T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
FEN FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

Kümeleme Analizi: K-Means, K-Medoids (PAM) ve Hiyerarşik Kümeleme

Yönlendirilmiş Grup Çalışması Ödev Raporu

Hazırlayanlar:
Ece Özgöcen
Çağan Arın Gürsu
Süleyman Rubar Oral

01.01.2026

Özet

Bu raporda kümeleme (denetimsiz öğrenme) kapsamında en yaygın yöntemlerden **K-Means**, **K-Medoids (PAM)** ve **Hiyerarşik Kümeleme** yaklaşımları sistematik olarak ele alınmıştır. İlk bölümde yöntemlerin matematiksel temeli, algoritma adımları, avantajları ve dezavantajları açıklanmış; ikinci bölümde ise sunum dosyasında üretilen analiz çıktıları (keşifsel veri analizi, standartlaştırma, elbow–silhouette grafikleri, farklı yöntemlerin sonuç karşılaştırmaları vb.) rapora görsellerle birlikte entegre edilmiştir. Kümeleme kalitesinin değerlendirilmesinde **Silhouette genişliği** ve **Dunn indeksi** gibi içsel ölçütlerin rolü vurgulanmış, küme sayısı seçimi ve yöntem seçimine ilişkin yorumlar sunulmuştur.

Anahtar kelimeler: Kümeleme, K-Means, K-Medoids, PAM, Hiyerarşik Kümeleme, Silhouette, Dunn.

İçindekiler

Şekil Listesi

Tablo Listesi

Bölüm 1

Kümeleme Analizine Giriş

1.1 Kümeleme Problemi

Kümeleme, gözlemleri önceden tanımlı bir sınıf etiketi olmadan, benzerliklerine göre gruplamayı hedefleyen bir denetimsiz öğrenme problemidir. Temel amaç iki parçalıdır: (i) aynı küme içindeki gözlemler birbirine mümkün olduğunca benzer olmalı (kompaktlık), (ii) farklı kümelerdeki gözlemler ise birbirinden mümkün olduğunca ayrışmalıdır (separation). Bu bakış açısı, birçok kümeleme indeksinin de temelini oluşturur .

1.2 Uzaklık / Benzerlik Ölçüleri

Kümeleme algoritmalarının önemli bir kısmı bir uzaklık ölçüsü üzerine kuruludur. Sık kullanılan metrikler:

- **Öklid uzaklığı**: $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$
- **Manhattan uzaklığı**: $d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} |x_j - y_j|$
- **Minkowski**: $d(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^{p} |x_j - y_j|^q\right)^{1/q}$

Değişken ölçekleri farklı olduğunda (örneğin biri binlerle, diğeri 0–1 aralığında) standartlaştırma yapılmadığı takdirde büyük ölçekli değişkenler mesafeyi domine eder. Bu nedenle özellikle K-Means gibi mesafe-temelli yöntemlerde z-skor standardizasyonu kritik bir ön adımdır.

1.3 Kümeleme Kalitesi: Kompaktlık ve Ayrışma

Bir kümeleme çözümünün iyi kabul edilebilmesi için hem küme içi benzerlik yüksek, hem de kümeler arası ayrışma yüksek olmalıdır. İçsel doğrulama ölçütleri genellikle bu iki bileşeni birlikte kullanır. Örneğin, kompaktlık küme içi ortalama uzaklıklar veya küme içi çap (diameter) ile; ayrışma ise kümeler arası minimum uzaklıklarla ölçülebilir.

Bölüm 2

K-Means Kümeleme

2.1 Yöntemin Mantığı ve Amaç Fonksiyonu

K-Means, veriyi k adet kümeye ayırır ve her küme için bir **merkez (centroid)** tanımlar. Amaç fonksiyonu, küme içi kareler toplamını (within-cluster sum of squares, WCSS) minimize etmektir:

$$\text{WCSS} = \sum_{r=1}^{k} \sum_{\mathbf{x}_i \in C_r} \|\mathbf{x}_i - \boldsymbol{\mu}_r\|^2 \tag{2.1}$$

Burada $\boldsymbol{\mu}_r$ küme r'nin centroid'idir. K-Means, özellikle kümelerin yaklaşık küresel (spherical) ve benzer yoğunlukta olduğu durumlarda etkili sonuç verir.

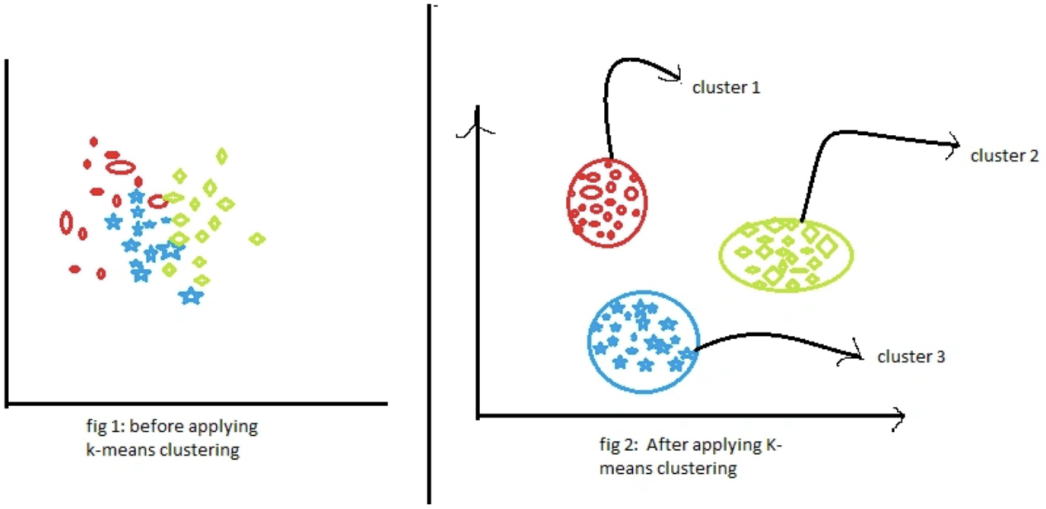
2.2 Algoritma Adımları

1. Küme sayısı k seçilir.
2. Başlangıç centroid'leri atanır (rastgele veya *k-means++*).
3. Her gözlem en yakın centroid'e atanır.
4. Her küme için centroid, o kümedeki gözlemlerin ortalaması olarak güncellenir.
5. Atamalar değişmeyene kadar 3–4. adımlar tekrarlanır.

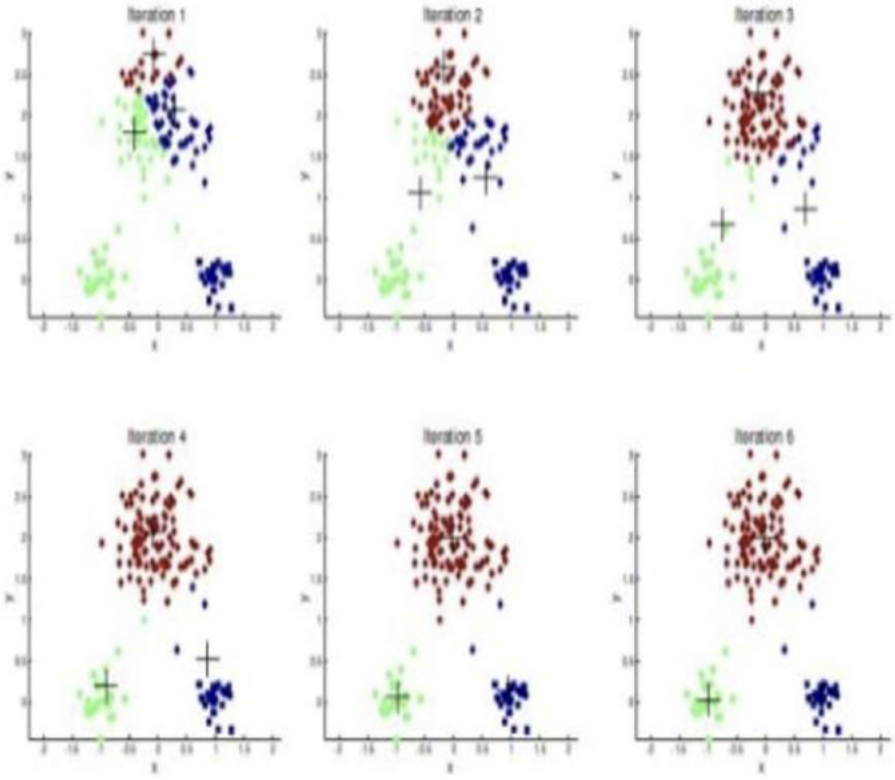
Algoritma yerel minimuma yakınsar; bu nedenle farklı başlangıçlarla birden fazla çalıştırma yaygın bir uygulamadır.

2.3 Avantajlar ve Dezavantajlar

- **Avantaj**: Uygulaması basit ve büyük veri setlerinde hızlıdır.
- **Dezavantajlar**: Aykırı değerlere duyarlıdır; başlangıç seçimine ve k seçimine bağlıdır.



Şekil 2.1: K-Means uygulanmadan önce ve sonra örnek küme görünümü.



Şekil 2.2: K-Means iterasyonları boyunca centroid güncellemelerine örnek görsel.

Bölüm 3

K-Medoids (PAM) Kümeleme

3.1 Yöntemin Mantığı

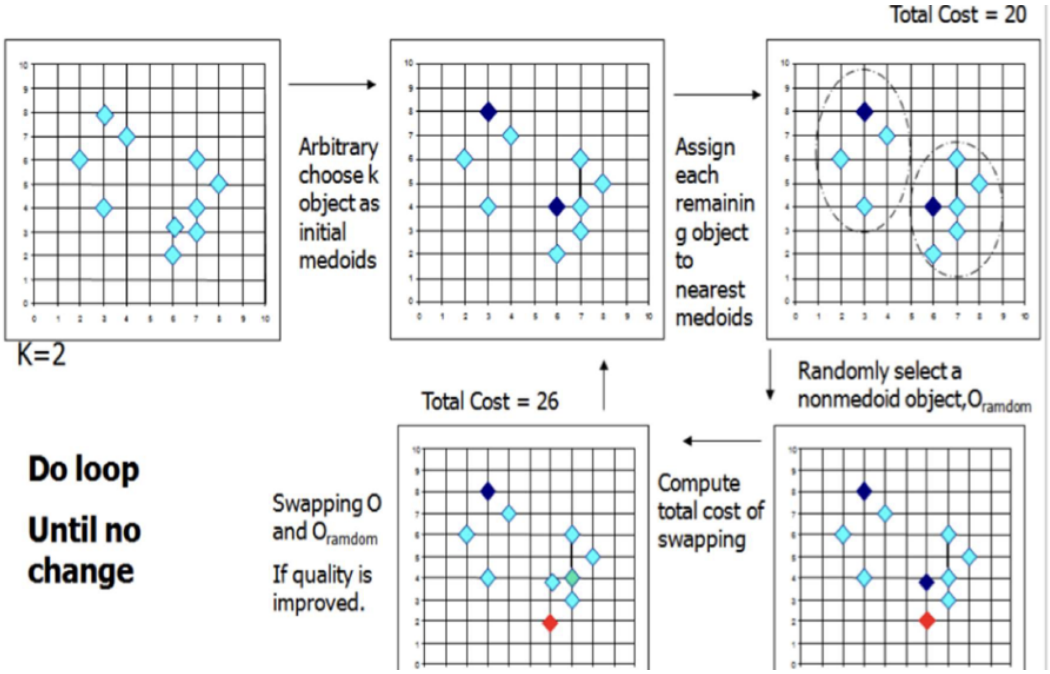
K-Medoids yönteminde her kümenin merkezi, veri setindeki **gerçek bir gözlem** olan *medoid* ile temsil edilir. Bu özellik, yöntemi K-Means'e göre aykırı değerlere daha dayanıklı hale getirir. Amaç fonksiyonu tipik olarak küme içi uzaklık toplamını minimize eder:

$$\sum_{r=1}^{k} \sum_{\mathbf{x}_i \in C_r} d(\mathbf{x}_i, \mathbf{m}_r) \tag{3.1}$$

3.2 PAM (Partitioning Around Medoids) Algoritması

PAM algoritması iki ana aşamadan oluşur: (i) başlangıç medoid'lerinin seçimi, (ii) takas (swap) operasyonları ile toplam maliyeti düşürecek iyileştirmelerin aranması.

1. k adet başlangıç medoid'i seçilir.
2. Her gözlem en yakın medoide atanır.
3. Medoid olmayan bir gözlem ile mevcut medoid'i takas etmenin toplam maliyete etkisi hesaplanır.
4. Maliyet azalıyor ise takas yapılır; değişim kalmayana kadar devam edilir.



Şekil 3.1: K-Medoids/PAM algoritmasının mantığını özetleyen örnek şema.

3.3 K-Means ve K-Medoids Karşılaştırması

Tablo 3.1: K-Means ve K-Medoids (PAM) karşılaştırması

Kriter	**K-Means**	**K-Medoids (PAM)**
Merkez tanımı	Ortalama (centroid)	Veri noktası (medoid)
Aykırı değerlere duyarlılık	Yüksek	Daha düşük
Hesaplama maliyeti	Daha hızlı	Daha maliyetli
Uzaklık türü	Genelde Öklid	Herhangi bir uzaklık
Yorumlanabilirlik	Orta (ortalama profil)	Yüksek (gerçek gözlem)

Bölüm 4

Hiyerarşik Kümeleme

4.1 Genel Bakış

Hiyerarşik kümeleme, gözlemler arasındaki benzerliği ağaç yapısında (dendrogram) ifade eder. Birleştirici (agglomerative) yaklaşımda her gözlem tek başına küme olarak başlar ve adım adım birleşerek daha büyük kümeler oluşturur. Bölücü (divisive) yaklaşımda ise tüm veri tek küme olarak başlar ve parçalanır. Pratikte en sık kullanılan agglomerative yaklaşımdır.

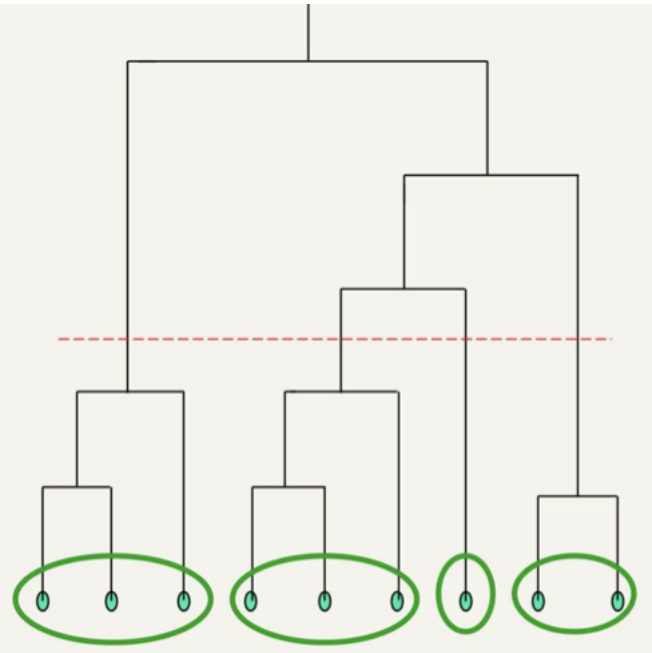
4.2 Bağlantı (Linkage) Yöntemleri

Bağlantı fonksiyonu iki küme arası uzaklığın nasıl tanımlandığını belirler:

- **Single linkage**: iki küme arasındaki en yakın iki noktanın uzaklığı (zincirleme etkisine açıktır).
- **Complete linkage**: iki küme arasındaki en uzak iki noktanın uzaklığı (daha kompakt kümeler).
- **Average linkage**: tüm çiftlerin ortalama uzaklığı.
- **Ward**: birleşme sonucunda küme içi varyans artışını minimize eder.

4.3 Zincirleme Etkisi (Chaining Effect)

Single linkage, veri uzayında birbirine yakın noktalar üzerinden "uzun zincir" şeklinde kümeler oluşturabilir. Bu durum dendrogramda tek bir kümenin çok sayıda gözlemi sırayla "yutması" şeklinde gözlenir ve pratikte yorumlamayı zorlaştırabilir.



Şekil 4.1: Dendrogram üzerinde kesim yüksekliğinin kümeleri belirlemesi örnek görsel.

Bölüm 5

Küme Doğrulama ve Optimal Küme Sayısı

5.1 Silhouette Genişliği

Silhouette analizi her gözlem için iki büyüklük tanımlar: $a(i)$ gözlemin kendi kümesindeki ortalama uzaklığı, $b(i)$ ise gözlemin en yakın diğer kümeye olan ortalama uzaklığıdır. Silhouette katsayısı:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{5.1}$$

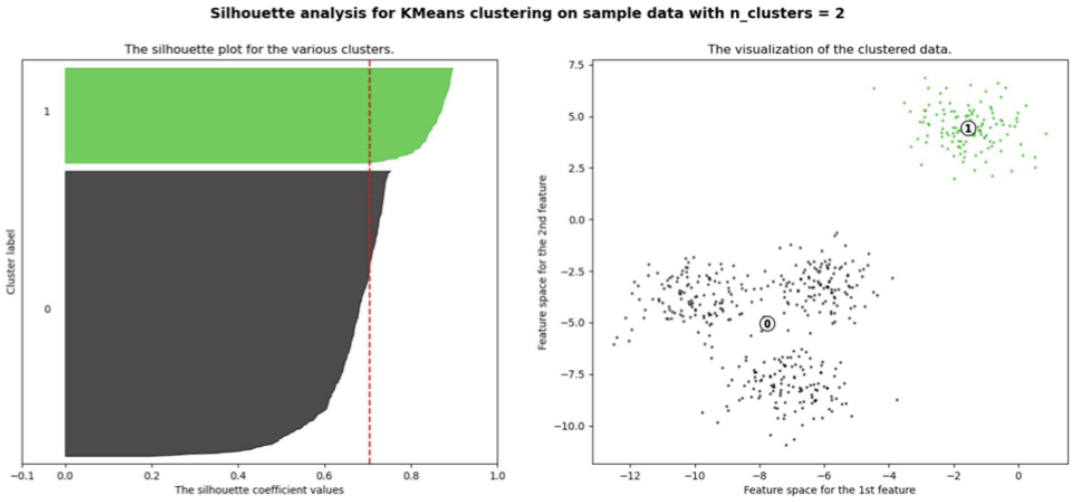
$s(i)$ değeri 1'e yakınsa iyi kümelenme, 0 civarı belirsizlik, negatif değer ise olası yanlış atamayı işaret eder.

5.2 Dunn İndeksi

Dunn indeksi ayrışma ve kompaktlığı birlikte ölçer:

$$D = \frac{\min\limits_{r \neq s} \delta(C_r, C_s)}{\max\limits_{t} \Delta(C_t)} \tag{5.2}$$

Burada $\delta(C_r, C_s)$ kümeler arası uzaklık, $\Delta(C_t)$ küme içi çaptır. Daha büyük Dunn değeri daha iyi ayrışma ve kompaktlık anlamına gelir.



Şekil 5.1: Silhouette analizi ve kümelenmiş veri görselleştirmesi örnek görsel.

Bölüm 6

Uygulama ve Bulgular

Bu bölümde veri setinin analiz çıktıları, rapor formatına uygun şekilde görsellerle birlikte sunulmuştur. Her bir alt başlıkta önce kısa yorum, ardından ilgili çıktı yer almaktadır.

6.1 Tanımlayıcı İstatistikler

Veri setinin temel özet istatistikleri (gözlem sayısı, ortalama, standart sapma, çeyreklikler ve maksimum) değişkenlerin ölçek farklılıklarını göstermesi açısından önemlidir. Özellikle bazı değişkenlerin dağılım genişliğinin çok yüksek olması, standardizasyon ihtiyacını desteklemektedir.

```
df.describe().T
```

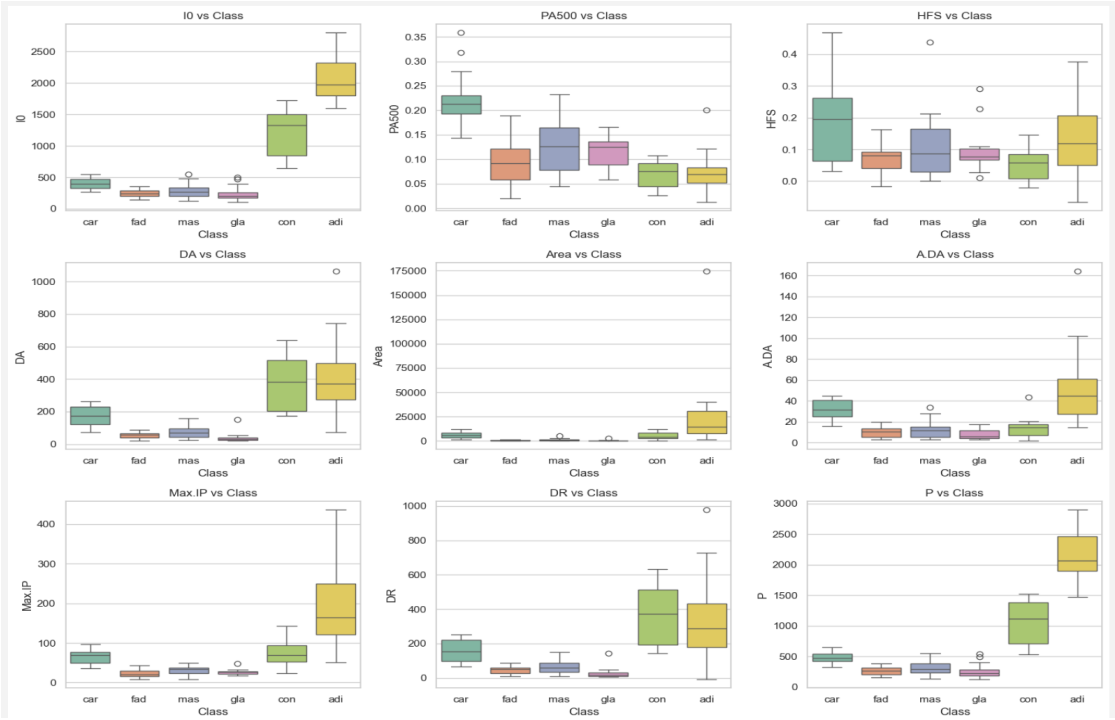
	count	mean	std	min	25%	50%	75%	max
I0	106.0	784.251618	753.950075	103.000000	250.000000	384.936489	1487.989626	2800.000000
PA500	106.0	0.120133	0.068596	0.012392	0.067413	0.105418	0.169602	0.358316
HFS	106.0	0.114691	0.101347	-0.066323	0.043982	0.086568	0.166504	0.467748
DA	106.0	190.568642	190.801448	19.647670	53.845470	120.777303	255.334809	1063.441427
Area	106.0	7335.155162	18580.314213	70.426239	409.647141	2219.581163	7615.204968	174480.476218
A.DA	106.0	23.473784	23.354672	1.595742	8.180321	16.133657	30.953294	164.071543
Max.IP	106.0	75.381258	81.345838	7.968783	26.893773	44.216040	83.671755	436.099640
DR	106.0	166.710575	181.309580	-9.257696	41.781258	97.832557	232.990070	977.552367
P	106.0	810.638127	763.019135	124.978561	270.215238	454.108153	1301.559438	2896.582483

Şekil 6.1: Tanımlayıcı istatistikler tablosu.

6.2 Boxplot Analizi

Sınıflara göre çizilen kutu grafikleri (boxplot), değişkenlerin sınıflar arasında ayırt ediciliğini ve aykırı değer davranışını görselleştirir. K-Means gibi yöntemler aykırı değerlere

duyarlı olduğundan, bu grafikler ön işleme kararlarını (ölçekleme, aykırı değer analizi vb.) yönlendirmek için kritiktir.



Şekil 6.2: Değişkenlerin sınıflara göre boxplot grafikleri.

6.3 Diğer Çıktılar

Aşağıda yer alan diğer görsel çıktılar, analize göre rapora eklenmiştir.

```
[6]: df['Class'].value_counts()

[6]: Class
     adi    22
     car    21
     mas    18
     gla    16
     fad    15
     con    14
     Name: count, dtype: int64
```

Şekil 6.3: Gözlem sayıları.

Yorum: Bu çıktıdaki bulgular Class değişkenindeki her kategoriye ait gözlem sayısını gösteren sınıf dağılımı göstermektedir.

Standartlaştırılmış Veri İstatistikleri

	I0	PA500	HFS	DA	Area	\
count	1.060000e+02	1.060000e+02	1.060000e+02	1.060000e+02	1.060000e+02	
mean	-5.865329e-17	-2.157603e-16	-2.890769e-16	-1.214961e-16	-2.932665e-17	
std	1.004751e+00	1.004751e+00	1.004751e+00	1.004751e+00	1.004751e+00	
min	-9.078691e-01	-1.578115e+00	-1.794562e+00	-9.000611e-01	-3.928481e-01	
25%	-7.119697e-01	-7.721970e-01	-7.010025e-01	-7.199772e-01	-3.745043e-01	
50%	-5.321468e-01	-2.155318e-01	-2.788055e-01	-3.675176e-01	-2.766302e-01	
75%	9.378356e-01	7.245996e-01	5.136792e-01	3.410553e-01	1.514400e-02	
max	2.686285e+00	3.488752e+00	3.500204e+00	4.596503e+00	9.038564e+00	

	A.DA	Max.IP	DR	P
count	1.060000e+02	1.060000e+02	1.060000e+02	1.060000e+02
mean	3.980045e-17	-2.545134e-16	3.435407e-16	-3.351617e-17
std	1.004751e+00	1.004751e+00	1.004751e+00	1.004751e+00
min	-9.412240e-01	-8.326514e-01	-9.751511e-01	-9.028828e-01
25%	-6.579462e-01	-5.988976e-01	-6.923121e-01	-7.116338e-01
50%	-3.157825e-01	-3.849401e-01	-3.816965e-01	-4.694820e-01
75%	3.217790e-01	1.024008e-01	3.672964e-01	6.464497e-01
max	6.048712e+00	4.455446e+00	4.493385e+00	2.746791e+00

Şekil 6.4: standartlaştırılmış Veri seti Tablosu.

Yorum:

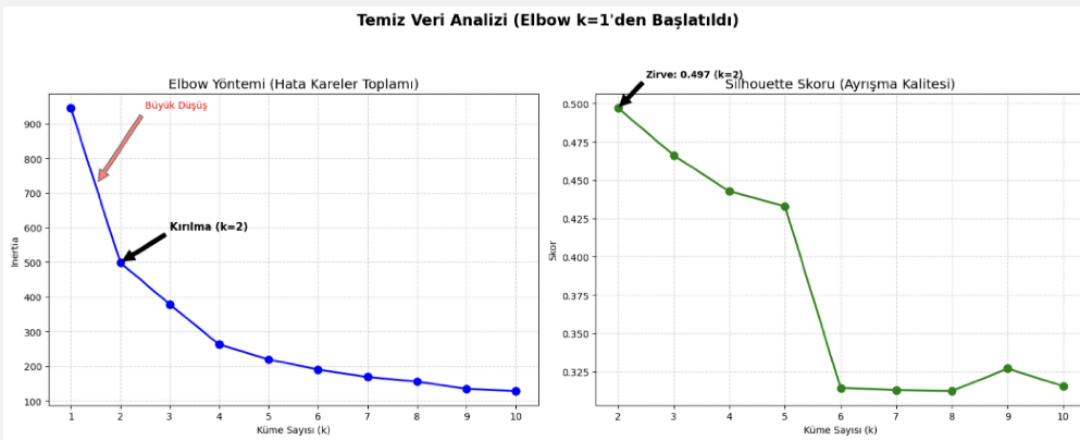
Korelasyon Matrisi Değerleri:

	I0	PA500	HFS	DA	Area	A.DA	Max.IP	\
I0	1.000000	-0.393647	0.028455	0.819606	0.560098	0.612070	0.823668	
PA500	-0.393647	1.000000	0.509019	-0.089817	0.083547	0.229837	-0.050401	
HFS	0.028455	0.509019	1.000000	0.106977	0.206059	0.356028	0.370827	
DA	0.819606	-0.089817	0.106977	1.000000	0.731132	0.648334	0.753227	
Area	0.560098	0.083547	0.206059	0.731132	1.000000	0.830172	0.735258	
A.DA	0.612070	0.229837	0.356028	0.648334	0.830172	1.000000	0.812815	
Max.IP	0.823668	-0.050401	0.370827	0.753227	0.735258	0.812815	1.000000	
DR	0.733252	-0.077054	0.011592	0.974202	0.675810	0.540695	0.600290	
P	0.988697	-0.345715	0.102362	0.774028	0.574073	0.679363	0.861837	

	DR	P
I0	0.733252	0.988697
PA500	-0.077054	-0.345715
HFS	0.011592	0.102362
DA	0.974202	0.774028
Area	0.675810	0.574073
A.DA	0.540695	0.679363
Max.IP	0.600290	0.861837
DR	1.000000	0.665987
P	0.665987	1.000000

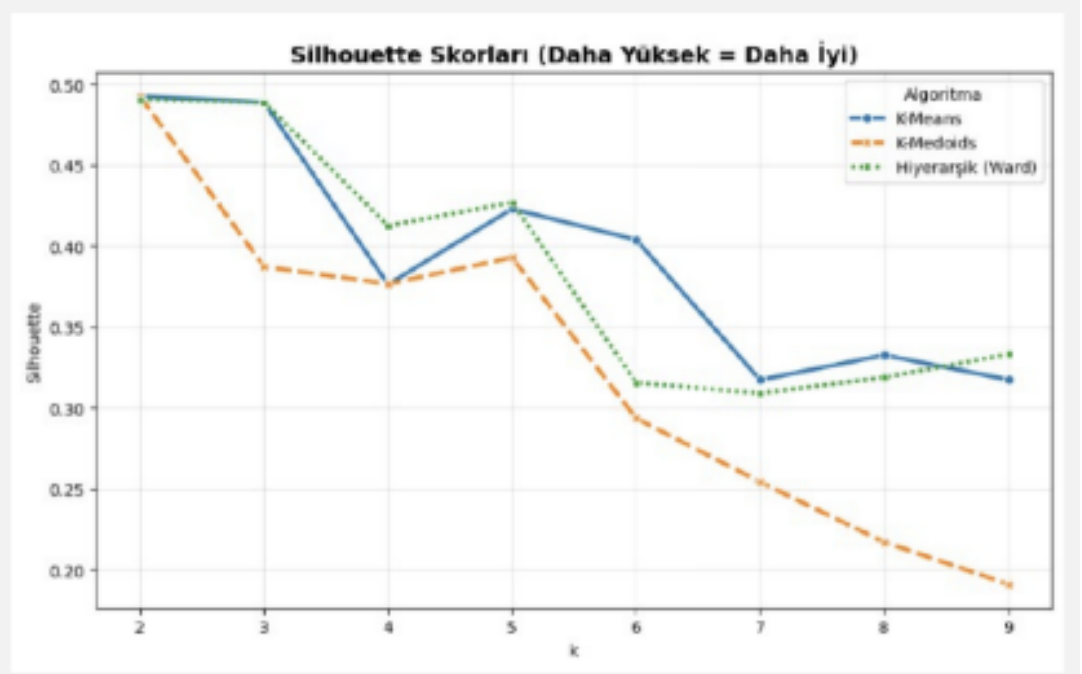
Şekil 6.5: Korelasyon Matrisi

Yorum: Korelasyon matrisi incelendiğinde, değişkenlerin bir kısmı arasında güçlü pozitif ilişkiler bulunduğu, özellikle spektral ve geometrik ölçümlerin birbirleriyle yüksek korelasyon gösterdiği görülmektedir; bu durum modelleme aşamasında çoklu bağlantı (multicollinearity) açısından dikkate alınmalıdır.



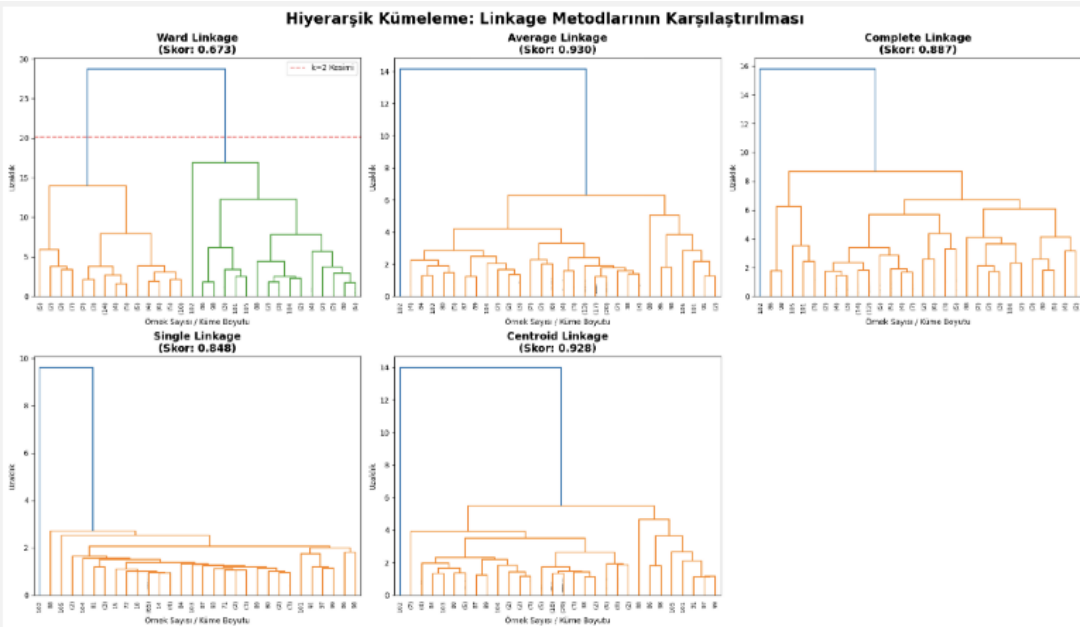
Şekil 6.6: Elbow ve silhouette grafikleri

Yorum: Elbow grafiği WCSS'nin k arttıkça nasıl azaldığını gösterir. Azalışın belirgin biçimde yavaşladığı nokta, uygun k için adaydır.Elbow grafiğinde WCSS değerinin k=2'den sonra belirgin biçimde daha yavaş azaldığı görülmektedir. Bu durum, küme sayısı olarak k=2'nin uygun bir aday olduğunu göstermektedir. Silhouette grafiği de düşük k değerlerinde daha yüksek ayrışma kalitesine işaret ederek bu bulguyu desteklemektedir.



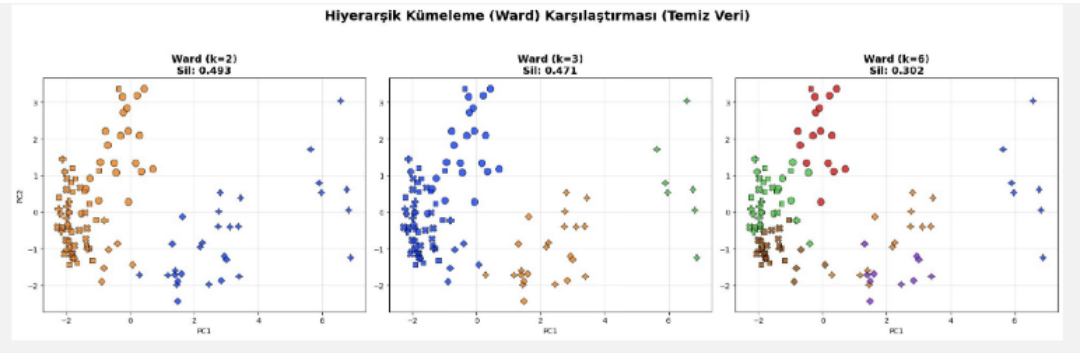
Şekil 6.7: Silhouette için ek grafik

Yorum: Bu çıktıdaki bulgulara göre,farklı kümeleme algoritmaları için hesaplanan silhouette skorları karşılaştırıldığında, düşük k değerlerinde genel olarak daha yüksek skorlar elde edildiği görülmektedir. Bulgular, küme sayısı ve yöntem seçiminin tek bir metriğe bağlı kalmadan, birlikte değerlendirilmesi gerektiğini göstermektedir.



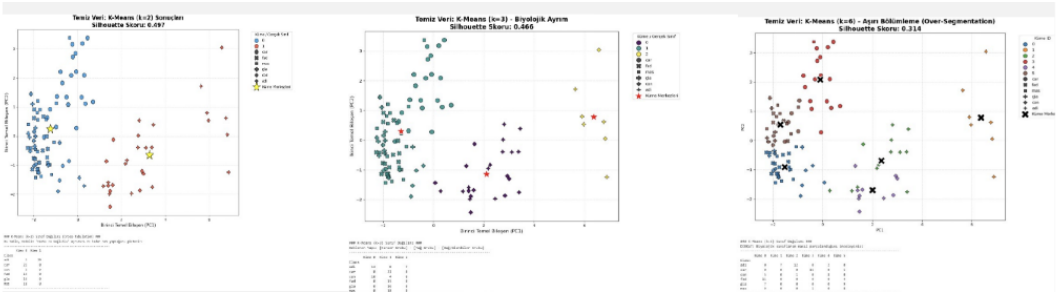
Şekil 6.8: Dendrogramlar üzerinde linkagelerin karşılaştırması

Yorum: Dendrogram, birleşme adımlarını ve uzaklık seviyelerini gösterir. Linkage seçimi yapıyı değiştirir; Ward genellikle daha kompakt kümeler üretir Single linkage zincirleme etkisi gösterirken, Ward linkage daha dengeli fakat ayrışma gücü görece daha düşüktür. Bu nedenle, veri yapısını daha iyi temsil eden yöntemler olarak Average ve Centroid linkage öne çıkmaktadır.



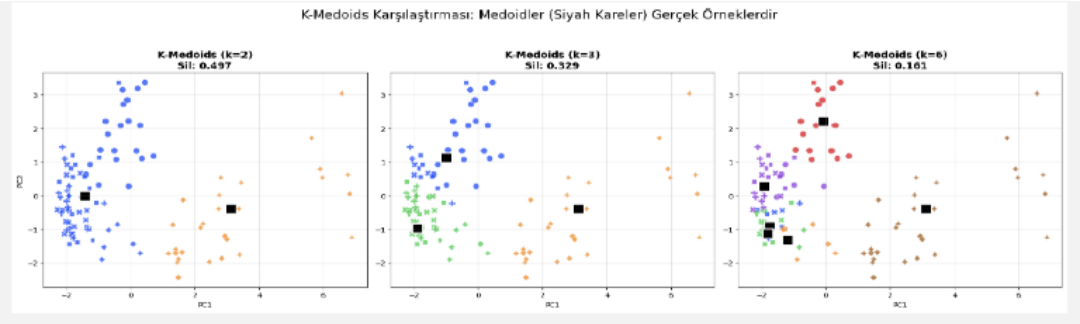
Şekil 6.9: Hiyerarşik kümeleme karşılaştırması

Yorum: Bu çıktıdaki bulgulara göre,bağlantı (linkage) yöntemleri 'Zincirleme Etkisi' (Chaining) nedeniyle veriyi uzayan zincirler gibi kümelemiştir ve dengesiz bir yapı oluşturmuştur. Ward metodu ise küme içi varyansı (intra-cluster variance) minimize etme prensibiyle çalıştığı için en kompakt ve biyolojik olarak en anlamlı (dengeli) grupları oluşturduğu için seçilmiştir."



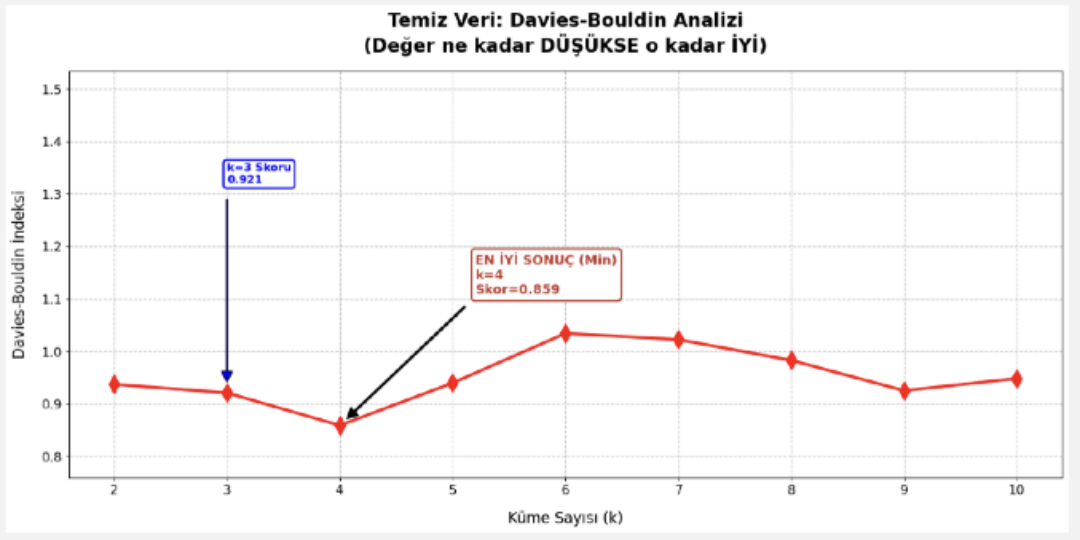
Şekil 6.10: K-Means kümeleme karşılaştırması

Yorum: K-Means sonuçları gösterilmektedir. Yöntem hızlıdır fakat aykırı değerlere ve başlangıç merkezlerine duyarlıdır.Farklı küme sayıları incelendiğinde, k=2 için silhouette skorunun en yüksek olduğu ve kümelerin daha net ayrıştığı görülmektedir. Küme sayısı arttıkça (özellikle k=6), aşırı bölünme (over-segmentation) oluşmakta ve küme kalitesi belirgin biçimde düşmektedir.



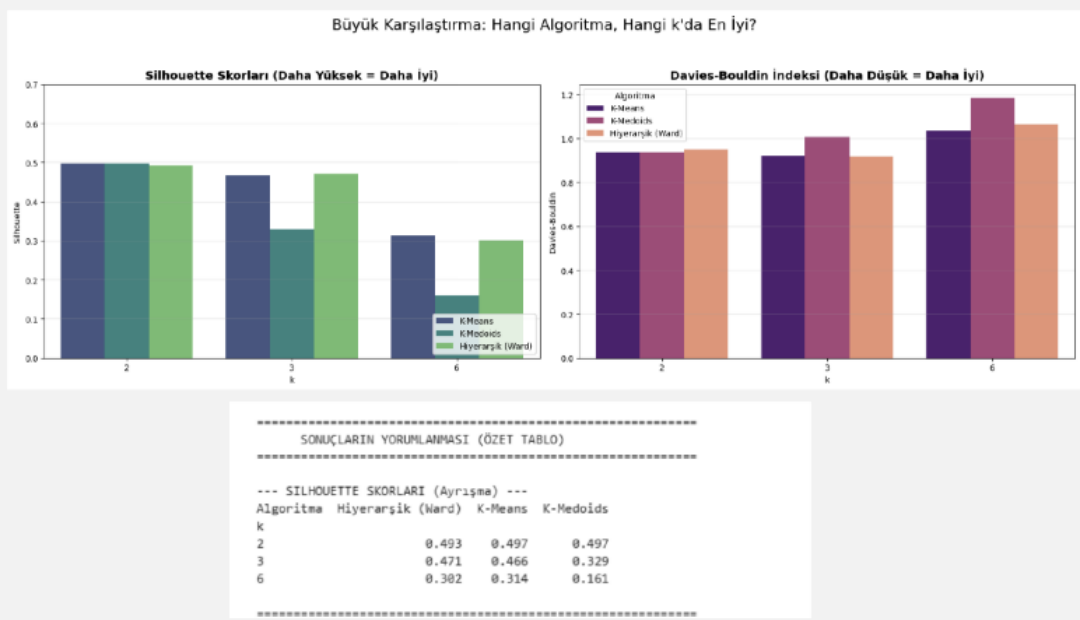
Şekil 6.11: K-Medoids kümeleme karşılaştırması

Yorum: K-Medoids/PAM sonuçları gösterilmektedir. Medoid'ler gerçek gözlemler olduğu için aykırı değerlere daha dayanıklıdır ve yorumlanabilirlik artar.K-medoids algoritmasında, k=2 için elde edilen silhouette skoru en yüksek değeri göstermektedir. Küme sayısı arttıkça skorların düşmesi, verinin doğal yapısının daha az sayıda küme ile daha iyi temsil edildiğini göstermektedir. Medoidlerin gerçek gözlemler olması yöntemin yorumlanabilirliğini artırmaktadır.



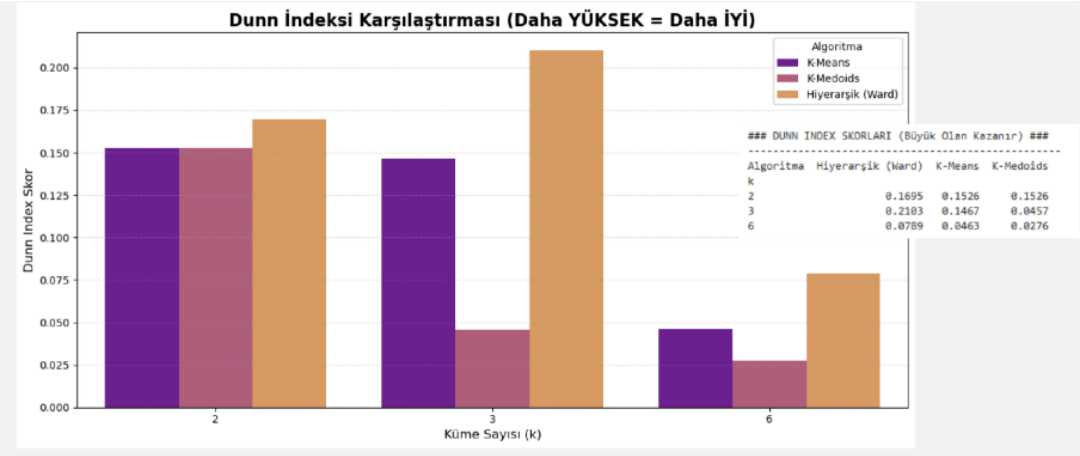
Şekil 6.12: Cluster Validation

Yorum: Davies–Bouldin indeksine göre en düşük değer k=4 için elde edilmiştir. Bu durum, küme içi benzerliğin yüksek ve kümeler arası ayrışmanın daha iyi olduğunu göstermektedir. Ancak sonuçlar diğer doğrulama ölçütleriyle birlikte değerlendirilmelidir..



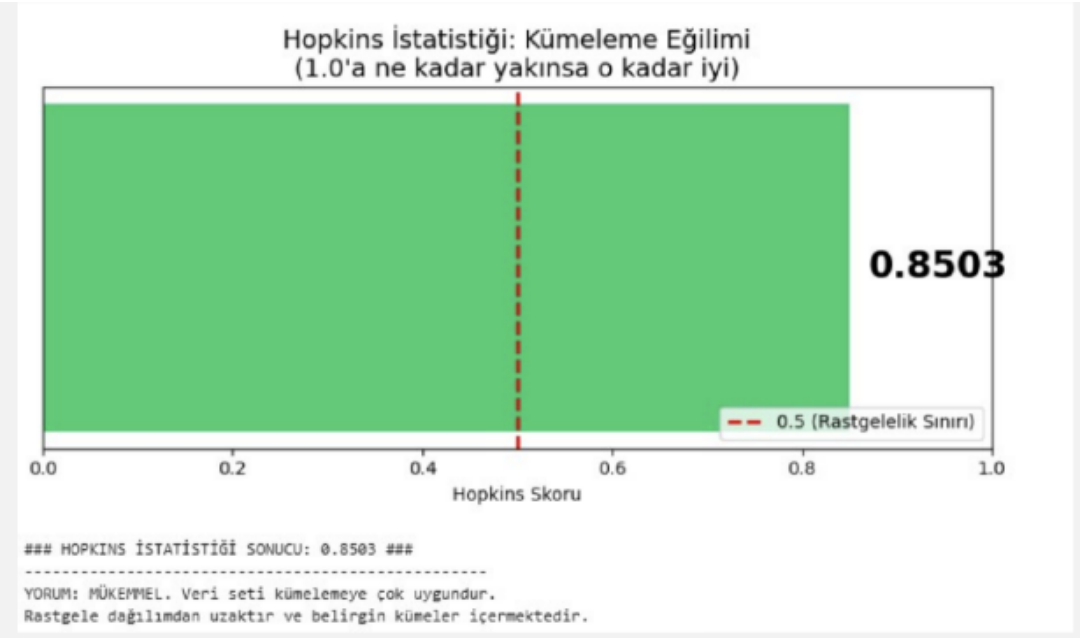
Şekil 6.13: Silhouette ve Davies-Bouldin karşılaştırması

Yorum: Silhouette, küme içi yakınlık ve kümeler arası ayrışmayı birlikte ölçer. Ortalama silhouette yüksekse kümeler iyi ayrışmış ve kompakt demektir.Silhouette skorları düşük k değerlerinde daha yüksek ayrışma kalitesine işaret ederken, Davies–Bouldin indeksi de benzer şekilde daha az küme sayısında daha iyi sonuçlar vermektedir. Her iki metrik birlikte değerlendirildiğinde, düşük k değerleri daha uygun görünmektedir.



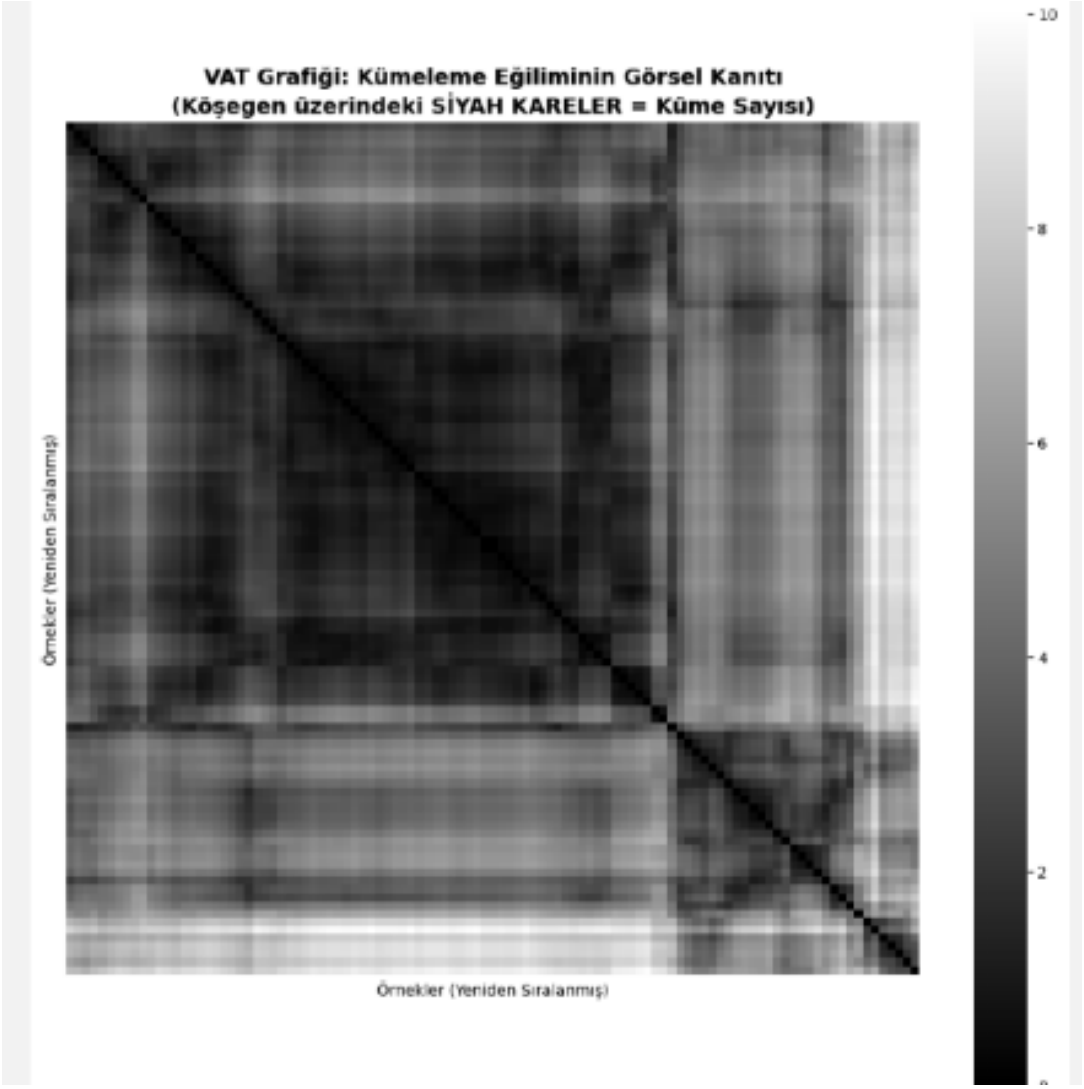
Şekil 6.14: Dunn Index karşılaştırması

Yorum: Dunn indeksi ayrışma/kompaktlığı birlikte ölçer. Daha büyük Dunn değeri daha iyi küme yapısına işaret eder.Dunn indeksine göre en yüksek skorlar hiyerarşik (Ward) yöntemi için elde edilmiştir. Küme sayısı arttıkça Dunn skorlarının düşmesi, kümeler arası ayrışmanın zayıfladığını ve aşırı bölünmenin oluştuğunu göstermektedir.



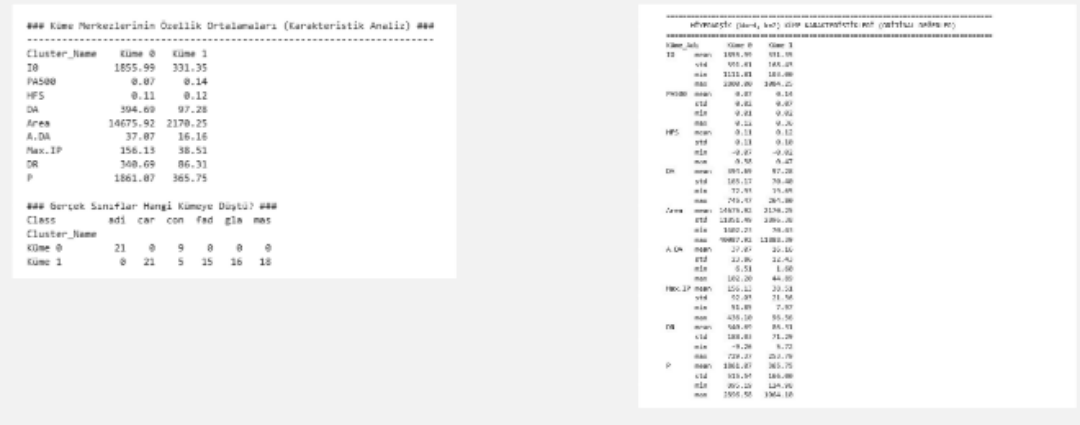
Şekil 6.15: Hopkins istatistiği

Yorum: Hopkins istatistiği verinin kümelenebilirliğini test eder. 0.5'e yakın değerler rastgeleliği, daha yüksek değerler kümelenme eğilimini destekler.Hesaplanan Hopkins skoru (0.85), veri setinin rastgele dağılmadığını ve yüksek düzeyde kümelenme eğilimi gösterdiğini ortaya koymaktadır. Bu sonuç, kümeleme analizlerinin uygulanması için verinin uygun olduğunu doğrulamaktadır.



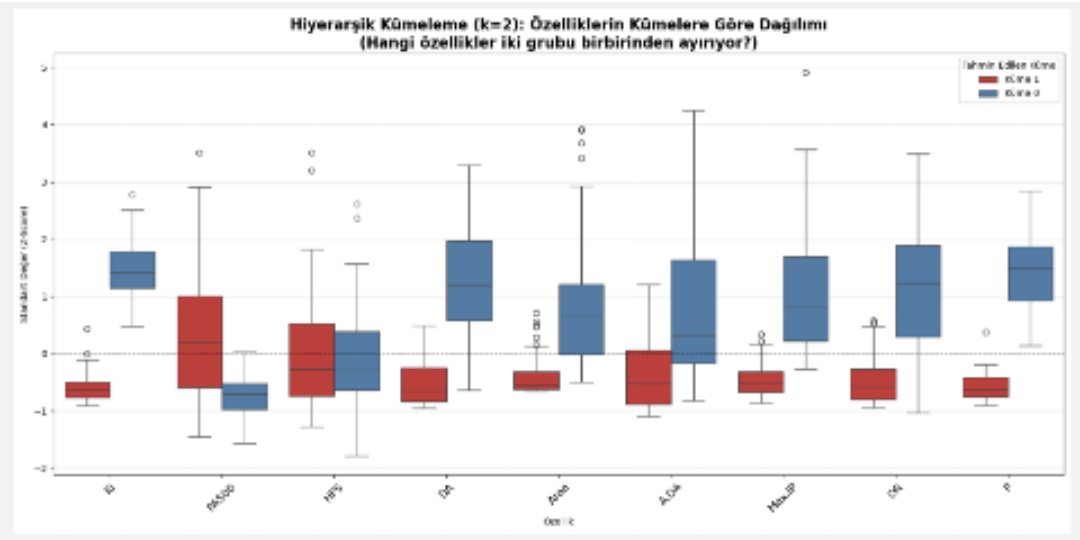
Şekil 6.16: VAT grafiği

Yorum: VAT grafiği uzaklık matrisindeki blok yapıyı görselleştirir. Belirgin bloklar doğal küme sayısı hakkında sezgisel ipucu verir.VAT grafiğinde köşegen boyunca gözlenen iki belirgin ana blok, verinin doğal olarak iki temel kümeye ayrılma eğiliminde olduğunu göstermektedir. Bu bulgu, Elbow yöntemiyle elde edilen k=2 sonucunu görsel olarak desteklemektedir.



Şekil 6.17: Hiyerarşik kümeleme tanımlayıcı istatistikleri

Yorum: Özet istatistikler değişkenlerin yayılımını ve ölçek farklarını gösterir. Ölçek farklılıkları standardizasyon ihtiyacını; uç değerler ise aykırı gözlem riskini işaret eder.Tanımlayıcı istatistikler incelendiğinde, kümeler arasında özellikle DA, Area, Max.IP ve P değişkenlerinde belirgin farklar olduğu görülmektedir. Bu durum, elde edilen kümelerin anlamlı ve ayırt edilebilir profiller oluşturduğunu göstermektedir.



Şekil 6.18: Kümelere göre boxplotlar

Yorum: Kutu grafikleri aykırı değerleri ve gruplar arası dağılım farklılıklarını gösterir. Hiyerarşik kümeleme (Ward yöntemi) sonucunda elde edilen boxplotlar, kümeler arasındaki ayrımın özellikle DA, Area, Max.IP, DR ve P değişkenlerinde belirgin olduğunu göstermektedir. Bu değişkenlerde medyan değerlerin ve dağılımların farklılaşması, hiyerarşik kümelemenin veriyi anlamlı ve yorumlanabilir iki ana grup altında başarılı bir şekilde ayırdığını ortaya koymaktadır.

Bölüm 7

Genel Değerlendirme ve Sonuç

Bu çalışmada, meme dokusu örneklerine ait elektriksel empedans ölçümleri kullanılarak denetimsiz öğrenme (unsupervised learning) yaklaşımları kapsamında kümeleme analizi gerçekleştirilmiştir. Veri setinin yapısal özelliklerini ortaya koymak ve doğal kümelenme eğilimini değerlendirmek amacıyla K-Means, K-Medoids (PAM) ve Hiyerarşik Kümeleme algoritmaları sistematik olarak karşılaştırılmıştır.

Analiz sürecinin ilk aşamasında uygulanan Hopkins istatistiği ve VAT grafiği, veri setinin rastgele bir yapıdan uzak olduğunu ve güçlü bir kümelenme eğilimi sergilediğini göstermiştir. Bu bulgu, kümeleme algoritmalarının uygulanabilirliğini desteklemiş ve sonraki adımlar için sağlam bir temel oluşturmuştur. Küme sayısının belirlenmesinde Elbow yöntemi, Silhouette genişliği ve Dunn indeksi birlikte değerlendirilmiş; bu ölçütlerin tamamı en tutarlı ve dengeli sonuçların $k = 2$ için elde edildiğini göstermiştir.

Yöntemler arası karşılaştırma sonucunda, Hiyerarşik Kümeleme kapsamında kullanılan Ward bağlantı yönteminin diğer alternatiflere kıyasla daha kompakt, daha dengeli ve yorumlanabilir kümeler ürettiği gözlemlenmiştir. Ward metodunun, özellikle tekli bağlantı (single linkage) yönteminde karşılaşılan zincirleme etkisini ortadan kaldırarak küme içi varyansı minimize etmesi, bu yöntemin tercih edilmesinde belirleyici olmuştur. Ayrıca, VAT grafiğinde gözlenen iki belirgin blok yapısının ve kümeleme sonuçlarının tıbbi bağlamdaki *hasta/sağlıklı* ayrımıyla örtüşmesi, elde edilen bulguların anlamlılığını güçlendirmiştir.

Sonuç olarak, bu çalışma hem farklı kümeleme yöntemlerinin karşılaştırmalı bir analizini sunmuş hem de içsel doğrulama ölçütlerinin kümeleme kalitesinin değerlendirilmesindeki kritik rolünü ortaya koymuştur. Elde edilen bulgular, Hiyerarşik Kümeleme (Ward Metodu) ve $k = 2$ seçiminin veri seti için en uygun yaklaşım olduğunu göstermektedir. Bu yönüyle çalışma, denetimsiz öğrenme yöntemlerinin tıbbi veri analizlerinde etkin bir şekilde kullanılabileceğini ortaya koymakta ve ileride yapılacak daha kapsamlı analizler için referans niteliği taşımaktadır.

Kaynakça

[1] Alboukadel Kassambara (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. STHDA (Statistical Tools for High-Throughput Data Analysis). Available at: `https://www.sthda.com`

[2] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.

[3] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.

[4] Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1), 95–104.

[5] Breast Tissue Data Set. Available at: `https://www.kaggle.com/datasets/ukveteran/breast-tissue-data-set`