahmin edememiştir ve uygun küme sayısı 8 olarak belirlenmiştir.

Çizelge 4.4. 3 kümeli farklı yoğunluklu veri setinin Ward yöntemine göre uygun küme sayıları ve indeks değerleri

Ward kümeleme yöntemi							
İNDEKS/ *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	0,7178	**131,7384**	0,0952	0,8627	32,0628	0,0477	1,2338
CH	4464,753	**11310,319**	9344,585	9924,537	11042,481	10849,945	10793,830
DB	0,6055	**0,5095**	0,6259	0,6770	0,7367	0,7549	0,7674
Silhouette	0,5827	**0,6498**	0,5809	0,5159	0,4594	0,4528	0,4453
Dunn	**0,0081**	0,0057	0,0052	0,0067	0,0045	0,0045	0,0058
Gamma	**3,7433**	1,7708	1,4896	1,2003	1,2328	1,3265	1,3787
Tau	66,5584	221,0749	292,0743	403,8183	552,5126	647,4287	**814,9571**
Ratkowsky	**0,5654**	0,5374	0,4692	0,4224	0,3877	0,3603	0,3377
Ptbiserial	0,6096	**0,6942**	0,6516	0,6285	0,5242	0,4718	0,4581
CCC	**140,3987**	130,5363	111,3459	103,9629	100,8689	94,9338	93,5094
C indeks	0,2260	0,1792	0,1608	0,1845	0,1494	0,1351	**0,1268**
Gplus	0,4552	1,7956	1,1097	1,9719	1,9178	0,5907	**0,3814**
McClain	**0,3137**	0,4333	0,5056	0,5413	0,7719	0,9348	0,9418
SDbw	1,4351	0,2594	0,3523	0,2886	0,2706	0,2213	**0,1809**
SD indeks	0,2172	**0,1514**	0,2454	0,2287	0,3479	0,5773	0,5631
Duda	0,2107	0,3606	0,4013	0,4542	0,4475	0,4292	**0,4199**
PseudoT^2	6386,1375	1402,4455	954,8238	1277,1458	824,7322	527,9390	**665,8987**
Gap	0,5344	**0,9211**	0,7132	0,4218	0,4261	0,2902	0,2746
Ball	195925,734	**39324,558**	22323,9635	12917,2084	7867,3998	5754,8587	4000,3673
Hubert	0	**0**	0	0	0	0	0
D indeks	10,5242	**5,5698**	4,8245	4,1523	3,5207	3,1841	2,9128

Çizelge 4.4 incelendiğinde hiyerarşik kümeleme yöntemlerinden Ward yöntemine göre, gerçekte 3 kümeli farklı yoğunluklu veri seti için Krzanowski-Lai (KL), Calinski-Harabasz (CH), Davies-Bouldin (DB), Silhouette, Ptbiserial, SD indeks, Gap, Ball, Hubert ve D indeks diğer indekslere göre daha başarılı olmakla birlikte uygun küme sayısını 3 olarak doğru tahmin etmişlerdir. Dunn, Gamma, Ratkowsky, CCC ve McClain indekslerine göre uygun küme sayısı 2 olarak belirlenmiştir. Tau, C indeks, Gplus, SDbw ve diğer iki yöntemin aksine Duda ve PseudoT^2 indeksleri uygun küme sayısını yaklaşık olarak bile tahmin edemeyerek, uygun küme sayısının 8 olduğunu tahmin etmiştir.

Çizelge 4.5. 3 kümeli farklı yoğunluklu veri setinin *k*-ortalama yöntemine göre uygun küme sayıları ve indeks değerleri

***k*-ortalama kümeleme yöntemi**							
İNDEKS/ *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	0,7024	**51,7313**	0,8229	0,9379	1,8389	16,0350	0,0130
CH	4499,156	**14141,125**	10992,513	11821,693	13248,811	10958,133	13873,203
DB	0,7075	**0,5025**	0,5991	0,6767	0,7397	0,7307	0,7441
Silhouette	0,5339	**0,6654**	0,6146	0,5662	0,5089	0,5025	0,4689
Dunn	0,0017	0,0050	**0,0085**	0,0024	0,0026	0,0045	0,0023
Gamma	**2,9881**	0,0261	0,1302	0,2107	0,3739	0,3465	0,3425
Tau	66,8873	233,4651	339,3708	476,4468	658,1356	836,5738	**954,4323**
Ratkowsky	**0,5532**	0,5409	0,4747	0,4270	0,3910	0,3640	0,3397
Ptbiserial	0,6080	**0,6749**	0,6671	0,5645	0,5589	0,5421	0,5193
CCC	**140,6446**	132,7559	117,0773	110,1021	107,2647	104,2204	99,2002
C indeks	0,2730	0,1720	0,1822	0,1681	0,1476	0,1598	**0,1475**
Gplus	**0,2462**	1,4497	1,3981	2,0457	0,7291	1,7115	0,7457
McClain	**0,3703**	0,4167	0,4867	0,5523	0,7512	0,7596	0,8418
SDbw	2,3525	0,2508	0,2970	0,2814	0,3019	0,2460	**0,2148**
SD indeks	0,2247	**0,1409**	0,1965	0,2451	0,3401	0,3708	0,4668
Duda	0,2506	**1,0268**	1,1499	0,8258	1,5988	1,5319	1,5226
PseudoT^2	5495,6355	**-36,1667**	-177,0434	184,5737	-236,3342	-297,9260	-263,5838
Gap	0,3952	**0,8092**	0,6145	0,5007	0,4677	0,3884	0,2480
Ball	194962,427	**37237,580**	19212,776	10948,137	6604,775	4453,714	3415,777
Hubert	0	**0**	0	0	0	0	0
D indeks	11,1901	**5,4492**	4,5801	3,8958	3,2918	2,9834	2,7936

Çizelge 4.5 incelendiğinde hiyerarşik olmayan kümeleme yöntemlerinden *k*-ortalama yöntemine göre, gerçekte 3 kümeli farklı yoğunluklu veri seti için Krzanowski-Lai (KL), Calinski-Harabasz (CH), Davies-Bouldin (DB), Silhouette, Ptbiserial, SD indeks, Duda, PseudoT^2, Gap, D indeks, Ball ve Hubert indeksleri diğerlerine göre daha başarılı olmakla birlikte uygun küme sayısını 3 olarak doğru tahmin etmişlerdir. Gamma, Ratkowsky, CCC, Gplus ve McClain indekslerine göre uygun küme sayısı 2 olarak belirlenmiştir. Tau, C indeks ve SDbw indeksleri uygun küme sayısını yaklaşık olarak bile tahmin edemeyerek, uygun küme sayısının 8 olduğunu tahmin etmiştir.

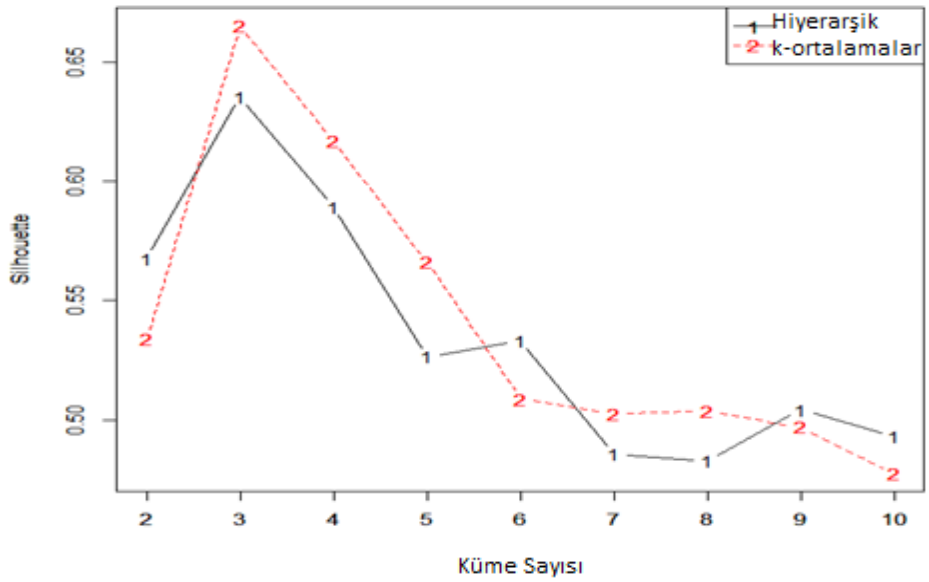
Çizelge 4.6'da tek bağlantı, tam bağlantı, Ward ve k-ortalama kümeleme yöntemlerine göre küme geçerlilik indeksleri kullanılarak belirlenen en uygun küme sayıları yer almaktadır. Çizelgede koyu (bold) işaretlenmiş değerler, farklı yoğunluklu veri setinde etiketli küme sayısını (orijinal küme sayısı) doğru tahmin eden indekslerin küme sayılarını göstermektedir.

Çizelge 4.6. 3 kümeli farklı yoğunluklu veri seti için en uygun küme sayıları

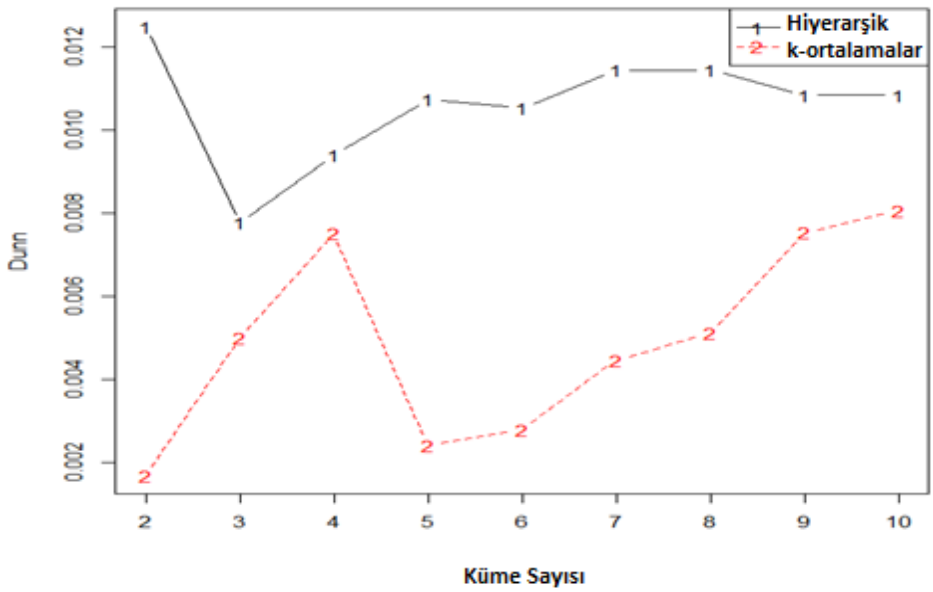
İNDEKS	Tek Bağlantı	Tam Bağlantı	Ward	K-Ortalama
KL	4	**3**	**3**	**3**
CH	**3**	**3**	**3**	**3**
DB	**3**	**3**	**3**	**3**
Silhouette	**3**	**3**	**3**	**3**
Dunn	2	2	2	4
Gamma	2	2	2	2
Tau	8	8	8	8
Ratkowsky	8	2	2	2
Ptbiserial	**3**	**3**	**3**	**3**
CCC	2	2	2	2
C indeks	2	8	8	8
Gplus	8	8	8	2
McClain	2	2	2	2
SDbw	8	8	8	8
SD indeks	**3**	**3**	**3**	**3**
Duda	2	5	8	**3**
PseudoT^2	2	5	8	2
Gap	2	2	**3**	**3**
Ball	**3**	**3**	**3**	**3**
Hubert	**3**	**3**	**3**	**3**
D indeks	**3**	**3**	**3**	**3**

Çizelge 4.6'ya göre, Krzanowski-Lai (KL), Calinski-Harabasz (CH), Davies-Bouldin (DB), Silhouette, Ptbiserial, SD indeks, Ball, D indeks ve Hubert indekslerinin hem hiyerarşik hem de k-ortalama yöntemine göre uygun küme sayısını doğru tahmin ederek diğer indekslere göre daha iyi performans sergilemişlerdir. Öte yandan Dunn, Gamma, Ratkowsky ve PseudoT^2 gibi bazı indeksler uygun küme sayılarını yaklaşık olarak bile tahmin edememişlerdir.

Şekil 4.2 ve Şekil 4.3'te farklı yoğunluklu veri setinin hiyerarşik ve k-ortalama yöntemine göre Silhouette ve Dunn indeks grafikleri görülmektedir.



Şekil 4.2. Farklı yoğunluklu veri seti için uygun küme sayılarının Silhouette indeksine göre gösterimi



Şekil 4.3. Farklı yoğunluklu veri seti için uygun küme sayılarının Dunn indeksine göre gösterimi

Çizelge 4.7. 4 kümeli iyi ayrılmış veri setinin tek bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri

Tek bağlantı kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	0,6230	0,3699	**8,9448**	3,9391	0,4954	1,1641	6,1256
CH	374,2187	231,7947	**825,8768**	643,2381	656,2895	585,8560	536,1128
DB	0,7404	0,5764	0,4525	**0,4368**	0,5782	0,5908	0,7172
Silhouette	0,5526	0,5124	**0,6782**	0,6105	0,5553	0,5044	0,4238
Dunn	0,0860	0,0860	**0,1389**	0,1389	0,0877	0,1020	0,1020
Gamma	0,7977	0,8114	**0,9846**	0,9837	0,9810	0,9800	0,9755
Tau	**56,7891**	48,3309	37,5680	32,0216	27,9003	24,4215	19,7340
Ratkowsky	**0,5193**	0,4501	0,4750	0,4257	0,3927	0,3645	0,3418
Ptbiserial	0,6878	0,7011	**0,8228**	0,8188	0,7944	0,7870	0,7690
CCC	36,2461	23,3910	**50,7004**	45,2755	44,8764	42,4331	40,5457
C indeks	0,3494	0,3398	0,2740	0,2740	**0,2704**	0,2609	0,2868
Gplus	10,3356	8,0377	6,1575	3,7086	2,6508	1,9873	**1,3362**
McClain	**0,3749**	0,3876	0,5099	0,5162	0,5523	0,5631	0,5917
SDbw	0,5627	0,3247	**0,0829**	0,1027	0,1291	0,1315	0,1362
SD indeks	2,3340	1,4361	**1,1119**	1,7982	1,9498	2,1996	2,7804
Duda	0,2924	0,2199	**0,8296**	0,6118	0,8174	0,8381	0,6323
PseudoT^2	222,6845	620,7717	**17,6622**	53,2961	15,4114	17,1916	48,2592
Gap	**-0,8337**	-1,2666	-0,6142	-1,0876	-1,2911	-1,4676	-1,6610
Ball	147,4426	86,1005	**17,1485**	13,2133	8,7813	7,0368	5,7724
Hubert	0,0018	0,0019	**0,0022**	0,0022	0,0022	0,0023	0,0023
D indeks	0,9137	0,8524	**0,4338**	0,4257	0,3877	0,3769	0,3661

Çizelge 4.7 incelendiğinde tek bağlantı yöntemine göre, gerçekte 4 kümeli iyi ayrılmış veri seti için Krzanowski-Lai (KL), Calinski-Harabazs (CH), Silhouette, Dunn, Gamma, Ptbiserial, CCC, SDbw, SD indeks, Duda, PseudoT^2, Ball, D indeks ve Hubert indeksleri diğerlerine göre daha başarılı olmakla birlikte uygun küme sayısını 4 olarak doğru tahmin etmişlerdir. Tau, Ratkowsky, McClain ve Gap indekslerine göre uygun küme sayısı 2 iken, Davies-Bouldin (DB) indeksine göre uygun küme sayısı 5, C indekse göre uygun küme sayısı 6 olarak belirlenmiştir. Bunun yanında Gplus indeksi oldukça düşük bir performans göstererek, uygun küme yaklaşık olarak bile tahmin edememiştir.

Çizelge 4.8. 4 kümeli iyi ayrılmış veri setinin tam bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri

Tam bağlantı kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	0,2568	**10,2322**	1,8132	1,1255	0,9706	0,7364	3,4950
CH	342,7251	846,2862	**888,4583**	787,1358	718,8650	663,8043	606,6076
DB	0,7733	0,5752	**0,4363**	0,7463	0,8871	0,8956	0,8136
Silhouette	0,5359	0,6660	**0,6896**	0,5864	0,4595	0,3803	0,3660
Dunn	0,1261	0,1300	**0,2036**	0,1020	0,0570	0,0603	0,0603
Gamma	0,7638	0,8081	0,9611	**0,9803**	0,9678	0,9743	0,9758
Tau	**53,4026**	45,3897	42,9820	35,1209	28,6514	21,4557	17,8930
Ratkowsky	0,5013	**0,5362**	0,4767	0,4294	0,3940	0,3660	0,3431
Ptbiserial	0,6587	0,8102	**0,8320**	0,7894	0,7250	0,7054	0,7035
CCC	35,3444	337,8251	**51,9834**	48,7285	46,4212	44,5303	42,6036
C indeks	0,3767	**0,2060**	0,2845	0,3075	0,3012	0,3031	0,3009
Gplus	11,5871	7,3678	6,5159	3,6175	2,0221	1,7968	**1,4536**
McClain	**0,3978**	0,5244	0,4961	0,5618	0,6842	0,7221	0,7254
SDbw	0,5506	0,1864	**0,0766**	0,1406	0,1429	0,1439	0,1637
SD indeks	2,3822	1,4539	**1,0832**	2,6018	3,1212	3,0501	3,1787
Duda	0,1737	0,3748	**0,6549**	0,6575	0,7458	0,4486	0,6954
PseudoT^2	818,4350	158,4493	**42,1501**	46,3689	30,6692	23,3537	25,8379
Gap	-0,8501	**-0,4305**	-0,6035	-0,7497	-1,0112	-1,3566	-1,4978
Ball	155,0334	**32,1285**	16,0506	10,9860	8,0694	6,2618	5,1407
Hubert	0,0018	**0,0021**	0,0022	0,0022	0,0023	0,0023	0,0023
D indeks	0,9415	**0,4787**	0,4261	0,3945	0,3681	0,3527	0,3426

Çizelge 4.8 incelendiğinde hiyerarşik kümeleme tekniklerinden tam bağlantı yöntemine göre, gerçekte 4 kümeli iyi ayrılmış veri seti için Calinski-Harabasz (CH), Davies-Bouldin (DB), Silhouette, Dunn, Ptbiserial, CCC, SDbw, SD indeks, Duda ve PseudoT^2 uygun küme sayısını 4 olarak doğru tahmin etmişlerdir. Tau ve McClain indekslerine göre uygun küme sayısı 2 iken, Krzanowski-Lai (KL), Ratkowsky, C indeks, Gap, Ball, D indeks ve Hubert indekslerine göre uygun küme sayısı 3 olarak belirlenmiştir. Gamma indeksine göre ise uygun küme sayısı 5'tir. Ayrıca Gplus indeksi yine oldukça düşük bir performans göstererek, yaklaşık olarak bile tahmin edemeyerek, uygun küme sayısını 8 olarak belirlemiştir.

Çizelge 4.9. 4 kümeli iyi ayrılmış veri setinin Ward yöntemine göre uygun küme sayıları ve indeks değerleri

Ward kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	0,5166	**8,8573**	2,0758	0,7803	1,3558	1,5188	6,7236
CH	400,3894	814,6713	**845,7463**	761,6082	689,9856	644,2103	615,2781
DB	0,6302	0,5972	**0,4483**	0,7544	0,8954	1,1518	1,1383
Silhouette	0,5635	0,6603	**0,6852**	0,5842	0,4625	0,3352	0,3494
Dunn	0,1033	0,1240	**0,1456**	0,0429	0,0314	0,0314	0,0314
Gamma	0,5773	0,6589	0,6817	**0,8873**	0,8816	0,8745	0,8610
Tau	39,7820	**48,4512**	36,2839	28,3567	17,6304	14,6204	12,3021
Ratkowsky	**0,5382**	0,5348	0,4756	0,4289	0,3934	0,3657	0,3432
Ptbiserial	0,6887	0,8058	**0,8276**	0,7737	0,6954	0,6009	0,5523
CCC	36,9623	37,3451	**51,1170**	48,1606	45,7245	44,0257	42,8411
C indeks	0,3037	0,2078	0,2157	0,1982	0,1896	0,1748	**0,1660**
Gplus	13,1654	9,7568	6,9900	4,0945	3,0322	1,7968	**1,7884**
McClain	**0,3684**	0,5307	0,5026	0,5898	0,7559	1,0387	1,2357
SDbw	0,4299	0,1899	**0,0795**	0,1700	0,0974	0,1308	0,1497
SD indeks	2,1279	1,5652	**1,1450**	2,6663	3,5883	4,2464	4,3117
Duda	0,3074	0,3592	**0,6598**	0,6469	0,7107	0,5676	0,4909
PseudoT^2	398,7781	167,7092	**41,7555**	48,0259	36,6377	46,4635	36,2966
Gap	-0,7938	**-0,3688**	-0,7039	-0,8563	-1,1286	-1,4892	-1,6013
Ball	141,6781	**33,1992**	16,7840	16,7840	11,3246	8,3831	5,0724
Hubert	0,0018	**0,0021**	0,0022	0,0022	0,0023	0,0024	0,0024
D indeks	0,8871	**0,4825**	0,4299	0,3921	0,3617	0,3424	0,3221

Çizelge 4.9 incelendiğinde tam bağlantı kümeleme yöntemine benzer şekilde, Ward yöntemine göre de, gerçekte 4 kümeli iyi ayrılmış veri seti için Calinski-Harabazs (CH), Davies-Bouldin (DB), Silhouette, Dunn, Ptbiserial, CCC, SDbw, SD indeks, Duda ve PseudoT^2 indeksleri diğerlerine göre daha başarılı olmakla birlikte uygun küme sayısını 4 olarak doğru tahmin etmişlerdir. Ratkowsky ve McClain indekslerine göre uygun küme sayısı 2 iken, Krzanowski-Lai (KL), Tau, Gap, Ball, D indeks ve Hubert indekslerine göre uygun küme sayısı 3 olarak belirlenmiştir. Gamma indeksine göre ise uygun küme sayısı 5'tir. Ayrıca tek bağlantı ve tam bağlantı yöntemlerinden farklı olarak C indeks ve Gplus indeksleri oldukça düşük bir performans göstererek, uygun küme sayısını 8 olarak belirlemiştir.

Çizelge 4.10. 4 kümeli iyi ayrılmış veri setinin *k*-ortalama yöntemine göre uygun küme sayıları ve indeks değerleri

***k*-Ortalama kümeleme yöntemi**							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	0,5834	**8,2373**	3,8250	0,5465	2,9747	0,2635	2,2108
CH	418,5999	847,7205	**888,4583**	817,6990	746,5827	744,8822	681,3959
DB	0,7456	0,5673	**0,4363**	0,7172	0,9399	1,0977	1,2302
Silhouette	0,5609	0,6660	**0,6896**	0,5973	0,4823	0,3698	0,3381
Dunn	0,0480	0,0538	**0,2036**	0,0550	0,0324	0,0297	0,0297
Gamma	0,5335	0,59809	0,7817	**0,9015**	0,8933	0,8812	0,8717
Tau	41,389	**50,4560**	40,0801	32,0214	25,4215	21,1503	17,5608
Ratkowsky	**0,5405**	0,5363	0,4767	0,4301	0,3945	0,3673	0,3442
Ptbiserial	0,7133	0,8104	**0,8320**	0,7733	0,6894	0,6057	0,5715
CCC	37,4444	37,8466	**51,9834**	49,3865	47,0659	46,4795	44,5540
C indeks	0,3060	0,2523	0,2845	0,2600	0,2463	0,2636	**0,2023**
Gplus	12,1654	9,5204	7,3341	5,8650	2,5211	1,9842	**1,4309**
McClain	0,3930	0,5239	**0,4961**	0,5893	0,7673	1,0022	1,1244
SDbw	0,5436	0,1851	**0,0766**	0,1215	0,1684	0,1607	0,1553
SD indeks	3,2315	1,8145	**1,2734**	2,7148	3,6705	3,9359	5,1970
Duda	**1,0469**	1,9562	1,9967	0,3806	2,0678	1,4589	1,0275
PseudoT^2	**-8,4269**	84,0758	-89,8502	105,7855	-46,4757	-29,5698	-2,0354
Gap	-0,7670	**-0,5185**	-0,6734	-0,9404	-1,1529	-1,4082	-1,5030
Ball	137,9259	**32,0815**	16,0506	10,6062	7,7897	5,6183	4,6059
Hubert	0,0018	**0,0021**	0,0022	0,0022	0,0023	0,0024	0,0024
D indeks	0,8875	**0,4782**	0,4261	0,3861	0,3626	0,3302	0,3159

Çizelge 4.10 incelendiğinde hiyerarşik olmayan kümeleme yöntemlerinden *k*-ortalama kümeleme yöntemine göre, gerçekte 4 kümeli iyi ayrılmış veri seti için Calinski-Harabasz (CH), Davies-Bouldin (DB), Silhouette, Dunn, Ptbiserial, CCC, SDbw, SD indeks ve McClain indeksleri diğerlerine göre daha başarılı olmakla birlikte uygun küme sayısını 4 olarak doğru tahmin etmişlerdir. Ratkowsky, Duda ve PseudoT^2 indekslerine göre uygun küme sayısı 2 iken, Krzanowski-Lai (KL), Tau, Gap, D indeks, Ball ve Hubert ve indekslerine göre uygun küme sayısı 3 olarak belirlenmiştir. Gamma indeksine göre ise uygun küme sayısı 5'tir. C indeks ve Gplus indeksleri diğer indekslere göre oldukça düşük bir performans göstererek, uygun küme sayısını 8 olarak belirlemiştir.

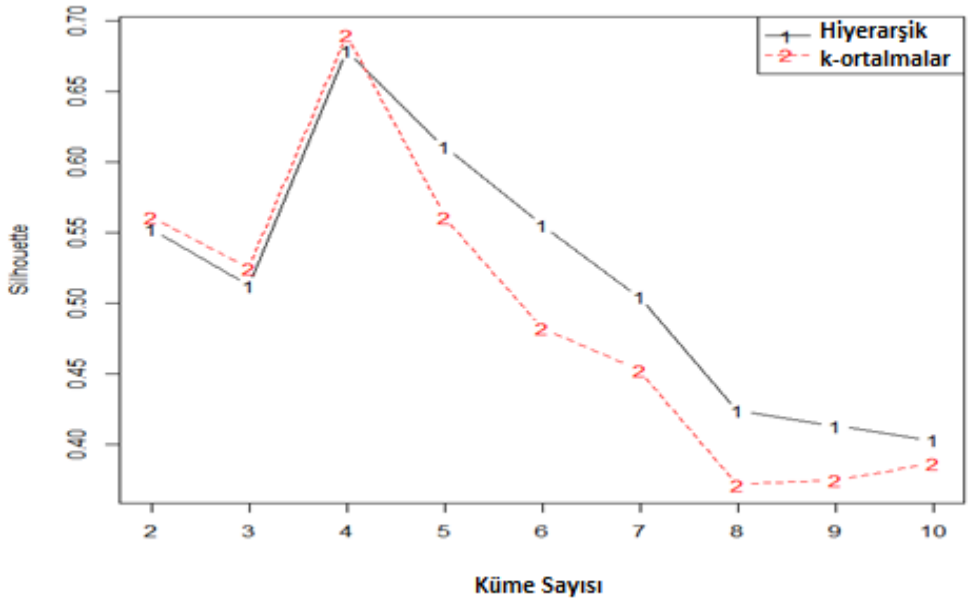
Çizelge 4.11'de tek bağlantı, tam bağlantı, Ward ve k-ortalama kümeleme yöntemlerine göre küme geçerlilik indeksleri kullanılarak belirlenen en uygun küme sayıları yer almaktadır. Çizelgede koyu (bold) işaretlenmiş değerler, iyi ayrılmış veri setinde etiketli küme sayısını (orijinal küme sayısı) doğru tahmin eden indekslerin küme sayılarını göstermektedir

Çizelge 4.11. 4 kümeli iyi ayrılmış veri seti için en uygun küme sayıları

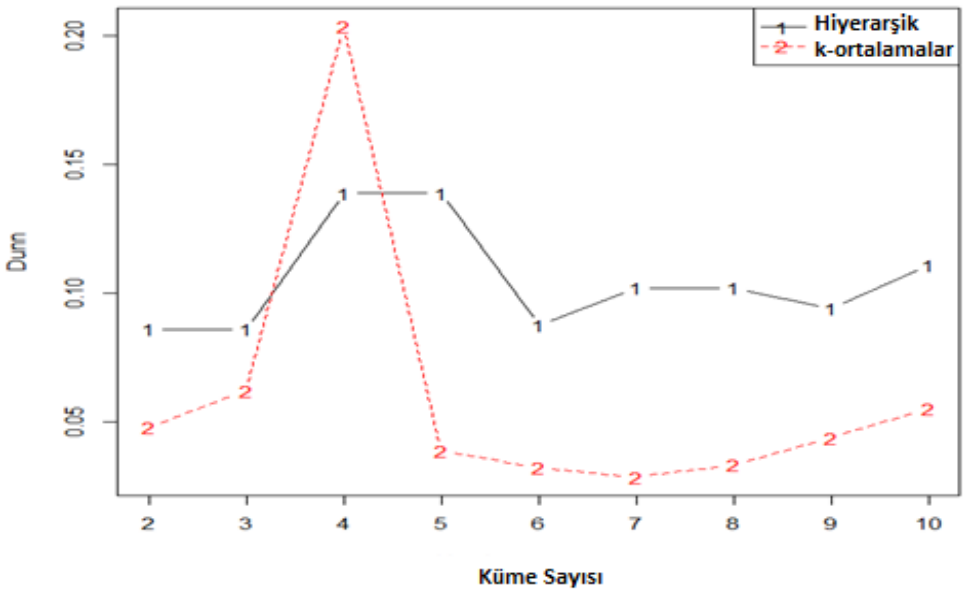
İNDEKS	Tek Bağlantı	Tam Bağlantı	Ward	K-Ortalama
KL	**4**	3	3	3
CH	**4**	**4**	**4**	**4**
DB	5	**4**	**4**	**4**
Silhouette	**4**	**4**	**4**	**4**
Dunn	**4**	**4**	**4**	**4**
Gamma	**4**	5	5	5
Tau	2	2	3	3
Ratkowsky	2	3	2	2
Ptbiserial	**4**	**4**	**4**	**4**
CCC	**4**	**4**	**4**	**4**
C indeks	6	3	8	8
Gplus	8	8	8	8
McClain	2	2	2	**4**
SDbw	**4**	**4**	**4**	**4**
SD indeks	**4**	**4**	**4**	**4**
Duda	**4**	**4**	**4**	2
PseudoT^2	**4**	**4**	**4**	2
Gap	2	3	3	**4**
Ball	**4**	3	3	**4**
Hubert	**4**	3	3	**4**
D indeks	**4**	3	3	**4**

Çizelge 4.11 incelendiğinde küme geçerlilik indekslerinin iyi ayrılmış veri setinin uygun küme sayısını belirlemede genel olarak daha başarılı olduklarını söyleyebiliriz. Calinski-Harabasz (CH), Davies-Bouldin (DB), Dunn, Silhouette, Ptbiserial, SD indeks gibi çoğu indeksin hem hiyerarşik hem de k-ortalama yöntemine göre uygun küme sayısını doğru tahmin ederek diğer indekslere göre daha iyi performans sergilemişlerdir. Öte yandan Gamma, Tau, Ratkowsky, C indeks, McClain ve PseudoT^2 indekslerinin uygun küme küme sayısını tahmin etmede başarısız olmuşlardır.

Şekil 4.4 ve Şekil 4.5'te iyi ayrılmış veri setinin hiyerarşik ve k-ortalama yöntemine göre Silhouette ve Dunn indeks grafikleri görülmektedir.



Şekil 4.4. İyi ayrılmış veri seti için uygun küme sayılarının Silhouette indeksine göre gösterimi



Şekil 4.5. İyi ayrılmış veri seti için uygun küme sayılarının Dunn indeksine göre gösterimi

Çizelge 4.12. Alt kümeli veri setinin tek bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri

Tek bağlantı kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	**1,5943**	1,1992	0,9993	1,0243	1,0056	0,9908	1,0523
CH	1902,8947	**2538,6238**	1693,7284	1270,1569	1020,5130	853,6092	732,5455
DB	0,5584	0,4699	**0,4410**	0,7264	0,6962	0,6776	0,7226
Silhouette	0,6394	**0,6604**	0,6024	0,4300	0,2596	0,1913	0,1228
Dunn	0,2636	**0,2695**	0,1322	0,1227	0,1312	0,1176	0,1026
Gamma	**1,4707**	-0,1917	-0,0582	0,0061	0,0054	0,0025	0,0096
Tau	5,6366	11,8144	11,8319	11,8408	11,8936	0,9409	**11,9630**
Ratkowsky	0,3697	**0,5098**	0,4416	0,3950	0,3608	0,3343	0,3128
Ptbiserial	0,7920	**0,8439**	0,8435	0,8427	0,8418	0,8409	0,8392
CCC	31,1386	**49,8896**	33,3962	22,9167	15,4102	9,5025	4,5623
C indeks	0,2990	**0,2697**	0,2697	0,2697	0,2949	0,2943	0,2941
Gplus	**0,8053**	231,3178	132,0795	2,2076	2,1475	7,4006	1,0635
McClain	**0,2234**	0,3264	00,3269	0,3277	0,3292	0,3306	0,3325
SDbw	0,3706	**0,1259**	0,1523	0,1413	0,2525	0,2241	0,1629
SD indeks	0,3243	**0,2899**	0,6705	0,8470	0,8115	0,8163	0,8591
Duda	0,5916	**0,4610**	0,5256	0,5830	0,5829	0,5828	0,5827
PseudoT^2	1101,5576	**-23,7696**	-14,5027	3,7979	3,3832	1,5422	5,9517
Gap	**0,9962**	0,8937	0,1257	0,0645	-0,0860	-0,2815	-1,1541
Ball	9748,9940	**3100,7921**	2322,1385	1856,3213	1540,0724	1314,8325	1148,3502
Hubert	0	0	**0**	0	0	0	0
D indeks	3,9012	2,7157	**2,7121**	2,7094	2,7040	2,6981	2,6943

Çizelge 4.12 incelendiğinde tek bağlantı yöntemine göre, alt kümeli veri seti için Krzanowski-Lai (KL), Gamma, Gplus, McClain ve Gap indexlerine göre uygun küme sayısını 2 iken, Calinski-Harabasz (CH), Silhouette, Dunn, Ratkowsky, CCC, C indeks, SDbw, SD indeks, Duda, PseudoT^2 ve Ball indekslerine göre uygun küme sayısı 3 olarak belirlenmiştir. Bunun yanı sıra Davies-Bouldin (DB), Hubert ve D indekse göre uygun küme sayısı 4'tür. Tau indeksine göre uygun küme sayısı 8'dir.

Çizelge 4.13. Alt kümeli veri setinin tam bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri

Tam bağlantı kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	2,9895	0,5894	2,7036	0,6946	**4,9697**	0,8640	0,8403
CH	1902,895	2063,514	2933,363	3396,322	4631,404	**4916,144**	4394,360
DB	0,5584	0,8548	**0,5198**	0,5303	0,5811	0,6953	0,8814
Silhouette	**0,6394**	0,5829	0,6045	0,6125	0,5981	0,5445	0,4975
Dunn	**0,2636**	0,0440	0,0562	0,0621	0,0341	0,0245	0,0248
Gamma	**0,9781**	10,8953	3,8949	1,1208	0,8201	0,3506	0,1568
Tau	1,2648	0,2710	1,0897	1,2777	2,1086	3,5595	**3,5949**
Ratkowsky	0,3697	**0,4957**	0,4641	0,4304	0,3973	0,3695	0,3459
Ptbiserial	**0,7920**	0,7469	0,7646	0,7442	0,6468	0,5373	0,5219
CCC	31,1386	42,2278	52,8596	56,6574	66,5957	**67,7667**	63,1766
C indeks	0,2990	0,2450	0,2661	0,2674	**0,2250**	0,2703	0,2704
Gplus	**5,6366**	9,9662	19,0724	28,4155	47,1153	59,5414	62,0697
McClain	**0,2234**	0,4764	0,4695	0,5044	0,6613	0,9389	0,9955
SDbw	0,3706	0,3147	0,0944	**0,0529**	0,0734	0,1382	0,1597
SD indeks	0,5299	0,5574	0,5004	**0,4396**	0,5884	1,0083	1,5849
Duda	0,5916	0,5253	0,5256	0,5714	0,5537	**0,4644**	0,4610
PseudoT^2	516,4107	223,2061	223,8285	374,2351	298,1809	**146,4622**	143,8309
Gap	**0,2857**	-0,3142	-0,0975	-0,0155	-0,0094	-0,1230	-0,3621
Ball	9748,9940	**3675,8153**	1440,5818	773,5319	388,7683	263,6859	221,3272
Hubert	0	**0**	0	0	0	0	0
D indeks	3,9012	**2,9699**	2,1105	1,6853	1,3467	1,2024	1,1801

Çizelge 4.13 incelendiğinde tam bağlantı yöntemine göre, alt kümeli veri seti için Silhouette, Dunn, Gamma, Gplus, Ptbiserial, McClain ve Gap indexlerine göre uygun küme sayısını 2 iken, Ratkowsky, Ball, Hubert ve D indekse göre uygun küme sayısı 3 olarak belirlenmiştir. Davies-Bouldin (DB), göre uygun küme sayısı 4; SDbw ve SD indekslerine göre 5'tir. Ayrıca Krzanowski-Lai (KL) ve C indekslerine göre uygun küme sayısı 6 iken, Calinski-Harabasz (CH), CCC, Duda ve PseudoT^2 uygun küme sayısı 7 olarak belirlenmiştir. Tau indeksine göre uygun küme sayısı 8'dir.

Çizelge 4.14. Alt kümeli veri setinin Ward yöntemine göre uygun küme sayıları ve indeks değerleri

Ward kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	1,5943	2,0442	1,4080	0,8411	**8,2099**	5,0945	0,1466
CH	1902,895	2538,624	3015,657	3534,899	4612,908	**4701,106**	4582,425
DB	0,5584	**0,4699**	0,6900	0,6429	0,5589	0,7428	0,7675
Silhouette	0,6394	**0,6604**	0,5868	0,5913	0,6015	0,5288	0,5248
Dunn	0,2636	**0,2695**	0,0324	0,0328	0,0454	0,0207	0,0207
Gamma	**0,2301**	0,2947	0,0210	0,0465	0,5811	0,4288	0,5205
Tau	0,8053	1,7530	0,6644	0,6808	2,4536	1,8448	**2,6601**
Ratkowsky	0,3697	**0,5098**	0,4553	0,4233	0,3969	0,3697	0,3463
Ptbiserial	0,7920	**0,8439**	0,6941	0,6783	0,6591	0,5396	0,4967
CCC	31,1386	49,8896	53,8742	58,0956	**66,4530**	66,1988	64,6228
C indeks	0,2990	0,2697	0,2430	**0,2131**	0,2533	0,2401	0,2240
Gplus	**5,6366**	11,8144	19,5531	29,4958	46,9349	57,0218	64,6431
McClain	**0,2234**	0,3264	0,5989	0,6215	0,6365	0,9445	1,0963
SDbw	0,3706	0,1259	0,1469	0,0895	**0,0792**	0,1137	0,1160
SD indeks	0,6642	**0,4449**	0,6375	0,5147	0,5800	1,1421	1,3004
Duda	0,4044	0,4495	0,2036	0,2438	**0,6474**	0,5399	0,6029
PseudoT^2	516,4107	445,6485	223,8285	215,7172	**307,7354**	210,0761	166,7048
Gap	-0,1026	**-0,0086**	-0,2296	-0,1892	-0,0840	-0,2494	-0,379
Ball	9748,9940	**3100,7921**	1405,1683	745,2012	390,2627	275,3373	212,5163
Hubert	0	**0**	0	0	0	0	0
D indeks	3,9012	**2,7157**	2,1101	1,6848	1,3505	1,2237	1,1487

Çizelge 4.14 incelendiğinde Ward yöntemine göre, alt kümeli veri seti için Silhouette, Dunn, Gamma, Tau, Gplus, Ptbiserial, McClain ve Gap indexlerine göre uygun küme sayısını 2 iken, Ratkowsky, Ball, Hubert ve D indekse göre uygun küme sayısı 3 olarak belirlenmiştir. Davies-Bouldin (DB) indeksine göre uygun küme sayısı 4, SDbw ve SD indekslerine göre 5'tir. Ayrıca Krzanowski-Lai (KL) ve C indekse göre uygun küme sayısı 6 iken, Calinski-Harabasz (CH), CCC, Duda ve PseudoT^2 uygun küme sayısı 7 olarak belirlenmiştir.

Çizelge 4.15. Alt kümeli veri setinin *k*-ortalama yöntemine göre uygun küme sayıları ve indeks değerleri

***k*-Ortalama kümeleme yöntemi**							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	1,5937	1,9618	3,4392	0,2269	**4,5060**	1,1306	1,6828
CH	1903,975	2541,673	3055,240	3177,317	**4875,903**	4231,923	4615,213
DB	0,5607	**0,4705**	0,6790	0,5779	0,5767	0,8313	0,8985
Silhouette	0,6389	**0,6605**	0,5889	0,5974	0,6055	0,5554	0,5043
Dunn	0,0646	**0,0893**	0,0122	0,0087	0,0125	0,0125	0,0109
Gamma	**3,7284**	0,8527	-0,0367	0,8949	-0,2954	0,7035	-0,1969
Tau	0,7999	1,7444	1,3423	0,4369	**9,0708**	1,9734	2,0740
Ratkowsky	0,3692	**0,5099**	0,4562	0,4201	0,3973	0,3684	0,3461
Ptbiserial	0,7917	**0,8438**	0,6952	0,6506	0,6389	0,6243	0,5226
CCC	31,1535	49,9348	54,3534	54,2691	**68,4317**	62,5225	64,8690
C indeks	0,2987	0,2909	0,2425	**0,2152**	0,2488	0,2476	0,2638
Gplus	**5,6387**	11,8262	19,7843	26,7083	49,5002	51,5244	65,0918
McClain	**0,2243**	0,3270	0,5965	0,6896	0,6716	0,7073	0,9797
SDbw	0,3773	0,1258	0,1430	0,1174	**0,0745**	0,1138	0,1327
SD indeks	0,5159	**0,3766**	0,5772	0,6886	0,5848	1,2858	1,2888
Duda	**3,7220**	0,5394	1,0382	0,2036	1,4241	0,5864	1,2467
PseudoT^2	**-638,4503**	544,0181	-18,2689	969,8281	-55,0929	225,0404	-50,8553
Gap	**-0,1041**	-0,1915	-0,2653	-0,4373	-0,1400	-0,4329	-0,4277
Ball	9745,3653	**3097,6815**	1388,7475	822,9785	370,0375	304,7144	211,0514
Hubert	0,0013	**0,0013**	0,0014	0,0014	0,0014	0,0015	0,0015
D indeks	0,3106	**0,1637**	0,1282	0,1132	0,0935	0,0863	0,0703

Çizelge 4.15 incelendiğinde *k*-ortalama yöntemine göre, alt kümeli veri seti için Gamma, Gplus, McClain, Duda, PseudoT^2 ve Gap indexlerine göre uygun küme sayısını 2 iken, Davies-Bouldin (DB), Silhouette, Dunn, Ratkowsky, Ptbiserial, SD indeks, Ball, Hubert ve D indekse göre uygun küme sayısı 3 olarak belirlenmiştir. C indekse göre uygun küme sayısı 5, Krzanowski-Lai (KL), Calinski-Harabasz (CH), Tau, CCC ve SDbw indekslerine göre uygun küme sayısı 6 olarak belirlenmiştir.

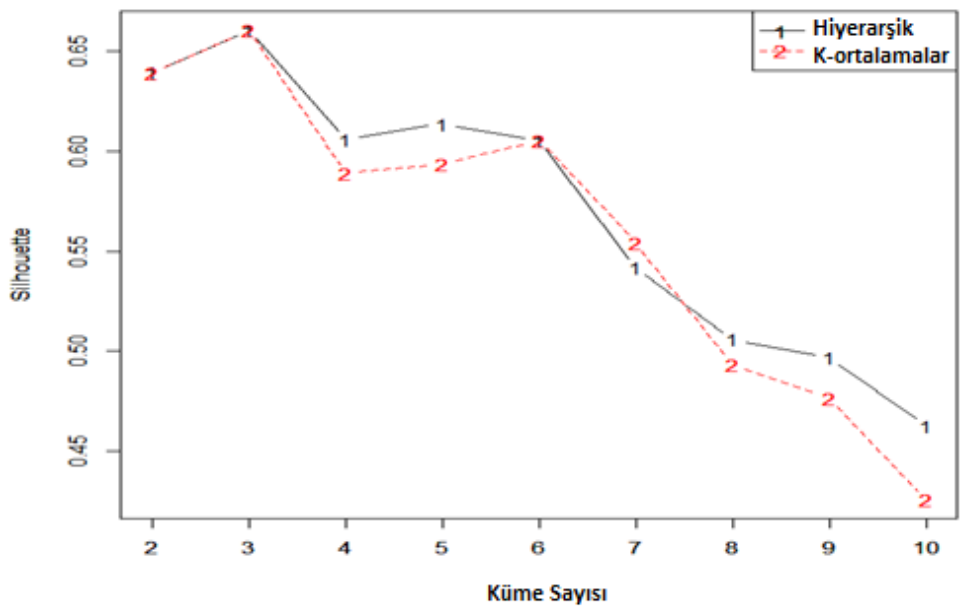
Çizelge 4.16'da tek bağlantı, tam bağlantı, Ward ve k-ortalama kümeleme yöntemlerine göre küme geçerlilik indeksleri kullanılarak belirlenen en uygun küme sayıları yer almaktadır. Çizelgede koyu (bold) işaretlenmiş değerler, alt kümeli veri setinde etiketli küme sayısını (orijinal küme sayısı) doğru tahmin eden indekslerin küme sayılarını, koyu ve altı çizili değerler ise alt küme sayısını göstermektedir.

Çizelge 4.16. Alt kümeli veri seti için en uygun küme sayıları

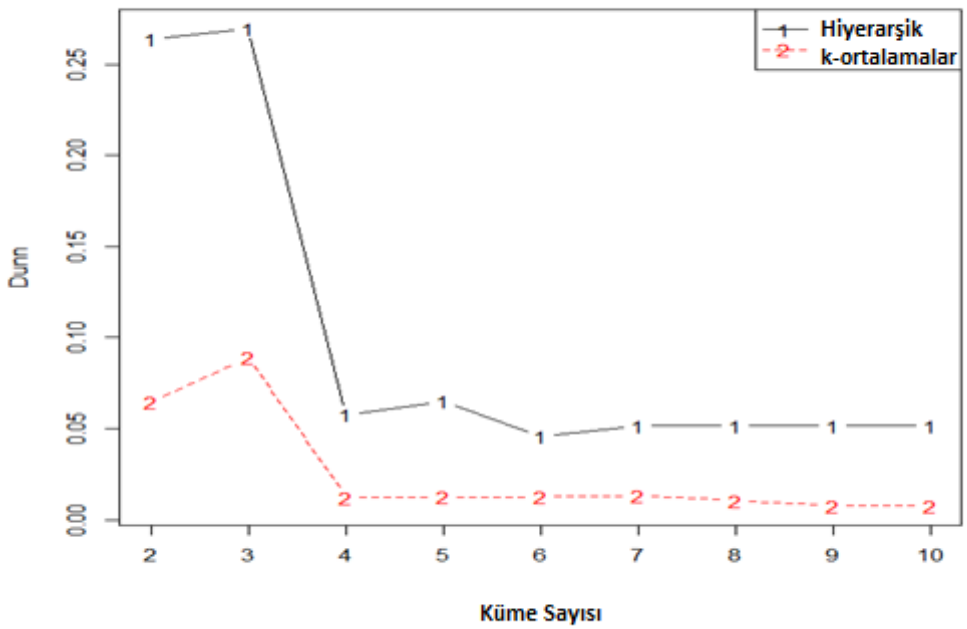
İNDEX	Tek Bağlantı	Tam Bağlantı	Ward	K-Ortalama
KL	2	**6**	**6**	**6**
CH	**3**	7	7	**6**
DB	4	4	**3**	**3**
Silhouette	**3**	2	**3**	**3**
Dunn	**3**	2	**3**	**3**
Gamma	2	2	2	2
Tau	8	8	8	**6**
Ratkowsky	**3**	**3**	**3**	**3**
Ptbiserial	**3**	2	**3**	**3**
CCC	**3**	7	**6**	**6**
C indeks	**3**	**6**	5	5
Gplus	2	2	2	2
McClain	2	2	2	2
SDbw	**3**	5	**6**	**6**
SD indeks	**3**	5	**3**	**3**
Duda	**3**	7	**6**	2
PseudoT^2	**3**	7	**6**	2
Gap	2	2	**3**	2
Ball	**3**	**3**	**3**	**3**
Hubert	4	**3**	**3**	**3**
D indeks	4	**3**	**3**	**3**

Çizelge 4.16 incelendiğinde, küme geçerlilik indeksleri gerçekte 6 kümeli olarak üretilen alt kümeli veri setinin uygun küme sayısını, sub-optimal küme sayısını tahmin ettiği görülmektedir. Bu durum göz önüne alındığında alt kümeli veri seti için Davies-Bouldin (DB), Dunn, Silhouette, Ratkowsky, Ptbiserial, Ball, Hubert ve D indeksin uygun küme sayını belirlemede diğer indekslere göre daha başarılı olduklarını söyleyebiliriz.

Şekil 4.6 ve Şekil 4.7'de alt kümeli veri setinin hiyerarşik ve k-ortalama yöntemine göre Silhouette ve Dunn indeks grafikleri görülmektedir.



Şekil 4.6. Alt kümeli veri seti için uygun küme sayılarının Silhouette indeksine göre gösterimi



Şekil 4.7. Alt kümeli veri seti için uygun küme sayılarının Dunn indeksine göre gösterimi

Çizelge 4.17. Cassini veri setinin tek bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri

Tek bağlantı kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	**1,5366**	0,2767	1,0032	1,0020	0,9985	1,0001	1,0011
CH	**3277,1370**	2682,1974	1788,0292	1342,0221	1074,5247	895,5950	767,6948
DB	**0,8457**	1,0048	1,1117	1,0451	1,1226	1,1202	1,1602
Silhouette	**0,4822**	0,3529	0,1353	0,0830	0,0073	0,0475	0,2051
Dunn	**0,0818**	0,0757	0,0370	0,0332	0,0325	0,0311	0,0304
Gamma	**0,6944**	0,0026	0,0016	0,0018	0,0008	0,0024	0,0007
Tau	1,2448	2,7073	2,1273	**55,8754**	7,7070	45,2038	10,7897
Ratkowsky	**0,3111**	0,2802	0,2428	0,2212	0,2020	0,1876	0,1756
Ptbiserial	**0,6680**	0,6200	0,6197	0,6194	0,6182	0,6178	0,6174
CCC	**1,9777**	6,3642	25,0160	37,4673	46,9314	54,6502	61,1899
C indeks	0,4080	0,3491	0,3491	0,3492	0,3491	0,3491	**0,3470**
Gplus	**2,3121**	3,1486	3,1493	3,1518	3,1545	3,1557	3,1567
McClain	**0,4525**	0,8181	0,8189	0,8201	0,8241	0,8254	0,8268
SDbw	1,6480	0,6535	0,6292	0,5968	0,5021	0,4697	**0,3812**
SD indeks	**2,9713**	3,3193	5,5946	5,1276	5,4538	5,5197	5,3291
Duda	0,5900	**0,9974**	0,9984	0,9982	0,9992	1,0024	0,9993
PseudoT^2	1040,9406	**1,2788**	1,6035	1,7722	0,7924	-2,4191	0,7206
Gap	-0,5218	**-0,5093**	-0,5162	-0,5216	-0,5208	-1,4451	-1,9199
Ball	982,1044	**480,7897**	360,5078	288,1807	239,9465	205,5873	179,8329
Hubert	0	**0**	0	0	0	0	0
D indeks	0,8266	**0,6827**	0,6825	0,6820	0,6813	0,6810	0,6808

Çizelge 4.17 incelendiğinde tek bağlantı yöntemine göre, Cassini veri seti için Krzanowski-Lai (KL), Calinski-Harabasz (CH), Davies-Bouldin (DB), Silhouette, Dunn, Gamma, Ratkowsky, Ptbiserial, CCC, McClain, SD indeks ve Gplus indekslerine göre uygun küme sayısı 2'dir. Bunun yanında Duda, PseudoT^2, Gap, Ball, Hubert ve D indeks diğerlerine göre daha başarılı bir performans göstererek uygun küme sayısını 3 olarak doğru tahmin etmişlerdir. Tau indeksine göre uygun küme sayısı 5 olarak tahmin edilmiştir. Ancak C indeks ve SDbw indekslerini uygun küme sayısını belirlemedeki performansları önemli derecede düşüktür. Bu indekslere göre uygun küme sayısı 8 olarak belirlenmiştir.

Çizelge 4.18. Cassini veri setinin tam bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri

Tam bağlantı kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	0,8564	5,8336	0,0863	**11,1880**	1,1995	0,5785	2,9268
CH	3277,137	2414,152	2510,656	**3943,670**	3711,767	3546,541	3321,784
DB	0,8457	**0,7588**	0,9648	0,7719	0,8015	0,9327	0,9928
Silhouette	**0,4822**	0,4303	0,3603	0,4670	0,4302	0,3764	0,3611
Dunn	**0,0818**	0,0184	0,0170	0,0221	0,0258	0,0278	0,0193
Gamma	**1,5116**	0,4880	2,0539	0,9204	0,6358	0,6072	0,6755
Tau	**2,6379**	0,8435	0,2328	0,7051	0,9432	0,9929	0,7766
Ratkowsky	0,3111	**0,4239**	0,4202	0,4012	0,3727	0,3504	0,3289
Ptbiserial	**0,6680**	0,6189	0,5821	0,6060	0,5877	0,5546	0,5351
CCC	1,9777	-11,6992	-8,6629	**15,5427**	12,4331	10,1360	6,8082
C indeks	0,4080	0,3912	0,3485	**0,3198**	0,3532	0,3580	0,3543
Gplus	**2,3121**	2,9339	4,0179	7,3231	8,4421	9,5363	10,3317
McClain	**0,4525**	0,6303	1,0714	1,2778	1,3862	1,5896	1,7217
SDbw	1,6480	1,2960	0,8570	0,3996	0,3959	0,3397	**0,3164**
SD indeks	2,3319	2,1239	2,2672	**1,8925**	2,5626	3,1327	3,5318
Duda	0,3979	**0,6719**	0,3272	0,5203	0,6109	0,6216	0,5963
PseudoT^2	1510,0861	**731,5757**	2193,5306	505,3213	361,1434	261,1168	303,3037
Gap	**-0,3113**	-1,0445	-1,1504	-1,1118	-1,3761	-1,4859	-1,6256
Ball	982,1044	**515,9755**	282,5720	124,0301	89,6586	68,0317	54,9454
Hubert	0	0	0	**0**	0	0	0
D indeks	0,8266	0,7145	0,6079	**0,4616**	0,4303	0,4050	0,3879

Çizelge 4.18 incelendiğinde tam bağlantı yöntemine göre, Cassini veri seti için Silhouette, Dunn, Gamma, Tau, Ptbiserial, McClain, Gap ve Gplus indekslerine göre uygun küme sayısı 2 iken'dir. Davies-Bouldin (DB), Ratkowsky, Duda, PseudoT^2 ve Ball, indeksleri diğerlerine göre daha başarılı bir performans göstererek uygun küme sayısını 3 olarak doğru tahmin etmişlerdir. Krzanowski-Lai (KL), Calinski-Harabasz (CH), CCC, C indeks, SD indeks, Hubert ve D indekse göre uygun küme sayısı 5 olarak tahmin edilmiştir. Öte yandan SDbw indeksinin uygun küme sayısını belirlemedeki performansı önemli derecede düşük olup, 8 küme olarak belirlenmiştir.

Çizelge 4.19. Cassini veri setinin Ward kümeleme yöntemine göre uygun küme sayıları ve indeks değerleri

Ward kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	1,5094	1,2982	0,3291	**12,0511**	0,5839	1,8058	0,7790
CH	3308,713	2682,197	2942,820	**4065,094**	3882,359	3658,768	3540,532
DB	0,8386	1,0048	0,8698	**0,7547**	0,8460	0,9610	1,0362
Silhouette	**0,4833**	0,3529	0,4220	0,4765	0,4497	0,4082	0,3831
Dunn	**0,0757**	0,0757	0,0124	0,0188	0,0201	0,0205	0,0236
Gamma	**0,6429**	1,7648	1,4977	0,9263	0,6671	0,7166	0,6531
Tau	**1,2376**	0,5811	0,3002	0,8492	0,9162	0,8152	1,0732
Ratkowsky	0,3138	0,2802	0,3929	**0,4030**	0,3758	0,3504	0,3309
Ptbiserial	**0,7015**	0,6884	0,6221	0,5307	0,5589	0,5215	0,5203
CCC	2,3673	-6,3642	-0,5000	**17,1925**	14,8282	11,7724	10,1216
C indeks	0,4081	0,3491	**0,3027**	0,3454	0,3457	0,3342	0,3632
Gplus	**2,3247**	3,1486	4,5374	7,5178	8,7841	9,8065	10,9462
McClain	**0,4529**	0,8181	1,0624	1,2588	1,4362	1,6088	1,7953
SDbw	1,5108	0,6535	0,5090	0,3728	0,3117	0,3255	**0,2842**
SD indeks	2,1739	2,7673	2,1553	**1,8556**	2,7302	3,2031	3,1910
Duda	0,6044	0,3615	0,4001	0,5187	**0,5994**	0,5821	0,6082
PseudoT^2	980,5562	1762,9975	1496,1803	553,0084	**339,5393**	350,3977	320,7826
Gap	**-0,3738**	-0,8842	-1,0052	-1,1156	-1,2691	-1,4535	-1,5892
Ball	976,7639	**480,7897**	250,2215	120,8182	86,1675	66,1578	51,8608
Hubert	0	0	0	**0**	0	0	0
D indeks	0,8244	0,6827	0,5639	**0,4573**	0,4235	0,3999	0,3777

Çizelge 4.19 incelendiğinde Ward yöntemine göre, Cassini veri seti için Silhouette, Dunn, Gamma, Tau, Ptbiserial, McClain, Gap ve Gplus indekslerine göre uygun küme sayısı 2 olarak belirlenmiştir. Dikkat edilecek olursa sadece Ball indeksi uygun küme sayısını 3 olarak doğru tahmin etmiştir. Krzanowski-Lai (KL), Calinski-Harabasz (CH), Davies-Bouldin (DB), Ratkowsky, CCC, SD indeks, Hubert ve D indekse göre uygun küme sayısı 5 olarak tahmin edilirken; Duda ve PseudoT^2'ye göre uygun küme sayısı 6'dır. Öte yandan SDbw indeksinin uygun küme sayısını belirlemedeki performansı önemli derecede düşük olup, 8 küme olarak belirlenmiştir.

Çizelge 4.20. Cassini veri setinin k-ortalama yöntemine göre uygun küme sayıları ve indeks değerleri

***k*-ortalama kümeleme yöntemi**							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	1,5256	5,6204	0,1709	0,3626	**6,7418**	1,1946	1,1946
CH	3758,883	2776,443	2836,955	2361,843	**3999,214**	3615,056	3787,946
DB	**0,8151**	0,9478	0,9097	0,9745	0,8226	0,9633	0,9984
Silhouette	**0,5027**	0,4289	0,4130	0,3838	0,4589	0,4134	0,3941
Dunn	0,0063	**0,0064**	0,0056	0,0056	0,0133	0,0133	0,0141
Gamma	**0,2112**	0,0246	0,1124	1,7501	0,6545	0,0650	0,4464
Tau	2,0474	0,6688	18,1702	0,1962	**2,3968**	0,5334	0,8797
Ratkowsky	0,3197	**0,4293**	0,3878	0,3572	0,3769	0,3520	0,3349
Ptbiserial	**0,6944**	0,6215	0,5989	0,5607	0,5805	0,5455	0,5139
CCC	7,7027	-4,5745	-2,4097	-11,1898	**16,4156**	11,1404	13,6521
C indeks	0,3912	0,3587	0,3047	**0,3039**	0,3477	0,3402	0,3765
Gplus	**2,5050**	3,2241	4,4101	4,7869	9,0184	9,7013	11,6412
McClain	**0,4703**	0,7703	1,0883	1,2705	1,4326	1,6488	1,8735
SDbw	1,8561	1,3312	0,5153	0,4902	0,3197	0,2908	**0,2592**
SD indeks	**2,3202**	2,5179	2,3342	2,9055	2,4892	4,1039	3,9087
Duda	**1,2680**	0,9759	0,8988	0,3634	0,6039	0,9388	1,8109
PseudoT^2	**-401,3221**	25,4084	76,5851	1748,3794	326,5758	42,5499	-275,8417
Gap	**-0,4432**	-1,0269	-1,2824	-1,6535	-1,3761	-1,6224	-1,7122
Ball	906,4877	**469,5319**	257,4414	189,7437	83,9290	66,8753	48,7645
Hubert	0	**0**	0	0	0	0	0
D indeks	0,7980	**0,6894**	0,5719	0,5388	0,4189	0,3992	0,3660

Çizelge 4.20 incelendiğinde *k*-ortalama yöntemine göre, Cassini veri seti için Davies-Bouldin (DB), Silhouette, Gamma, Ptbiserial, McClain, Gap, Gplus, SD indeks, Duda ve PseudoT^2 indekslerine göre uygun küme sayısı 2 olarak belirlenmiştir. Dikkat edilecek olursa Dunn, Ball, Hubert ve D indeks diğerlerine göre daha başarılı olarak uygun küme sayısını 3 olarak doğru tahmin etmişlerdir. Bunun yanında, C indekse göre uygun küme sayısı 5, Krzanowski-Lai (KL), Calinski-Harabasz (CH), Tau, Ratkowsky ve CCC indekslerine göre uygun küme sayısı 6'dır. SDbw indeksinin uygun küme sayısını belirlemedeki performansı önemli derecede düşük olup, 8 küme olarak belirlenmiştir.

Çizelge 4.21'da tek bağlantı, tam bağlantı, Ward ve k-ortalama kümeleme yöntemlerine göre küme geçerlilik indeksleri kullanılarak belirlenen en uygun küme sayıları yer almaktadır. Çizelgede koyu (bold) işaretlenmiş değerler, Cassini veri setinde etiketli küme sayısını (orijinal küme sayısı) doğru tahmin eden indekslerin küme sayılarını göstermektedir.

Çizelge 4.21. Cassini veri seti için en uygun küme sayıları

İNDEKS	**Tek Bağlantı**	**Tam Bağlantı**	**Ward**	***k*-Ortalama**
KL	2	5	5	6
CH	2	5	5	6
DB	2	**3**	5	2
Silhouette	2	2	2	2
Dunn	2	2	2	**3**
Gamma	2	2	2	2
Tau	5	2	2	2
Ratkowsky	2	**3**	5	6
Ptbiserial	2	2	2	2
CCC	2	5	5	2
C indeks	8	5	4	6
Gplus	2	2	2	5
McClain	2	2	2	2
SDbw	8	8	8	8
SD indeks	2	5	5	2
Duda	**3**	**3**	6	2
PseudoT^2	**3**	**3**	6	2
Gap	**3**	2	2	2
Ball	**3**	**3**	**3**	**3**
Hubert	**3**	5	5	**3**
D indeks	**3**	5	5	**3**

Çizelge 4.21 incelendiğinde, küme geçerlilik indekslerinin genel olarak uygun küme sayılarını belirlemede başarısız olduğu görülmektedir. Küme geçerlilik indekslerinin diğer veri setlerinin uygun küme sayısını belirlemedeki performansları göz önüne alındığında, bu başarısızlığın Cassini veri setinin küme yapısındaki yanlış değerlendirmeden kaynaklandığı düşünülmektedir.

Çizelge 4.22. İris veri setinin tek bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri

Tek bağlantı kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	**16,7156**	0,8118	2,8501	0,6846	0,9318	1,0414	1,0207
CH	277,9947	**502,8216**	187,4765	169,1470	139,0601	116,9933	100,7182
DB	**0,4360**	0,4989	0,6348	0,5812	0,5775	0,5560	0,6497
Silhouette	0,5121	**0,6867**	0,2886	0,2905	0,2348	0,1528	0,0854
Dunn	**0,3389**	0,1691	0,1518	0,1560	0,1523	0,1504	0,1341
Gamma	**0,9587**	0,9583	0,9484	0,9450	0,9408	0,9348	0,9244
Tau	**2770,853**	2796,561	2795,249	2807,087	2800,907	2797,920	2796,076
Ratkowsky	**0,5535**	0,4650	0,4046	0,3728	0,3428	0,3205	0,3003
Ptbiserial	**0,8358**	0,8348	0,8250	0,8166	0,8112	0,8048	0,7938
CCC	**35,7286**	29,2661	24,2234	22,6013	20,0869	17,9022	16,0180
C indeks	0,2718	0,2630	**0,2615**	0,2868	0,2837	0,2834	0,2825
Gplus	**57,0657**	58,0159	71,8255	76,8320	82,5891	90,9681	105,2659
McClain	**0,2622**	0,2760	0,2872	0,3172	0,3268	0,3341	0,3475
SDbw	0,1578	0,1019	0,0762	0,1075	0,0473	0,0516	**0,0347**
SD indeks	**1,0333**	1,2389	1,9314	1,9634	2,9039	2,8079	2,8608
Duda	**0,9108**	1,0401	0,8389	1,0364	1,5340	1,0599	1,1497
PseudoT^2	**9,5959**	-3,7023	18,2461	-3,1948	-16,7093	-5,0878	-11,5870
Gap	**1,0126**	-0,1975	-0,1938	-0,5547	-0,7290	-0,6869	-1,4596
Ball	**77,4735**	47,4931	35,1059	24,0507	19,4840	16,4735	14,2786
Hubert	0,0019	**0,0020**	0,0020	0,0020	0,0020	0,0020	0,0020
D indeks	0,8535	**0,8219**	0,8125	0,7605	0,7478	0,7396	0,7323

Çizelge 4.22 incelendiğinde tek bağlantı yöntemine göre, İris veri seti için Calinski-Harabasz (CH), Silhouette, Hubert ve D indeks diğerlerine göre daha başarılı bir performans göstererek uygun küme sayısını 3 olarak doğru tahmin etmişlerdir. Bunun yanında C indekse göre uygun küme sayısı 4 ve SDbw indeksine göre uygun küme sayısı 8 olarak belirlenmiştir. Diğer indekslere göre uygun küme sayısı 2'dir.

Çizelge 4.23. İris veri setinin tam bağlantı yöntemine göre uygun küme sayıları ve indeks değerleri

Tam bağlantı kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	1,9652	5,3598	**54,0377**	0,0263	7,1653	0,5308	2,4071
CH	280,8392	485,9050	**495,1816**	414,3925	455,4931	423,7198	414,7146
DB	0,7027	**0,7025**	0,7289	0,9838	1,0524	1,0030	1,0738
Silhouette	0,5136	**0,5160**	0,4998	0,3462	0,3382	0,3298	0,3240
Dunn	0,0824	0,1033	0,1365	0,1000	0,1311	0,1346	**0,1529**
Gamma	0,7472	0,8928	**0,9261**	0,8589	0,8919	0,9020	0,9115
Tau	2475,495	**2649,840**	2495,851	2206,153	1728,103	1664,993	1384,061
Ratkowsky	**0,4729**	0,4922	0,4387	0,4026	0,3738	0,3482	0,3275
Ptbiserial	0,6369	**0,7203**	0,6948	0,6073	0,5295	0,5212	0,4753
CCC	30,4441	**35,8668**	35,6036	33,0698	33,9870	32,4873	32,4873
C indeks	0,3723	**0,3163**	0,3465	0,3758	0,4032	0,3982	0,4118
Gplus	353,1090	139,9284	87,9342	149,0951	88,5252	77,1718	**58,7781**
McClain	**0,4228**	0,4964	0,5734	0,7936	1,0742	1,1037	1,3191
SDbw	0,8976	0,2350	0,1503	0,5055	0,3126	0,2284	**0,0357**
SD indeks	**1,8326**	1,6226	1,9103	3,4597	3,5342	3,6106	3,9101
Duda	0,1460	0,5582	**0,5932**	0,5452	0,5656	0,6480	2,1863
PseudoT^2	444,4821	55,4060	**32,9134**	48,3914	19,9691	19,5552	-11,9371
Gap	-0,2356	**0,1343**	-0,1465	-0,3669	-0,3256	-0,5714	-0,6911
Ball	117,5765	**29,8417**	15,2432	10,9620	6,7533	5,1835	3,9719
Hubert	0,0015	**0,0020**	0,0022	0,0022	0,0023	0,0023	0,0023
D indeks	1,1446	**0,6722**	0,5832	0,5513	0,4778	0,4530	0,4239

Çizelge 4.23 incelendiğinde tam bağlantı yöntemine göre, İris veri seti için Silhouette, Tau, Ptbiserial, CCC, C indeks Gap, Ball, Hubert ve D indeks diğerlerine göre daha başarılı bir performans göstererek uygun küme sayısını 3 olarak doğru tahmin etmişlerdir. Krzanowski-Lai (KL), Calinski-Harabasz (CH), Gamma, Duda ve PseudoT^2 indekslerine göre uygun küme sayısı 4 olarak tahmin edilmiştir. Öte yandan Dunn, Gplus ve SDbw indekslerinin uygun küme sayısını belirlemedeki performansı önemli derecede düşük olup, 8 küme olarak belirlenmiştir.

Çizelge 4.24. İris veri setinin Ward yöntemine göre uygun küme sayıları ve indeks değerleri

Ward kümeleme yöntemi							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	**5,6700**	4,0887	1,5933	1,7570	1,6748	0,9004	3,4043
CH	502,8216	**556,8795**	514,9703	488,9985	461,5442	434,3388	418,7252
DB	**0,4360**	0,7218	0,8567	0,8752	0,9985	1,0688	1,0215
Silhouette	0,5513	**0,6867**	0,4899	0,4860	0,3576	0,3479	0,3495
Dunn	**0,3389**	0,1378	0,1235	0,1243	0,1311	0,1311	0,1508
Gamma	**0,9587**	0,9158	0,9092	0,9257	0,8985	0,9024	0,9161
Tau	**2770,853**	2623,023	2243,920	2114,589	1754,865	1478,367	1418,726
Ratkowsky	**0,5535**	0,4952	0,4405	0,3989	0,3745	0,3501	0,3292
Ptbiserial	**0,8358**	0,7194	0,6327	0,6121	0,5353	0,4892	0,4814
CCC	35,7286	**37,5635**	36,0912	35,1108	34,1499	33,2106	32,6054
C indeks	**0,2718**	0,3425	0,3256	0,2987	0,3995	0,3693	0,4040
Gplus	57,0657	106,4747	97,1612	74,0282	83,9267	68,9558	**56,6177**
McClain	**0,2622**	0,5158	0,7212	0,7751	1,0456	1,2553	1,2797
SDbw	0,1578	0,2928	0,1529	0,0544	0,1116	0,0892	**0,0350**
SD indeks	**1,2979**	1,7450	2,5920	2,9482	3,3035	3,9564	3,9828
Duda	0,4599	0,5123	0,4640	0,5377	**0,6399**	0,5573	0,6108
PseudoT^2	115,0825	59,0272	39,2717	41,2762	**20,2558**	19,0679	19,1183
Gap	**0,1282**	0,1264	-0,2948	-0,3444	-0,5854	-0,9505	-1,0964
Ball	77,4735	**26,4818**	14,7081	9,4050	6,6700	5,0634	3,9356
Hubert	0,0019	**0,0021**	0,0021	0,0022	0,0023	0,0023	0,0023
D indeks	0,8535	**0,6502**	0,5621	0,5119	0,4715	0,4425	0,4188

Çizelge 4.24 incelendiğinde Ward yöntemine göre, İris veri seti için Calinski-Harabasz (CH), Silhouette, CCC, Ball, Hubert ve D indeks uygun küme sayısını 3 olarak doğru tahmin etmiştir. Krzanowski-Lai (KL), Davies-Bouldin (DB), Dunn, Gamma, Tau, Ratkowsky, SD indeks, C indeks, McClain ve Gap indekslerine göre uygun küme sayısı 2 olarak tahmin edilirken; Duda ve PseudoT^2'ye göre uygun küme sayısı 6'dır. Öte yandan Gplus ve SDbw indeksinin uygun küme sayısını belirlemedeki performansı önemli derecede düşük olup, 8 küme olarak belirlenmiştir.

Çizelge 4.25. İris veri setinin k-ortalama yöntemine göre uygun küme sayıları ve indeks değerleri

***k*-ortalama kümeleme yöntemi**							
İNDEKS / *k*	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KL	5,9068	3,5663	**7,2495**	0,4117	0,6156	1,6869	5,3825
CH	513,9245	**561,6278**	530,7658	459,5058	433,4067	443,3948	440,6205
DB	**0,4744**	0,7256	0,8436	0,9987	1,0923	1,0070	1,0403
Silhouette	0,5528	**0,6810**	0,4981	0,3728	0,3263	0,3462	0,3519
Dunn	0,0765	0,0988	**0,1365**	0,0624	0,0739	0,0872	0,0974
Gamma	**0,9563**	0,9137	0,9157	0,8716	0,8762	0,9070	0,9200
Tau	**2775,771**	2603,635	2242,016	1887,013	1583,285	1600,174	1345,745
Ratkowsky	**0,5462**	0,4967	0,4413	0,4067	0,3737	0,3498	0,3302
Ptbiserial	**0,8345**	0,7146	0,6361	0,5521	0,5023	0,5119	0,4690
CCC	35,9428	**37,6701**	36,4682	34,3409	33,3747	33,4645	33,2318
C indeks	**0,2728**	0,3450	0,3211	0,3327	0,3594	0,3965	0,4007
Gplus	60,6542	108,4455	90,0451	115,8259	94,5785	70,5580	**51,3011**
McClain	**0,2723**	0,5255	0,7120	0,9903	1,2099	1,1407	1,3416
SDbw	0,1618	0,2257	0,3186	0,1542	0,1158	0,1341	**0,0713**
SD indeks	**1,2820**	1,7259	2,4707	3,1993	3,3704	3,4409	3,8586
Duda	**1,9253**	1,1915	0,5112	1,1340	0,8469	3,9365	0,9269
PseudoT^2	**-52,8667**	-9,3224	45,9014	-5,1981	5,2430	-17,9033	1,5780
Gap	**0,1448**	0,0250	-0,2748	-0,4927	-0,7702	-1,0016	-1,0134
Ball	76,1740	**26,2838**	14,3071	9,9645	7,0760	4,9652	3,7486
Hubert	0,0019	**0,0021**	0,0021	0,0022	0,0023	0,0023	0,0024
D indeks	0,8556	**0,6480**	0,5574	0,5148	0,4829	0,4428	0,4123

Çizelge 4.25 incelendiğinde *k*-ortalama yöntemine göre, İris veri seti için Calinski-Harabasz (CH), Silhouette, CCC, Ball, Hubert ve D indeks diğerlerine göre daha başarılı bir performans göstererek uygun küme sayısını 3 olarak doğru tahmin etmişlerdir. Bunun yanında Krzanowski-Lai (KL) ve Dunn indeksine göre uygun küme sayısı 4 ve Gplus ve SDbw indeksine göre uygun küme sayısı 8 olarak belirlenmiştir. Diğer indekslere göre uygun küme sayısı 2'dir.

Çizelge 4.26'da tek bağlantı, tam bağlantı, Ward ve k-ortalama kümeleme yöntemlerine göre küme geçerlilik indeksleri kullanılarak belirlenen en uygun küme sayıları yer almaktadır. Çizelgede koyu (bold) işaretlenmiş değerler, İris veri setinde etiketli küme sayısını (orijinal küme sayısı) doğru tahmin eden indekslerin küme sayılarını göstermektedir

Çizelge 4.26. İris veri seti için en uygun küme sayıları

İNDEKS	**Tek Bağlantı**	**Tam Bağlantı**	**Ward**	***k*-Ortalama**
KL	2	4	2	4
CH	**3**	4	**3**	**3**
DB	2	**3**	2	2
Silhouette	**3**	**3**	**3**	**3**
Dunn	2	8	2	4
Gamma	2	4	2	2
Tau	2	3	2	2
Ratkowsky	2	2	2	2
Ptbiserial	2	**3**	2	2
CCC	2	**3**	**3**	**3**
C indeks	4	**3**	2	2
Gplus	2	8	8	8
McClain	2	2	2	2
SDbw	8	8	8	8
SD indeks	2	2	2	2
Duda	2	4	6	2
PseudoT^2	2	4	6	2
Gap	2	**3**	2	2
Ball	2	**3**	**3**	**3**
Hubert	**3**	**3**	**3**	**3**
D indeks	**3**	**3**	**3**	**3**

Çizelge 4.26 incelendiğinde gerçekte 3 kümeli olan İris veri seti için Calinski-Harabasz (CH), Silhouette, CCC, Ball, Hubert ve D indeks hem hiyerarşik hem de k-ortalama yöntemine göre uygun küme sayısını doğru tahmin ederek diğer indekslere göre daha iyi performans sergilemişlerdir. Öte yandan Gplus ve SDbw indeksleri uygun küme sayısını tahmin etmede başarısız olmuşlardır.

4.2. Düzey 2 Bölgelerinin Kadın İşgücü İstatistikleri Bakımından Kümelenmesi

Ekonomik kalkınmanın en önemli koşulu üretim faktörlerinin olabildiğince yüksek düzeyde ve verimli olarak kullanılmasıdır. Ancak, bu faktörlerin başında gelen işgücünün kullanımında dünyanın bütün ülkelerinde erkeklerin lehine dengesiz bir dağılım söz konusudur [38]. Dünya işgücü piyasasında toplumsal cinsiyet eşitsizliğinin temel göstergeleri, kadın işsizlik oranları ve kadınların özellikle ekonomik anlamda güçlenmeleri için çok önemli olan işgücüne katılım oranlarıdır. Kadınların işgücüne katılımı, bir ülke için sürdürülebilir kalkınmanın en önemli unsurlarından biri olmasına karşın, Türkiye'de kadınların işgücüne katılım oranları erkeklere nazaran düşük olup, yıllara göre azalma eğilimi göstermektedir.

Kadınların işgücüne katılımını belirleyen birçok faktör bulunmaktadır. Araştırmalara göre sosyo-kültürel faktörler, kadınların eğitim düzeyi, düşük ücretler, medeni durum ve çocuk sayısı, kentleşme, kayıt dışılık oranı ve ekonomik krizler Türkiye'de kadınların işgücü piyasasına katılımlarında önemli belirleyici faktörlerdir. Bu faktörler içerisinden eğitim düzeyi, kadınların işgücü piyasasında daha fazla yer edinmeleri ve verimliliklerini belirlemede kritik düzeyde öneme sahiptir. Kadınların eğitim düzeyi arttıkça işgücüne katılımı da artmaktadır. Eğitim düzeyindeki artış ayrıca kadınların kendilerine duydukları saygıyı güçlendirmekte, fiziksel, duygusal ve ekonomik güven sağlamakta ve onlara rekabetçi çalışma becerileri edindirmektedir. Tahminlere göre eğitim düzeyindeki artışın kadınlar için istihdam edilebilirlik olasılığını, ilkokul mezunu kadınlar için %3'ten, yüksekokul ve fakülte mezunlarında %73'e çıkarmaktadır [20].

Çalışmanın bu bölümünde, Türkiye İstatistik Kurumu'nun (TÜİK) yayınladığı "İşgücü İstatistikleri Veri Tabanı" ve "Bölgesel İstatistikler Veri Tabanı"ndan derlenen 26 İstatistiki Bölge Birimi Sınıflandırması (İBBS) Düzey 2 bölgesinin 2014 yılına ait kadın eğitim ve işgücü istatistikleri kullanılarak, illerin kümelenmesine yer verilmiştir. Ayrıca değişkenler arasındaki ölçek farkını ortadan kaldırmak için herhangi bir değişkene ait kadın nüfusu, toplam kadın nüfusuna oranlanmıştır.

Küme geçerlilik indekslerinin hesaplanmasında k-ortalamalar kümeleme tekniği ve Öklid uzaklık ölçüsü kullanılmıştır. Ayrıca kümeleme sonuçlarının değerlendirilmesi için küme geçerlilik indeksleri ile belirlenen uygun küme sonuçlarına göre sınıflanan Düzey-2 bölgelerine, küme üyeliklerine göre diskriminant analizi uygulanarak doğru sınıflandırma oranları belirlenmiştir. Analizde kullanılan değişkenler Çizelge 4.27'de verilmiştir.

Çizelge 4.27. Analizde kullanılan değişkenler

Eğitim durumu
X_1: Okuma yazma bilmeyen kadın nüfusu oranı
X_2: İlkokul, ilköğretim, ortaokul ve dengi meslek okulu mezunu kadın nüfusu oranı
X_3: Lise ve dengi meslek okulu mezunu kadın nüfusu oranı
X_4: Yüksekokul veya fakülte (üniversite) mezunu kadın nüfusu oranı
İşgücüne katılım
X_5: Okuma yazma bilmeyen kadın nüfusun işgücüne katılım oranı
X_6: İlkokul, ilköğretim, ortaokul ve dengi meslek okulu mezunu kadın nüfusun işgücüne katılım oranı
X_7: Lise ve dengi meslek okulu mezunu kadın nüfusun işgücüne katılım oranı
X_8: Yüksekokul veya fakülte (üniversite) mezunu kadın nüfusun işgücüne katılım oranı
İstihdam
X_9: Okuma yazma bilmeyen kadın nüfusun istihdam oranı
X_{10}: İlkokul, ilköğretim, ortaokul ve dengi meslek okulu mezunu kadın nüfusun istihdam oranı
X_{11}: Lise ve dengi meslek okulu mezunu kadın nüfusun istihdam oranı
X_{12}: Yüksekokul veya fakülte (üniversite) mezunu kadın nüfusun istihdam oranı
İşsizlik
X_{13}: Okuma yazma bilmeyen kadın nüfusun işsizlik oranı
X_{14}: İlkokul, ilköğretim, ortaokul ve dengi meslek okulu mezunu kadın nüfusun işsizlik oranı
X_{15}: Lise ve dengi meslek okulu mezunu kadın nüfusun işsizlik oranı
X_{16}: Yüksekokul veya fakülte (üniversite) mezunu kadın nüfusun işsizlik oranı
İstihdam edilen faaliyet kolu
X_{17}: Tarım sektöründe istihdam edilen kadın nüfusu oranı
X_{18}: Sanayi sektöründe istihdam edilen kadın nüfusu oranı
X_{19}: Hizmet sektöründe istihdam edilen kadın nüfusu oranı
İşteki durum
X_{20}: Ücretli veya yevmiyeli olarak istihdam edilen kadın nüfusu oranı
X_{21}: Kendi hesabına çalışan veya işveren olarak istihdam edilen kadın nüfusu oranı
X_{22}: Ücretsiz aile işçisi olarak istihdam edilen kadın nüfusu oranı

Çizelge 4.28'da uygun küme sayısının belirlenmesinde kullanılan 21 içsel küme geçerlilik indeksinin indeks değerleri ve geçerli küme sayıları verilmektedir.

Çizelge 4.28. Küme geçerlilik indekslerinin değerleri ve uygun küme sayıları

k	**KL**	**CH**	**Dunn**	**Silhouette**	**Gamma**	**Tau**	**Ratkowsky**
2	2,7641	**23,7954**	0,2107	**0,3700**	0,5876	47,6862	0,3462
3	2,4852	23,2480	0,2933	0,3277	0,7312	**53,2062**	**0,3466**
4	1,2258	20,6120	0,2933	0,2977	0,7991	50,5354	0,3218
5	1,0326	19,8723	0,3518	0,2781	0,8038	40,3262	0,3271
6	**2,8579**	21,0840	**0,4574**	0,2998	**0,9041**	39,4892	0,3335
7	1,3157	19,2263	0,4574	0,2499	0,8815	30,2523	0,3170
8	0,9984	17,5476	0,4572	0,2996	0,8933	30,6585	0,3053
9	0,7758	16,4139	0,4572	0,2812	0,8764	23,8646	0,2922
10	1,1468	16,2951	0,4192	0,2658	0,8999	21,5262	0,2789
k	**PtBiserial**	**CCC**	**DB**	**Gplus**	**C indeks**	**McClain**	**SDbw**
2	0,5057	**11,076**	**0,8724**	16,7354	0,3696	**0,6445**	0,4428
3	**0,5440**	7,1564	0,8873	9,7815	0,4274	1,0353	0,2812
4	0,5310	5,9952	0,8913	6,3538	0,3645	1,3699	0,2340
5	0,4703	5,1780	0,9295	4,9231	0,4604	2,0187	0,1920
6	0,4783	4,1457	0,9508	2,0954	0,4683	2,2250	0,1647
7	0,4137	3,4593	1,0009	2,0338	0,4458	3,1052	0,1470
8	0,4180	4,3696	1,0337	1,8308	0,1941	3,0642	0,1131
9	0,3671	5,2890	1,1013	1,6831	0,1863	4,0929	0,1046
10	0,3511	4,0479	1,0604	**1,1969**	**0,1702**	4,6058	**0,0892**
k	**SD indeks**	**Duda**	**PseudoT^2**	**Gap**	**Ball**	**Hubert**	**D indeks**
2	**6,8411**	**0,7972**	**3,5612**	**-0,6462**	1,9523	0,1503	0,3636
3	7,1745	0,5818	6,4685	-0,7000	**1,0945**	**0,2150**	**0,2963**
4	7,9223	1,0384	-0,2590	-1,5456	0,5101	0,2627	0,2637
5	8,5575	0,7851	1,9162	-1,6293	0,3250	0,2666	0,2369
6	8,9260	1,2327	-0,9440	-1,7620	0,2067	0,2777	0,2142
7	10,1174	5,9371	-4,1578	-1,5132	0,1571	0,2795	0,1999
8	10,1387	0,5875	2,8090	-1,8283	0,1242	0,2829	0,1829
9	11,2064	0,3463	3,7757	-1,7316	0,0990	0,2829	0,1719
10	11,2913	0,6848	1,3810	-1,4671	0,0765	0,3037	0,1593

k-ortalamalar kümeleme analizi sonucunda hesaplanan Krzanowski-Lai (KL), Calinski-Harabasz (CH), Dunn, Silhouette, Gamma, Tau, PtBiserial ve CCC indekslerine göre maksimum indeks değerine ulaşılan küme sayısı, uygun küme sayısı olarak alınır. Çizelge 4.28 incelendiğinde, Krzanowski-Lai (KL), Dunn ve Gamma indeksleri en büyük değerini $k = 6$ küme için sırasıyla 2,8579, 0,4574 ve 0,9041 olarak almıştır ve küme sayısı arttıkça indeks değerleri azalmaktadır. Benzer şekilde Calinski-Harabazs (CH), Silhouette ve CCC indeksleri uygun küme sayısı $k = 2$ iken, Tau, Raskowsky ve PtBiserial indekslerine göre uygun küme sayısı $k = 3$ olarak belirlenmiştir.

Uygun küme sayısının belirlenmesinde kullanılan Davies-Bouldin (DB), Gplus, C indeks, McClain, SD indeks ve Sdbw indekslerine göre minimum indeks değerine ulaşılan küme sayısı uygun küme sayısı olarak belirlenebilmektedir. Çizelge 4.28 incelendiğinde McClain, Davies-Bouldin (DB) ve SD indekse göre uygun küme sayısı k = 2'dir. Bunun yanında C indeks, Gplus ve Sdbw indeks değerleri, küme sayısı arttıkça ciddi şekilde küçülmektedir. Çizelge 4.28'de küme sayısı k = 2,3,4,…,10 alındığından söz konusu geçerlilik indeksleri için uygun küme sayısı k = 10 olarak belirlenmiştir. Ancak analiz sonucunda k = 10 küme için 0,1702 ile en küçük değerini alan C indeksinin değeri, küme sayısı arttıkça azalış göstererek, küme sayısı k = 21 olduğunda 0,1702 değerinden 0,0830 değerine kadar düşüş göstermiştir. Benzer şekilde, Gplus ve Sdbw indeksleri k = 10 küme için sırasıyla, 1,1969 ve 0,0892 indeks değerleri ile en küçük değerini almış ve küme sayısı arttıkça azalış göstererek, küme sayısı k = 24 olduğunda 0,0830 ve 0,0065 değerine kadar düşmüştür. Söz konusu indekslerin uygun küme sayısının belirlenmesinde etkili olmadığı açıktır.

Diğer bazı küme geçerlilik indeksleri için uygun küme sayısı, indeks değerlerinin indeks için hesaplanan kritik değer ile karşılaştırmasına dayanır. Analiz sonucu elde edilen kritik değerler Çizelge 4.29'da verilmiştir.

Çizelge 4.29. Uygun küme sayısının belirlenmesinde kullanılan indeks kritik değerleri

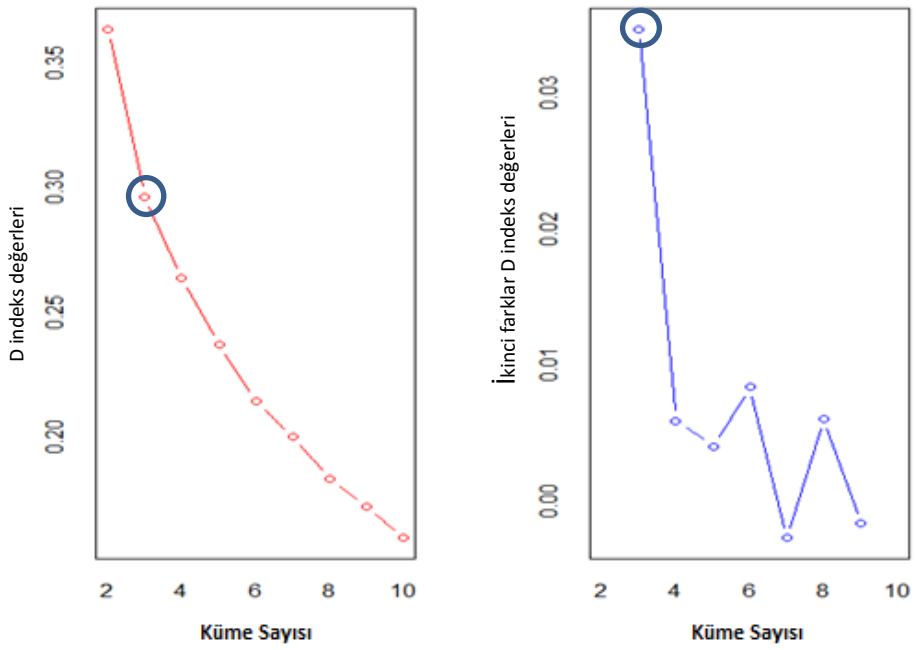
k	**Duda**	**PseudoT^2**	**Gap**
2	0,7308	5,1573	0,0874
3	0,6997	3,8628	0,8933
4	0,5674	5,3371	0,1433
5	0,6524	3,7290	0,2136
6	0,6028	3,2950	0,0691
7	0,5674	3,8122	0,5778
8	0,6028	2,6360	0,2069
9	0,5195	1,8501	-0,0126
10	0,5195	2,7752	0,0724

Duda indeksine göre uygun küme sayısı, Duda indeks değeri > Duda kritik değer kriterini sağlayan en küçük küme sayısıdır. Duda indeksinin k = 2 küme için indeks değeri 0,7972 iken, indeks kritik değeri 0,7308'dir. Bu durumda Duda indeksine göre uygun küme sayısı 2 olarak belirlenebilir.

PseudoT^2 indeksine göre uygun küme sayısının belirlenmesindeki kriter ise Duda indeksinin tam tersidir. PseudoT^2 indeks değeri< PseudoT^2 kritik değer kriterini sağlayan en küçük küme sayısı uygun küme sayısıdır. PseudoT^2 indeksinin k = 2 küme için indeks değeri 3,5612 iken, indeks kritik değeri 5,1573 değerini almıştır ve uygun küme sayısı 2 olarak belirlenmektedir.

Gap istatistiğine göre uygun küme sayısı, 0'dan büyük veya 0'a eşit en küçük küme sayısına ulaşılan kritik değerin elde edildiği küme sayısıdır. k = 2 küme için Gap kritik değeri 0,0874 değeri ile bu kriteri sağlamaktadır. Böylece gap indeksine göre uygun küme sayısı 2 olarak belirlenmiştir.

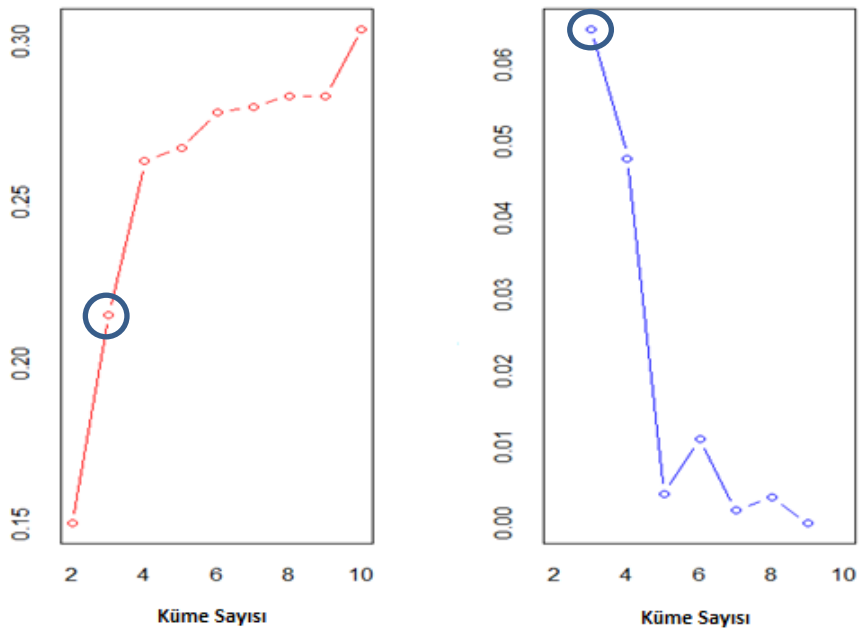
D indeks ve Hubert indeksleri uygun küme sayısının belirlenmesinde kullanılan grafiksel yöntemlerdir. Şekil 4.8 ve Şekil 4.9 de bu iki indeksin grafikleri verilmektedir.



(a) (b)

Şekil 4.8. D indeks grafikleri

Şekil 4.8 (b) incelendiğinde daire içine alınan zirve noktası, D indeksi için uygun küme sayının k = 3 olduğunu göstermektedir. Çizelge 4.28 ve şekil 4.8 (a)'da görüldüğü üzere 0,2963 indeks değeri ile uygun küme sayısı 3 olarak belirlenmiştir.



(a) (b)

Şekil 4.9. Hubert istatistiği grafiği

Şekil 4.9 (b) incelendiğinde daire içine alınan zirve noktası, Hubert istatistiği için uygun küme sayının k = 3 olduğunu göstermektedir. Çizelge 4.28 ve şekil 4.9 (a)'da görüldüğü üzere 0,2150 indeks değeri ile uygun küme sayısı 3 olarak belirlenmiştir.

k-ortalama kümeleme yöntemine göre belirlenen ve yukarıda ayrıntılı olarak açıklanan küme geçerlilik indekslerine göre uygun küme sayısı, Calinski-Harabazs (CH), Silhouette, CCC, Davies-Bouldin (DB), McClain, SD indeks, Duda, PseudoT^2 ve Gap indekslerine göre k = 2; Tau, Ratkowsky, PtBiserial, Ball, D indeks ve Hubert indekslerine göre k = 3; Krzanowski-Lai (KL), Dunn ve Gamma indekslerine göre k = 6 ve Gplus, C indeks ve SDbw indekslerine göre k =10 olarak belirlenmiştir. Bu sonuçlara göre Düzey 2 bölgelerinin kümelere dağılımı aşağıda verilmiştir.

Çizelge 4.30'te *k*-ortalama kümeleme yöntemi ile belirlenen ve uygun küme sayısı k = 2 olan küme geçerlilik indeksleri sonucunda elde edilen illerin kümelere dağılımı yer almaktadır.

Çizelge 4.30. Uygun küme sayısı $k = 2$ olan iller

KÜMELER	Düzey 2 Bölgeleri
Küme 1	TR10 (İstanbul) TR21 (Tekirdağ, Edirne, Kırklareli) TR22 (Balıkesir, Çanakkale) TR31 (İzmir) TR32 (Aydın, Denizli, Muğla) TR41 (Bursa, Eskişehir, Bilecik) TR42 (Kocaeli, Sakarya, Düzce, Bolu, Yalova) TR51 (Ankara) TR52 (Konya, Karaman) TR61 (Antalya, Isparta, Burdur) TR62 (Adana, Mersin) TR63 (Hatay, Kahramanmaraş, Osmaniye) TR72 (Kayseri, Sivas, Yozgat) TRC1 (Gaziantep, Adıyaman, Kilis)
Küme 2	TR33 (Manisa, Afyon, Kütahya, Uşak) TR71 (Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir) TR81 (Zonguldak, Karabük, Bartın) TR82 (Kastamonu, Çankırı, Sinop) TR83 (Samsun, Tokat, Çorum, Amasya) TR90 (Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane) TRA1 (Erzurum, Erzincan, Bayburt) TRA2 (Ağrı, Kars, Iğdır, Ardahan) TRB1 (Malatya, Elazığ, Bingöl, Tunceli) TRB2 (Van, Muş, Bitlis, Hakkari) TRC2 (Şanlıurfa, Diyarbakır) TRC3 (Mardin, Batman, Şırnak, Siirt)

Çizelge 4.30 incelendiğinde, 1. Kümede TR10 (İstanbul), TR31 (İzmir), TR41 (Bursa, Eskişehir, Bilecik) ve TR51 (Ankara) gibi gelişmiş bölgelerin yanında TR52 (Konya, Karaman), TR63 (Hatay, Kahramanmaraş, Osmaniye) ve TR72 (Kayseri, Sivas, Yozgat) gibi orta gelişmiş bölgeler yer alırken, 2. kümede TR83 (Samsun, Tokat, Çorum, Amasya), TR90 (Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane), TRA1 (Erzurum, Erzincan, Bayburt), TRB2 (Van, Muş, Bitlis, Hakkari) ve TRC3 (Mardin, Batman, Şırnak, Siirt) gibi az gelişmiş bölgeler yer almaktadır. 1. Kümede yer alan bölgelerdeki kadınların eğitim düzeyleri ve buna bağlı olarak istihdam ve işsizlik oranları 2. Kümede yer alan bölgelerden önemli derecede fazladır. 2. Kümede ücretsiz aile işçisi, tarım, hayvancılık, su ürünleri ve ormancılık sektöründe istihdam edilen kadın nüfusu oranları 1. Kümeye göre çok daha fazla olduğu görülmektedir. Diskriminant analizi sonucunda tekrar sınıflanan bölgelerin doğru sınıflandırma oranı %100 olarak belirlenmiştir.

Çizelge 4.31'de *k*-ortalama kümeleme yöntemine göre belirlenen ve uygun küme sayısı *k* = 3 olan küme geçerlilik indeksleri sonucunda elde edilen illerin kümelere dağılımı yer almaktadır.

Çizelge 4.31. Uygun küme sayısı *k* = 3 olan iller

KÜMELER	Düzey 2 Bölgeleri
Küme 1	TR10 (İstanbul) TR21 (Tekirdağ, Edirne, Kırklareli) TR31 (İzmir) TR41 (Bursa, Eskişehir, Bilecik) TR51 (Ankara)
Küme 2	TR22 (Balıkesir, Çanakkale) TR32 (Aydın, Denizli, Muğla) TR42 (Kocaeli, Sakarya, Düzce, Bolu, Yalova) TR52 (Konya, Karaman) TR61 (Antalya, Isparta, Burdur) TR62 (Adana, Mersin) TR63 (Hatay, Kahramanmaraş, Osmaniye) TR72 (Kayseri, Sivas, Yozgat) TRC1 (Gaziantep, Adıyaman, Kilis) TRC3 (Mardin, Batman, Şırnak, Siirt)
Küme 3	TR33 (Manisa, Afyon, Kütahya, Uşak) TR71 (Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir) TR81 (Zonguldak, Karabük, Bartın) TR82 (Kastamonu, Çankırı, Sinop) TR83 (Samsun, Tokat, Çorum, Amasya) TR90 (Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane) TRA1 (Erzurum, Erzincan, Bayburt) TRA2 (Ağrı, Kars, Iğdır, Ardahan) TRB1 (Malatya, Elazığ, Bingöl, Tunceli) TRB2 (Van, Muş, Bitlis, Hakkari) TRC2 (Şanlıurfa, Diyarbakır)

Çizelge 4.31 incelendiğinde, 1. Kümede yer alan TR10 (İstanbul), TR21 (Tekirdağ, Edirne, Kırklareli), TR31 (İzmir), TR41 (Bursa, Eskişehir, Bilecik) ve TR51 (Ankara), Türkiye'de eğitim düzeyi yüksek olan kadınların oransal olarak en yaygın olduğu bölgelerden oluşmaktadır. Bunun yanında eğitim düzeyi düşük kadınlara ilişkin işsizlik oranının en yüksek olduğu bölgelerde 1. Kümede yer almaktadır. 2. ve 3. kümede kadınların eğitim düzeyi düşük (lise altı) kadınların işsizlik oranları daha düşük iken, 3. Kümede lise ve üniversite mezunu kadınların işsizlik oranları 1. Kümedeki bölgelerden büyük ölçüde yüksek oranlara sahiptir. Bunun nedeni 3. Kümede yer alan bölgelerde kadın istihdamının büyük bir bölünün tarım sektöründe sağlanmasıdır.

Ayrıca diskriminant analizi sonucunda, *k*-ortalama yöntemine göre Tau, Ratkowsky, PtBiserial, Ball, D indeks ve Hubert indekslerinin verdiği uygun küme sayısına göre elde edilen kümeleme sonuçlarının doğru sınıflama yüzdesi %92,3 olarak tespit edilmiştir. Diskriminant analizi sonucuna göre 3. Kümede yer alan TR33 (Manisa, Afyon, Kütahya, Uşak) ve TR71 (Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir) yanlış sınıflandırılmıştır ve analiz sonucunda 2. Kümede yer almışlardır.

Çizelge 4.32'de *k*-ortalama kümeleme yöntemine göre belirlenen ve uygun küme sayısı $k = 6$ olan küme geçerlilik indeksleri sonucunda elde edilen illerin kümelere dağılımı yer almaktadır.

Çizelge 4.32. Uygun küme sayısı $k = 6$ olan iller

KÜMELER	Düzey 2 Bölgeleri
Küme 1	TR10 (İstanbul) TR31 (İzmir) TR41 (Bursa, Eskişehir, Bilecik) TR51 (Ankara)
Küme 2	TR21 (Tekirdağ, Edirne, Kırklareli) TR32 (Aydın, Denizli, Muğla) TR42 (Kocaeli, Sakarya, Düzce, Bolu, Yalova) TR61 (Antalya, Isparta, Burdur)
Küme 3	TR22 (Balıkesir, Çanakkale) TR52 (Konya, Karaman) TR62 (Adana, Mersin) TR63 (Hatay, Kahramanmaraş, Osmaniye) TR72 (Kayseri, Sivas, Yozgat) TRC1 (Gaziantep, Adıyaman, Kilis)
Küme 4	TR33 (Manisa, Afyon, Kütahya, Uşak) TR71 (Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir) TR81 (Zonguldak, Karabük, Bartın) TR82 (Kastamonu, Çankırı, Sinop) TR83 (Samsun, Tokat, Çorum, Amasya) TR90 (Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane) TRB1 (Malatya, Elazığ, Bingöl, Tunceli)
Küme 5	TRA1 (Erzurum, Erzincan, Bayburt) TRA2 (Ağrı, Kars, Iğdır, Ardahan) TRB2 (Van, Muş, Bitlis, Hakkari)
Küme 6	TRC2 (Şanlıurfa, Diyarbakır) TRC3 (Mardin, Batman, Şırnak, Siirt)

Çizelge 4.32 incelendiğinde, *k*-ortalama kümeleme yöntemine göre Krzanowski-Lai (KL), Dunn ve Gamma indekslerine göre uygun küme sayısı 6 olarak belirlenmiştir. Buna göre 1. Kümede TR10 (İstanbul), TR31 (İzmir), TR41 (Bursa, Eskişehir, Bilecik) ve TR51 (Ankara), kadınların eğitim düzeyine göre işsizlik ve istihdam oranının yüksek olduğu bölgeler yer alırken, 6. Kümede kadınların eğitim düzeyine göre işsizlik ve istihdam oranının çok düşük olduğu TRC2 (Şanlıurfa, Diyarbakır) ve TRC3 (Mardin, Batman, Şırnak, Siirt) bölgeler bulunmaktadır. Diğer illerde kendi aralarında 4 küme oluşturmuştur. Bölgelerin diskriminant analizi ile tekrar sınıflandırılması sonucunda doğru sınıflandırma oranı %100 olarak belirlenmiştir.

Çizelge 4.33'de *k*-ortalama kümeleme yöntemine göre belirlenen ve uygun küme sayısı k = 10 olan küme geçerlilik indeksleri sonucunda elde edilen illerin kümelere dağılımı yer almaktadır.

Çizelge 4.33. Uygun küme sayısı k = 10 olan iller

KÜMELER	Düzey 2 Bölgeleri
Küme 1	TR10 (İstanbul) TR51 (Ankara)
Küme 2	TR52 (Konya, Karaman) TR63 (Hatay, Kahramanmaraş, Osmaniye) TR72 (Kayseri, Sivas, Yozgat)
Küme 3	TR33 (Manisa, Afyon, Kütahya, Uşak) TR71 (Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir) TR83 (Samsun, Tokat, Çorum, Amasya) TRB1 (Malatya, Elazığ, Bingöl, Tunceli) TRC1 (Gaziantep, Adıyaman, Kilis)
Küme 4	TRA1 (Erzurum, Erzincan, Bayburt) TRA2 (Ağrı, Kars, Iğdır, Ardahan) TRB2 (Van, Muş, Bitlis, Hakkari)
Küme 5	TR22 (Balıkesir, Çanakkale) TR32 (Aydın, Denizli, Muğla) TR61 (Antalya, Isparta, Burdur)
Küme 6	TR81 (Zonguldak, Karabük, Bartın) TR82 (Kastamonu, Çankırı, Sinop) TR90 (Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane)
Küme 7	TRC2 (Şanlıurfa, Diyarbakır)
Küme 8	TRC3 (Mardin, Batman, Şırnak, Siirt)
Küme 9	TR42 (Kocaeli, Sakarya, Düzce, Bolu, Yalova) TR62 (Adana, Mersin)
Küme 10	TR21 (Tekirdağ, Edirne, Kırklareli) TR31 (İzmir) TR41 (Bursa, Eskişehir, Bilecik)

Çizelge 4.33 incelendiğinde, TRC2 (Şanlıurfa, Diyarbakır) ve TRC3 (Mardin, Batman, Şırnak, Siirt) bölgelerinin tek başlarına bir küme oluşturdukları, diğer bölgelerinde kendi aralarında 8 küme oluşturduğu gözlemlenmektedir. 1. Kümede TR10 (İstanbul) ve TR51 (Ankara) yer alırken, 10. Kümede büyük illerimizden TR21 (Tekirdağ, Edirne, Kırklareli), TR31 (İzmir) ve TR41 (Bursa, Eskişehir, Bilecik) kendi aralarında bir küme oluşturduğu gözlenmektedir. Ayrıca diskriminant analizi sonucunda, *k*-ortalama yöntemine göre Gplus, C indeks ve SDbw indekslerine göre elde edilen uygun küme sayısına göre kümeleme sonuçlarının doğru sınıflama yüzdesi %65 olarak tespit edilmiştir.

Düzey 2 bölgelerinin sınıflandırılmasında kullanılan küme geçerlilik indekslerinin doğru sınıflandırma oranlarını bir tablo ile gösterelim.

Çizelge 4.34. Küme geçerlilik indekslerinin doğru sınıflama oranları

İNDEKS	***k* = 2**	**İNDEKS**	***k* = 3**	**İNDEKS**	***k* = 6**	**İNDEKS**	***k* = 10**
CH	%100	**Tau**	%92,3	**KL**	%100	**Gplus**	%65
Silhouette	%100	**Ratkowsky**	%92,3	**Dunn**	%100	**C indeks**	%65
DB	%100	**Ptbiserial**	%92,3	**Gamma**	%100	**SDbw**	%65
CCC	%100	**Ball**	%92,3				
McClain	%100	**Hubert**	%92,3				
SD	%100	**D indeks**	%92,3				
indeks	%100						
Duda	%100						
PseudoT^2	%100						
Gap							

Çizelge 4.34 incelendiğinde kadın işgücü istatistiklerine göre Düzey 2 bölgelerinin kümelenmesinde, uygun küme sayısının $k = 2$ ve $k = 6$ olarak tahmin eden küme geçerlilik indekslerinin performanslarının, $k = 10$ küme olarak tahmin eden indekslerin performansından oldukça yüksek olduğu görülmektedir.

5. SONUÇ VE ÖNERİLER

Kümeleme analizinde anlamlı ve geçerli sonuçlara ulaşabilmek için uygun küme sayısının belirlenmesi birçok araştırmacının karşılaştığı en önemli sorunlardan biridir. Özellikle *k*-ortalama kümeleme yöntemi gibi bazı kümeleme algoritmalarında analizin başında küme sayısının belirlenmesi gerekmektedir. Uygun küme sayısının belirlenmesinde küme geçerlilik indeksleri kullanılmaktadır. Çalışmada ilk olarak uygun küme sayısının belirlenmesinde kullanılan 21 küme geçerlilik indeksi, R ortamında elde edilen farklı kümelenmiş yapay veri seti, geleneksel hiyerarşik kümeleme yöntemlerinden tek bağlantı yöntemi, tam bağlantı yöntemi ve Ward yöntemi, hiyerarşik olmayan kümeleme yöntemlerinden ise *k*-ortalama yöntemine göre karşılaştırılmıştır.

3 kümeli farklı yoğunluklu veri seti için yapılan karşılaştırmada, Krzanowski-Lai (KL), Calinski-Harabasz (CH), Davies-Bouldin (DB), Silhouette, Ptbiserial, SD indeks, Ball, D indeks ve Hubert indekslerinin hem hiyerarşik hem de k-ortalama yöntemine göre uygun küme sayısını doğru tahmin ederek diğer indekslere göre daha iyi performans sergiledikleri görülmüştür. Öte yandan Dunn, Gamma, Ratkowsky ve PseudoT^2 gibi bazı indeksler uygun küme sayılarını yaklaşık olarak bile tahmin edememişlerdir.

4 kümeli iyi ayrılmış veri seti için yapılan karşılaştırmada, küme geçerlilik indekslerinin uygun küme sayısını belirlemede diğer veri setlerine göre genel olarak daha başarılı olduklarını söyleyebiliriz. Calinski-Harabasz (CH), Davies-Bouldin (DB), Dunn, Silhouette, Ptbiserial, SD indeks gibi çoğu indeksin hem hiyerarşik hem de k-ortalama yöntemine göre uygun küme sayısını doğru tahmin ederek diğer indekslere göre daha iyi performans sergilemişlerdir. Gamma, Tau, Ratkowsky, C indeks, McClain ve PseudoT^2 indekslerinin uygun küme sayısını tahmin etmede oldukça başarısız olmuşlardır.

Gerçekte 6 kümeli olarak üretilen alt kümeli veri seti için yapılan karşılaştırmada, indekslerin uygun küme sayısı olarak sub-optimal küme sayısını tahmin ettiği görülmektedir. Bu durum göz önüne alındığında alt kümeli veri seti için Davies-Bouldin (DB), Dunn, Silhouette, Ratkowsky, Ptbiserial, Ball, Hubert ve D indeksin uygun küme sayını belirlemede diğer indekslere göre daha başarılı olduklarını söyleyebiliriz.

Cassini veri seti için yapılan karşılaştırmada, küme geçerlilik indekslerinin genel olarak uygun küme sayılarını belirlemede başarısız olduğu görülmektedir. Küme geçerlilik indekslerinin diğer veri setlerinin uygun küme sayısını belirlemedeki performansları göz önüne alındığında, bu başarısızlığın Cassini veri setinin küme yapısındaki yanlış değerlendirmeden kaynaklandığı düşünülmektedir.

İris veri seti için yapılan karşılaştırmada, Calinski-Harabasz (CH), Silhouette, CCC, Ball, Hubert ve D indeks uygun küme sayısını belirlemede diğer indekslere göre daha iyi performans sergilemişlerdir. Gplus ve SDbw indeksleri uygun küme sayısını tahmin etmede başarısız olmuşlardır.

Yapay veri uygulamalarında genel olarak Calinski-Harabasz(CH), Davies-Bouldin (DB), Silhouette, Ptbiserial ve SD indeksin uygun küme sayısını belirlemede diğerlerine oranla daha güvenilir olduğu görülmektedir. Krzanowski-Lai(KL), Gamma, Ratkowsky, CCC, McClain, Duda ve PseudoT^2 indeksleri tek bağlantı, tam bağlantı, ward ve *k*-ortalama kümeleme yöntemleri için uygun küme sayısını belirlemede değişken performanslar göstermekte ve genelde başarılı olamamaktadır. Bunun yanında Tau, Gplus, C indeks ve SDbw indeksleri uygun küme sayısının yaklaşık olarak bile tahmin edemeyerek başarısız olmuşlardır.

Uygulamanın ikinci bölümünde ise Eurostat tarafından belirlenen 26 İstatistiki Bölge Birimi Sınıflandırması (İBBS) Düzey 2 bölgesinin 2014 yılına ait kadın eğitim ve işgücü istatistikleri kullanılarak, 21 küme geçerlilik indeksinin k-ortalama kümeleme yöntemine göre belirlenen uygun küme sayısına göre kümeleme performansları değerlendirildi.

k-ortalama kümeleme yöntemine göre belirlenen Calinski-Harabazs (CH), Silhouette, CCC, Davies-Bouldin (DB), McClain, SD indeks, Duda, PseudoT^2 ve Gap indekslerine göre İstanbul, İzmir, Ankara gibi gelişmiş bölgelerin ve Konya, Kayseri, Adana gibi orta gelişmiş bölgeler bir kümede, Samsun, Trabzon az gelişmiş bölgelerin bir kümede yer aldığı, kadınların eğitim düzeyleri ve buna bağlı olarak istihdam ve işsizlik oranlarına dayalı genel bir sınıflama elde edildi. Diskriminant analizi sonucunda tekrar sınıflanan bölgelerin doğru sınıflandırma oranı %100 olarak belirlendi.

Tau, Ratkowsky, PtBiserial, Ball, D indeks ve Hubert indekslerine göre iyi, orta ve az gelişmiş bölgelerden oluşan 3 küme elde edilmiştir. İstanbul, İzmir, Bursa, Eskişehir ve Ankara Türkiye'de eğitim düzeyi yüksek olan kadınların oransal olarak en yaygın olduğu bölgelerden olmasına karşın eğitim düzeyi düşük kadınlara ilişkin işsizlik oranının en yüksek olduğu bölgelerdir. Trabzon, Erzurum, özelliklede de Şırnak ve Mardin gibi az gelişmiş bölgelerde lise ve üniversite mezunu kadınların işsizlik oranları, iyi ve orta gelişmiş bölgelerden oldukça yüksek oranlara sahiptir ve bu bölgelerde kadın istihdamının büyük bir bölünün ücretsiz aile işçisi, tarım, hayvancılık ve su ürünleri sektöründe sağlanmasıdır. Ayrıca diskriminant analizi sonucunda kümeleme sonuçlarının doğru sınıflama yüzdesi %92,3 olarak tespit edilmiştir.

Krzanowski-Lai (KL), Dunn ve Gamma indekslerine göre uygun küme sayısı 6 olarak belirlenmiştir. Bölgelerin diskriminant analizi ile tekrar sınıflandırılması sonucunda doğru sınıflandırma oranı %100 olarak belirlenmiştir.

Gplus, C indeks ve SDbw indekslerine uygun küme sayısını belirlemedeki performansı diğer indekslere göre önemli derecede düşüktür. Diskriminant analizi sonucunda, *k*-ortalama yöntemine göre bu indekslerinin verdiği uygun küme sayısına göre elde edilen kümeleme sonuçlarının doğru sınıflama yüzdesi %65 olarak tespit edilmiştir.

Araştırma sonuçlarına göre incelenen küme geçerlilik indekslerinden bazıları yapay veri setlerinin mevcut küme sayılarını tahmin etmede oldukça başarısız olduğu gözlenmiştir. Bununla birlikte, küme geçerlilik indekslerinin farklı kümeleme algoritmalarına göre farklı performans göstermeleri muhtemel bir durumdur. Ancak küme geçerlilik indeksinin herhangi bir kümeleme yöntemine göre veri setinin uygun küme sayısını ne kadar doğru verdiği, o indeksin uygun küme sayısını belirlemedeki başarını göstermektedir. Hem yapay veri setleri hem de kadın işgücü ve eğitim istatistikleri kullanılarak yapılan karşılaştırmada Calinski-Harabasz (CH), Davies-Bouldin (DB), Silhouette, Dunn ve SD indeksin uygun küme sayısını belirlemede diğerler indekslere oranla daha güvenilir ve doğru sonuçlar verdiği görülmektedir. Krzanowski-Lai(KL), Gamma, Ratkowsky, Ptbiserial, CCC, McClain, Tau, Duda ve PseudoT^2 indeksleri yapay verilerde tek bağlantı, tam bağlantı, ward ve *k*-ortalama kümeleme yöntemleri için uygun küme sayısını belirlemede değişken performanslar göstermekte ve genelde başarılı olamadıkları tespit edilmiştir. Ayrıca Gplus, C indeks ve SDbw indeksleri hem yapay veriler hem de gerçek hayat verisi kullanılarak

yapılan karşılaştırmalarda uygun küme sayısının yaklaşık olarak bile tahmin edemeyerek başarısız olmuşlardır.

Literatürde mevcut küme geçerlilik indekslerinin kümelerin şekli, hacmi ve uzaydaki serpilmelerini dikkate alarak uygun küme sayısını mümkün olduğunca doğru verecek şekilde sınanmaları önemlidir. Elde edilen bu sonuçlar ışığında uygun küme sayısının belirlenmesinde en güvenilir geçerlilik indeksi Silhouette olmakla birlikte, SD indeks, Calinski-Harabasz (CH) ve grafik yöntem olarak başarılı olan Hubert ve D indekste kullanılabilir. Küme geçerlilik indekslerinin diğer farklı kümeleme algoritmaları için karşılaştırılması daha genelleştirilebilir önerilere ulaşılmasını sağlayacaktır.

KAYNAKLAR

1. Anderberg, M.R. (1973). *Cluster Analysis for Applications*. London: Academic Press Inc., 359.

2. Atbaş, A.C. (2008). *Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma.* Yüksek Lisans Tezi, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

3. Aydın, N., Seven, A.N. (2015). İl Nüfus ve Vatandaşlık Müdürlüklerinin İş Yoğunluğuna Göre Hibrid Kümeleme İle Sınıflandırılması. *Yönetim ve Ekonomi Araştırmaları Dergisi,* 13 (2), 181-201.

4. Baker, F.B. and Hubert, L.J. (1975). Measuring the Power of Hierarchical Cluster Analysis. *Journal of the American Statistical Association,* 70(349), 31–38.

5. Ball, G.H. and Hall, D.J. (1965). ISODATA: A Novel Method of Data Analysis and Pattern Classification. California: *Stanford Research Institute,* Menlo Park.

6. Bolshakova, N. and Azuaje, F. (2003). Cluster Validation Techniques for Genome Expression Data. *Signal Processing*, 83(4), 825-833.

7. Calinski, T. and Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics – Theory and Methods,* 3(1), 1–27.

8. Charrad, M., Ghazzali, N., Boiteau, V. and Nicnafs, A., (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61(6), 4-22.

9. Çakmak, Z. (1999). Kümeleme Analizinde Geçerlilik Problemi ve Kümeleme Sonuçlarının Değerlendirilmesi. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi,* 189-205.

10. Davies, D.L. and Bouldin, D.W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1(2), 224–227.

11. Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons.

12. Dudoit, S. and Fridlyand, J. (2002). A Prediction-based Resampling Method for Estimating the Number of Clusters in a Dataset. *Genome Biology*, 3(7), 1-21.

13. Dunn, J. (1974). *Well Separated Clusters and Optimal Fuzzy Partitions.* Journal Cybernetics, 4(1), 95–104.

14. Erişoğlu, M. (2011). *Uzaklık Ölçülerinin Kümeleme Analizine olan Etkilerinin İncelenmesi ve Geliştirilmesi.* Doktora Tezi, Çukurova Üniversitesi Fen Bilimleri Enstitüsü, Adana.

15. Everitt, B. (1974). *Cluster Analysis.* London: Heinmann, 122.

16. Friedman, H.P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of American Statistical Association,* 62, 1159-1178.

17. Gan, G., Ma, C. and Wu J. (2007). Data Clustering: Theory, Algorithms, and Applications. *Society For Industrial & Applied Mathematics,* U.S. 466.

18. Gilbert, E.W. (1958). Pioneer maps of health and disease in England. *Geographical Journal*, 124, 172-183.

19. Guerra, L., Robles, V., Bielza, C.and Larranaga, P. (2012). A Comparison Study of Clustering Quality Indices Using Outliers and Noise. *Intelligent Data Analysis,* 16, 703-715.

20. Günsoy, G. ve Özsoy, C. (2012). Türkiye'de Kadın İşgücü, Eğitim ve Büyüme İlişkisinin VAR Analizi. *Finans Politik & Ekonomik Yorumlar,* 49(568), 21-39.

21. Gordon, A.D. (1999). *Classification*. 2nd edn. NewYork: Chapman & Hall, NY. 256.

22. Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2000). Quality Scheme Assessment in the Clustering Process. *In Principles of Data Mining and Knowledge Discovery. volume 1910 of Lecture Notes in Computer Science,* pp. 265–276. Springer-Verlag, Berlin Heidelberg. Proceedings of the 4th European Conference, PKDD 2000, Lyon, France, September 13–16.

23. Halkidi, M., Batistakis, Y., Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of İntelligent Information Systems,* 17(2/3) , 107-145.

24. Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). Clustering validity assessment: Finding the optimal partitioning of a data set. *in ICDM,* Washington, DC, USA, 187–194.

25. Hubert, L.J. and Levin, J.R. (1976). A General Statistical Framework for Assessing Categorical Clustering in Free Recall. *Psychological Bulletin,* 83(6), 1072–1080.

26. Hubert, L.J. and Arabie, P. (1985). Comparing Partitions. *Journal of Classification,* 2(1), 193–218.

27. Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J. (2010). Understanding of Internal Clustering Validation Measures. *IEEE International Conference on Data Mining,* 35, 911-916.

28. Krzanowski, W.J. and Lai, Y.T. (1988). A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. *Biometrics,* 44(1), 23–34.

29. Koldere, Y. (2008). *Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi.* Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

30. Kovacs, F., Legany, C. and Babos, A. (2005). Cluster Validity Measurement Techniques. *6th International Symposium of Hungarian Researchers on Computational Intelligence*, Hungary.

31. Lebart, L., Morineau, A. and Piron, M. (2000). *Statistique Exploratoire Multidimensionnelle*. Paris: Dunod.

32. Legány, C., Juhász, S. and Babos, A. (2006). Cluster Validity Measurement Techniques. Proceeding of the 5th. WSEAS Int.Conf. on Artificial, *Knowledge Engineering and Data bases*, Madrid, Spain, February 15-17, 388-393.

33. Maulik, U. and Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE PAMI*, 24, 1650–1654.

34. McClain, J.O. and Rao, V.R. (1975). CLUSTISZ: A Program to Test for The Quality of Clustering of a Set of Objects. *Journal of Marketing Research,* 12(4), 456–460.

35. Milligan, G.W. (1980). An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika,* 45(3), 325–342.

36. Milligan, G.W. (1981). A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis. *Psychometrika,* 46(2), 187–199.

37. Milligan, G.W. and Cooper M.C. (1985). An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika,* 50, 159–179.

38. Özer, M. ve Biçerli, K. (2003-2004). Türkiye'de Kadın İşgücünün Panel Veri Analizi. *Anadolu Üniversitesi Sosyal Bilimler Dergisi,* 3(1), 55-85.

39. Ratkowsky, D.A. and Lance, G.N. (1978). A Criterion for Determining the Number of Groups in a Classification. *Australian Computer Journal,* 10(3), 115–117.

40. Rohlf, F.J. (1974). Methods of Comparing Classifications. *Annual Review of Ecology and Systematics,* 5, 101–113.

41. Rousseeuw, P. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics,* 20, 53–65.

42. Sarle, W.S. (1983). Cubic Clustering Criterion. *SAS Technical Report A-108,* SAS Institute Inc. Cary, NC.

43. Sharma, S. (1996). *Applied Multivariate Techniques*. New York: John Wiley and Sons Inc.

44. Tatlıdil, H. (1996). *Uygulamalı Çok Değişkenli Analiz.* Ankara: Cem Web Ofset, 329-343.

45. Theodoridis, S. and Koutroubas, K. (2008). Pattern Recognition. *Academic Press, 4.*

46. Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the Number of Clusters in a Data Set Via the Gap Statistic. *Journal of the Royal Statistical Society* B, 63(2), 411–423.

47. İnternet: Türkiye İstatistik Kurumu Bölgesel İstatistikler Veri Tabanı, URL: http://www.webcitation.org/query?url=https%3A%2F%2Fbiruni.tuik.gov.tr%2Fbolgeselistatistik%2FsorguSayfa.do%3Ftarget%3Ddegisken&date=2016-07-24, Son Erişim Tarihi: 24.07.2016.

48. İnternet: Türkiye İstatistik Kurumu İşgücü İstatistikleri Veri Tabanı, URL: http://www.webcitation.org/query?url=https%3A%2F%2Fbiruni.tuik.gov.tr%2Fmedas%2F%3Fkn%3D72%26locale%3Dtr&date=2016-07-24, Son Erişim Tarihi: 24.07.2016.

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : HACIOĞLU, Hazan Kübra
Uyruğu : T.C.
Doğum tarihi ve yeri : 04.07.1989, RİZE
Medeni hali : Bekar
Telefon : 0 (542) 625 18 53
e-mail : hazanhacioglu@gmail.com

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek lisans	Gazi Üniversitesi /İstatistik Bölümü	-
Lisans	Eskişehir Osmangazi Üniversitesi / İstatistik	2013
Lise	Necat Sağbaş Anadolu Lisesi	2007

Yabancı Dil

İngilizce

Hobiler

Kitap okuma, Tiyatro ve Sinema, Müzik

GAZİ GELECEKTİR...