

1. Əsas Anlayışlar

1.1 Clustering Nədir?

Clustering — verilənlərdə təbii qruplaşmaları aşkar etmək üçün istifadə olunan **unsupervised learning** metodudur. Bu, verilənlərin daxili quruluşunu başa düşmək və onları oxşarlıq kriteriyalarına görə qruplara ayırmaq məqsədini güdür.

Məsələn, marketinqdə müştəriləri davranışlarına görə qruplaşdıraraq fərqli seqmentlər yaratmaqla, hər seqment üçün spesifik strategiyalar qurmaq mümkündür. Bu, effektiv resurs istifadəsi və daha hədəfli kampaniyalar deməkdir.

Riyazi əsaslar:

Verilənlər toplusu

$$X = \{x_1, x_2, \dots, x_n\}$$

üçün məqsəd k klaster tapmaqdır ki, burada:

- **Klasterlər biri-birindən ayrı olsun:**

$$C_i \cap C_j = \emptyset, \quad i \neq j$$

- **Hər nöqtə yalnız bir klasterdə olsun:**

$$\bigcup_{i=1}^k C_i = X$$

- **Klaster daxilindəki məsafə minimal, klasterlərarası məsafə maksimal olsun:**

$$\min \sum_{i=1}^k \sum_{x \in C_i} d(x, \mu_i)$$

Burada μ_i — klasterin mərkəzi, d isə məsafə funksiyasıdır.

1.2 Niyə Hierarchical Clustering?

Hierarchical clustering-in ən böyük üstünlüyü, klaster sayını əvvəlcədən bilmədən, verilənlərin təbii qruplaşmasını və onların arasındakı əlaqələri **hiyerarşik strukturda** göstərməsidir. Bu, verilənlərin təhlilində çox zəngin informasiya verir.

Misal üçün, müştəri segmentasiyasında, müştərilər ilk mərhələdə əsas qruplara, sonra isə həmin qruplar daxilində alt qruplara bölünə bilər. Bu struktur dendrogramda aydın şəkildə vizuallaşdırılır.

Bu metod həmçinin çox səviyyəli araşdırma üçün idealdır, çünki müxtəlif kəsirlərdə fərqli sayda klasterlər yaratmaq mümkündür və verilənlərin strukturunu dərinlən analiz etməyə imkan verir.

Lakin, **hesablama mürəkkəbliyi** yüksəkdir, böyük verilənlərdə çox vaxt tələb edir və həssas parametrlərlə işləyir. Buna görə də, kiçik və orta ölçülü verilənlər üçün daha əlverişlidir.

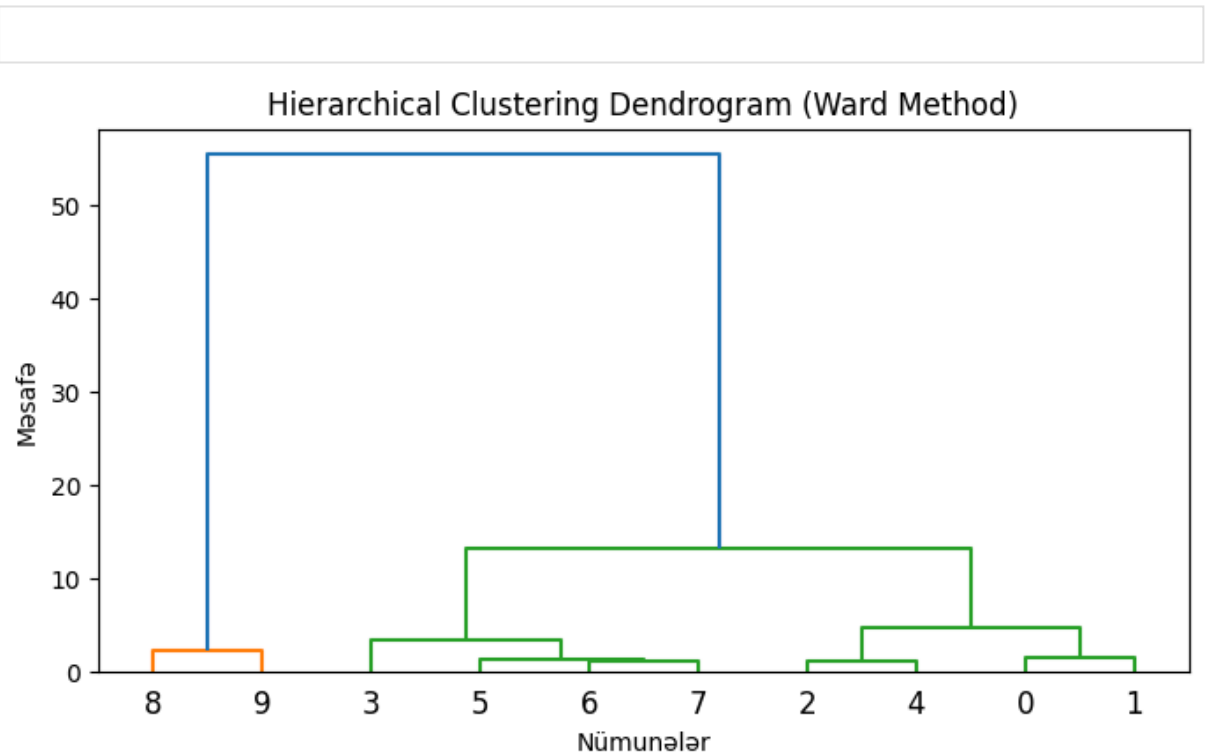
1.3 Hansı Problemlər Üçün Uyğundur?

Hierarchical clustering əsasən:

- **Kiçik və orta ölçülü verilənlər** üçün uyğundur, çünki böyük verilənlərdə hesablama xərci çox yüksəlir.
- **Verilənlərin təbii hiyerarşik quruluşunu öyrənmək** lazım olduqda — məsələn, genetik tədqiqatlarda filogenetik ağacların yaradılması.
- **Klaster sayı əvvəlcədən məlum olmadıqda**, dendrogramdan uyğun klaster sayını seçmək mümkün olur.
- **Müştəri segmentasiyası, sənəd qruplaşdırılması, anomaliya aşkarlanması kimi müxtəlif sahələrdə** istifadə olunur.

 **Qeyd:** Böyük dataset-lər üçün əvvəlcə **PCA** ilə ölçü azaldıb, sonra hierarşik clustering tətbiq etmək çox vaxt daha səmərəlidir.

In [2]:



2. Hierarchical Clustering Növləri

2.1 Agglomerative Clustering (Yuxarıdan Aşağıya)

Bu metodda:

- **Başlanğıc:** Hər nöqtə öz klasteridir.
- **Addım-addım:** Ən yaxın klasterlər tapılır və birləşdirilir.
- Proses tək klaster qalana qədər davam edir.

Məsafənin təyin olunması və linkage metodu çox önəmlidir.

Məsafəni düzgün seçmək klasterlərin keyfiyyətinə birbaşa təsir göstərir.

Bu metod **bottom-up (aşağıdan yuxarıya)** yanaşmadır.

2.2 Divisive Clustering (Aşağıdan Yuxarıya)

Bu metodda:

- **Başlanğıc:** Bütün nöqtələr bir klasterdədir.
- Klasterlər addım-addım bölünür.
- Proses istənilən klaster sayına çatana qədər davam edir.

Bu üsul **top-down (yuxarıdan aşağıya)** yanaşmadır və adətən çox hesablama tələb etdiyi üçün daha az yayılmışdır.

3. Verilənlərin Hazırlanması

3.1 Missing Data ?

- Əskik verilər clustering keyfiyyətini azalda bilər.
- Onların ya **silinməsi**, ya da **müvafiq dəyərlərlə doldurulması (imputation)** tövsiyə olunur.
- İmputasiya üçün ortalama, median, ya da model əsaslı metodlar istifadə oluna bilər.

3.2 Ölçüləndirmə (Scaling)

- Clustering məsafə əsaslı olduğundan, xüsusiyyətlərin fərqli ölçülərdə olması nəticələri çarpıdırır.
- **StandardScaler** (ortalama 0, standart sapma 1), **MinMaxScaler** kimi üsullarla verilənlər eyni ölçüyə gətirilməlidir.

3.3 Ölçü Seçimi və Reduksiya

- Yüksək ölçülü verilənlərdə (məsələn, yüzlərlə xüsusiyyət) clustering effektiv olmaya bilər.
- Bunun üçün **PCA** və ya **t-SNE** kimi ölçü azaldıcı metodlar tətbiq edilir.

4. Linkage Metodları

Linkage metodu iki klaster arasındaki məsafəni ölçür və bu, klasterlərin necə birləşdiriləcəyini müəyyən edir.

4.1 Single Linkage

- İki klaster arasındaki **ən yaxın nöqtələrin məsafəsini** götürür:

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

- Zəncir effekti** yaradır; uzun, sarkıq klasterlər əmələ gələ bilər.
- Gürültüyə və outlier-lərə qarşı həssasdır.

4.2 Complete Linkage

- İki klaster arasındaki **ən uzaq nöqtələrin məsafəsini** götürür:

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y)$$

- Daha kompakt, bərk klasterlər yaradır.
- Outlier-lərə qarşı Single Linkage-dən daha davamlıdır.

4.3 Average Linkage

- Bütün nöqtələr arasında məsafələrin **ortalamasını** hesablayır:

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

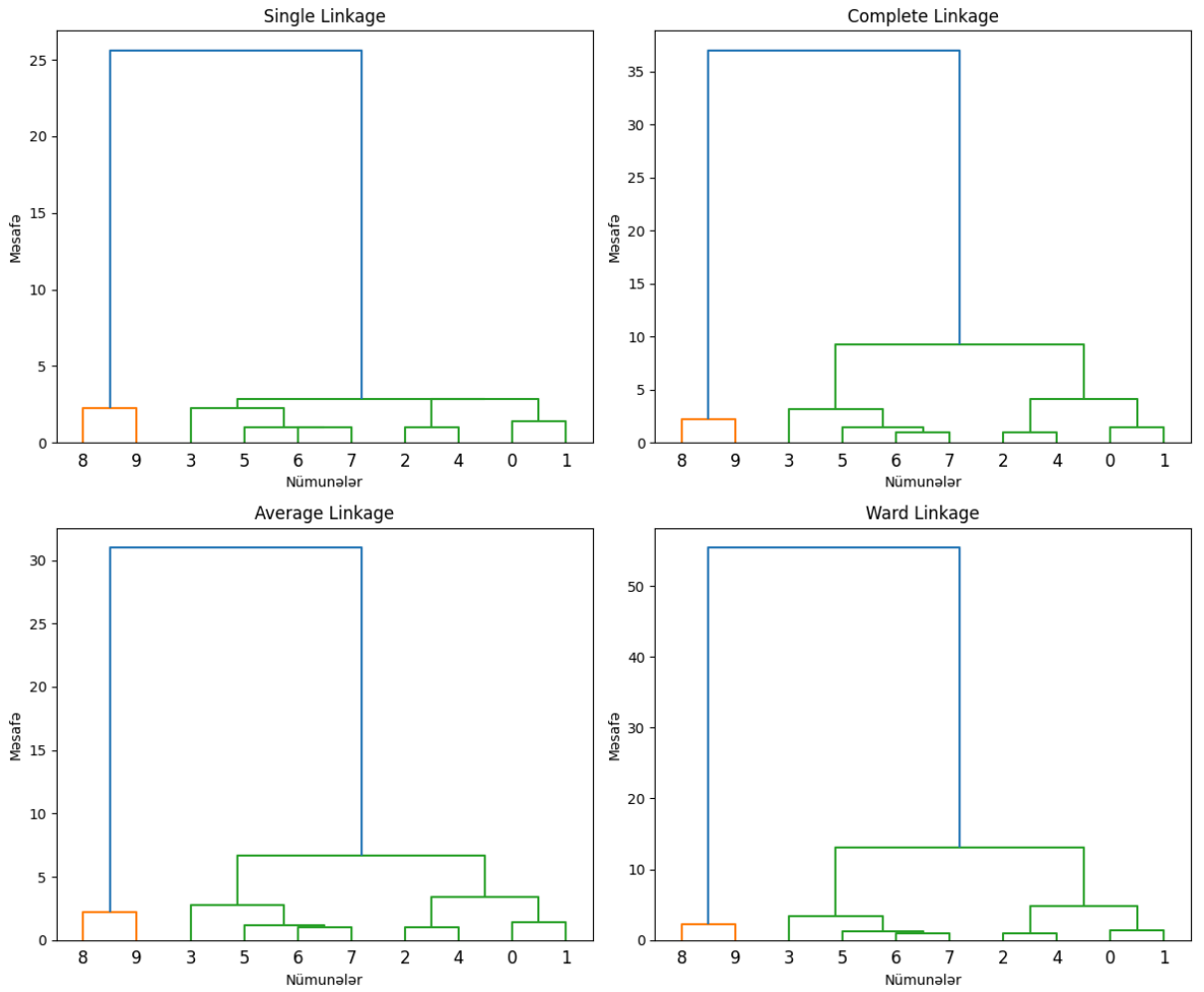
- Zəncir və kompaktlıq arasında balans yaradır.

4.4 Ward's Method

- Hər bir birləşdirmədə **klaster daxilində varyans artımını minimal edir**.
- Formul:

$$\Delta = \frac{|C_i||C_j|}{|C_i| + |C_j|} \|\bar{x}_i - \bar{x}_j\|^2$$

- Çox vaxt ən yaxşı nəticə verir, homogen və yuvarlaq klasterlər yaradır.



5. Dendrogram 🌳

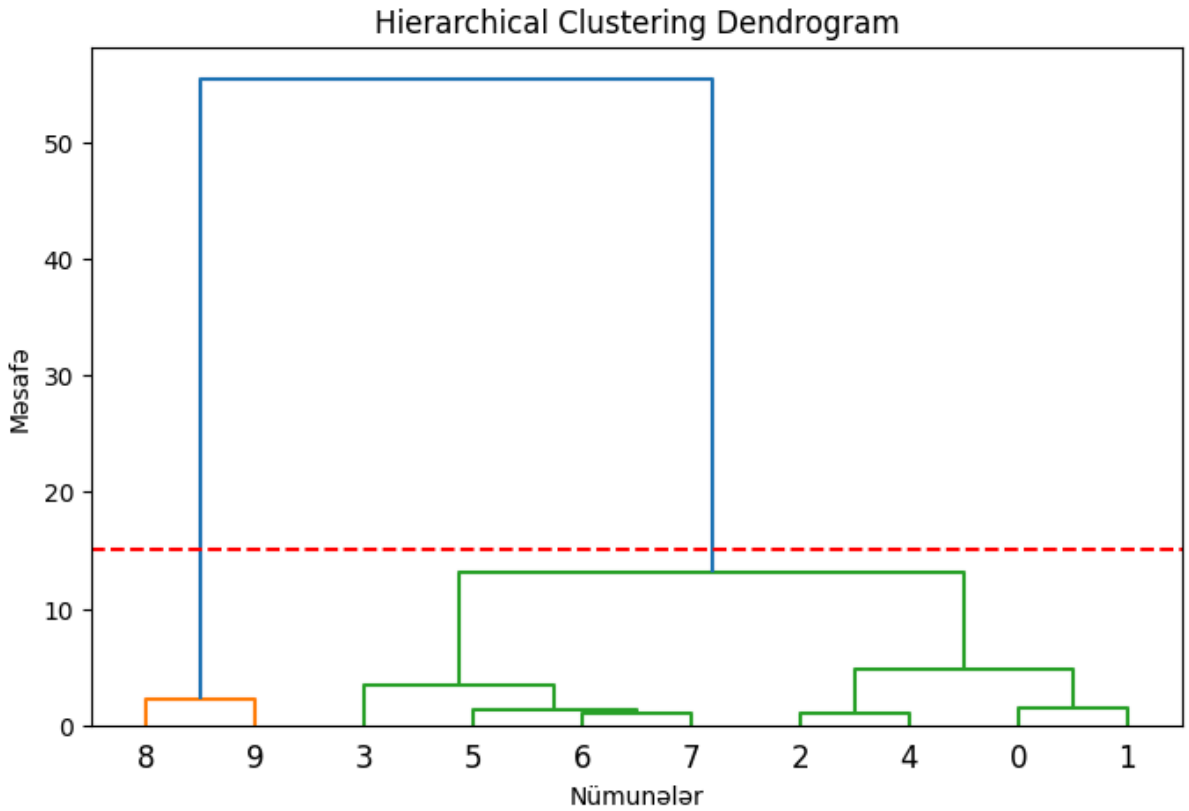
5.1 Dendrogram Nədir? 📊

- **Hierarchical clustering** nəticəsində yaranan **ağac şəklində vizuallaşdırma**dır.
- **Vertikal ox** – klasterlərin birləşmə məsafəsini göstərir.
- **Horizontal ox** – fərdi nöqtələri və ya müşahidələri göstərir.

5.2 Dendrogram ilə Klaster Sayının Müəyyənləşdirilməsi ✂️

- Dendrogramda **böyük məsafə fərqlərinin** olduğu yerlərdən üfüqi kəsik çəkilir.
- Kəsikdən **aşağıda qalan qruplar** klasterləri təşkil edir.
- Bu üsul, **optimal klaster sayını təyin etməyə** kömək edir.

In [8]:



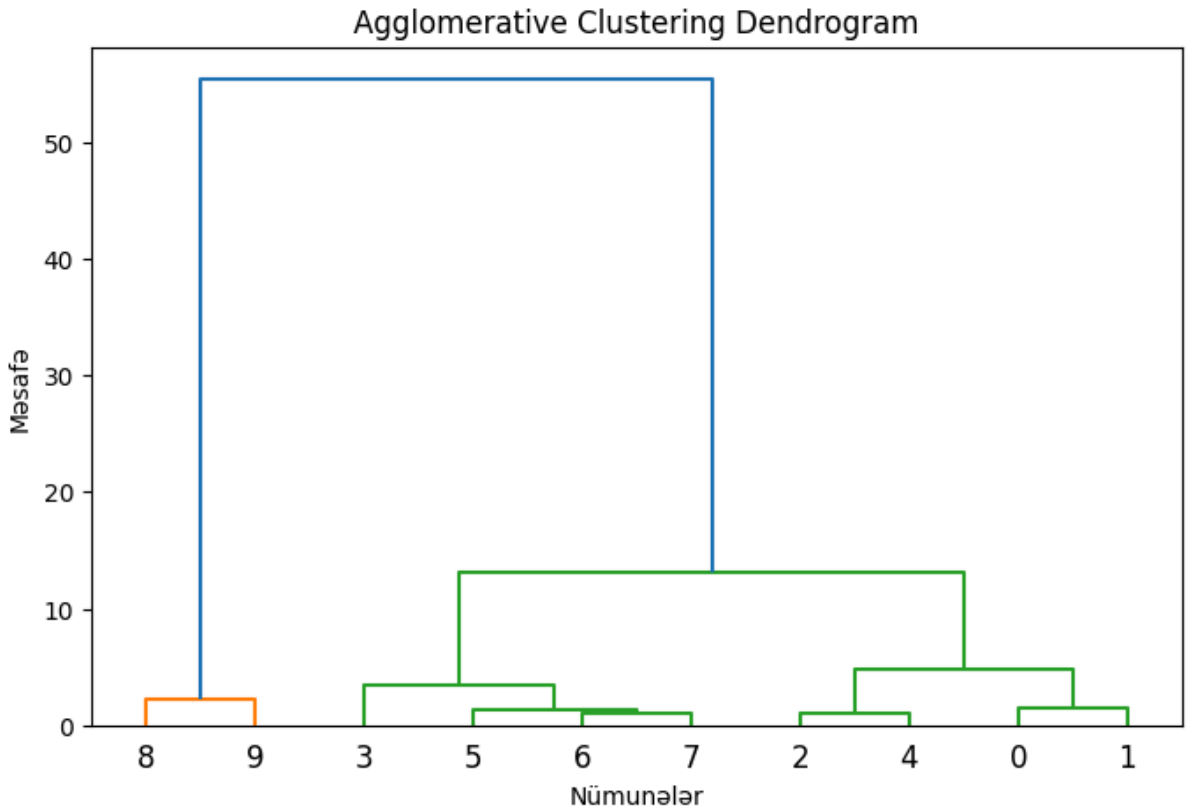
Klaster etiketləri: [2 2 2 2 2 2 2 2 1 1]

6. Alqoritmin İşləmə Prosesi

6.1 Agglomerative Clustering Addımları

- **Başlanğıc:** Hər nöqtə **ayrılıqda** bir klaster kimi götürülür.
- **Məsafə hesablanması:** Klasterlər arasındakı məsafələr seçilmiş **linkage metoduna** görə ölçülür.
- **Birləşdirmə:** Ən **yaxın** məsafəyə sahib olan iki klaster birləşdirilir.
- **Məsafələrin yenilənməsi:** Yeni yaranmış klasterin digər klasterlərlə məsafəsi yenidən hesablanır.
- **Təkrarlanma:** Bu proses **yalnız bir klaster qalana** qədər davam edir.
- **Nəticə:** Bütün proses boyu birləşmə ardıcılığı dendrogram vasitəsilə göstərilə bilər.






In [9]:



6.2 Divisive Clustering Addımları

- **Başlanğıc:** Bütün nöqtələr **tək bir klasterdə** yerləşir.
- **Bölünmə:** Klaster, ən çox **heterojen** (fərqli) hissələrinə bölünür.
- **Addım-addım təkrar:** Hər mərhələdə seçilmiş klaster yenidən kiçik klasterlərə ayrılır.
- **Dayanma şərti:** İstənilən **klaster sayına** çatdıqda və ya daha bölünmənin mənalı olmadığı halda proses dayandırılır.

7. Tətbiqlər

- **Müştəri segmentasiyası**  : Müştəriləri **oxşar davranış və xüsusiyyətlərinə** görə qruplaşdırmaq, hədəfli marketing strategiyaları qurmaq üçün istifadə olunur.
- **Genetik tədqiqat**  : Növlərin və ya fərdlərin **genetik yaxınlıq dərəcəsinə** görə qruplaşdırılması, təkamül əlaqələrinin araşdırılması üçün tətbiq edilir.
- **Görüntü emalı**  : Obyektlərin formalarına, rənglərinə və ya teksturalarına görə qruplaşdırılması, şəkil tanıma və obyekt aşkarlanmasında istifadə olunur.
- **Anomaliya aşkarlanması**  : Normadan fərqli olan **nadir nümunələrin** müəyyən edilməsi, təhlükəsizlik və fırıldaqçılıq aşkarlanmasında geniş istifadə olunur.
- **Mətn sənədlərinin qruplaşdırılması**  : Oxşar mövzulu sənədlərin və ya xəbərlərin **qruplara ayrılması**, axtarış sistemlərinin və tövsiyə motorlarının təkmilləşdirilməsində tətbiq edilir.

8. Performans və Qiymətləndirmə

8.1 Klaster Keyfiyyət Ölçüləri

Silhouette Skoru

- Hər nöqtənin öz klasterinə nə qədər uyğun olduğunu göstərir.
- **Hesablanması:**

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

Burada:

- **a(i):** Nöqtənin öz klasteri daxilində ortalama məsafəsi
- **b(i):** Ən yaxın digər klasterdəki ortalama məsafə
- Skor **1**-ə yaxın olduqda, klasterin keyfiyyəti yüksəkdir.
- **0** ətrafında isə nöqtə iki klaster arasında yerləşir.
- **Mənfi** skor isə nöqtənin yanlış klasterə düşdüyünü göstərir.

Davies-Bouldin İndeksi və Dunn İndeksi

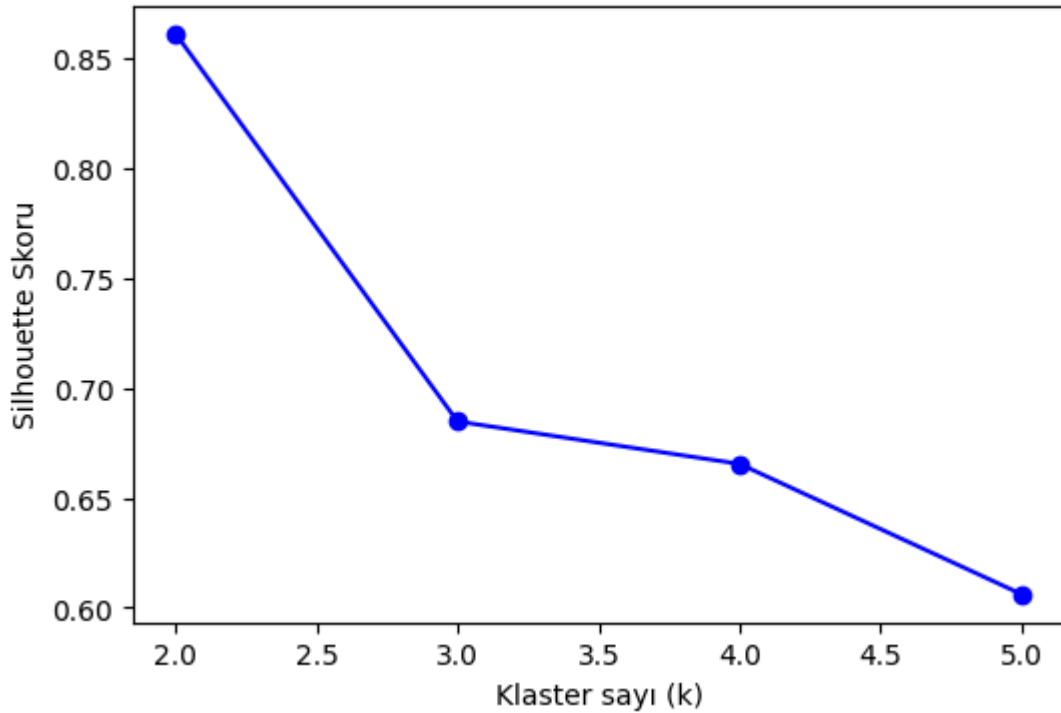
- **Davies-Bouldin (DB) İndeksi:** Klasterlərin sıxlığı və bir-birindən ayrılma dərəcəsini ölçür.
 - **Aşağı DB** indeksi daha yaxşı nəticə deməkdir.
- **Dunn İndeksi:** Klasterlər arasındakı minimal məsafənin klaster daxilindəki maksimal məsafəyə nisbətini ölçür.
 - **Yüksək Dunn** indeksi daha yaxşı qruplaşma göstərir.

8.2 Klaster Sayının Müəyyənləşdirilməsi

- **Dendrogram** üzərində böyük məsafə fərqi olduğu yer seçilir.
- Müxtəlif klaster sayları üçün **silhouette skorları** analiz edilir.
- **Elbow metodu:** Total məsafə və ya varyans dəyişmələrində "dirsək" nöqtəsi axtarılır.

In [11]:

Silhouette Skoru ilə Optimal Klaster Sayının Müəyyənləşdirilməsi



In [12]:

```
Davies-Bouldin indeksi: 0.32851132700017094  
Dunn indeksi: 0.24253562503633297
```

9. Problemlər və Həllər ⚠️

9.1 Böyük Verilənlərdə Çətinliklər 📊

- **Hierarchical clustering**-in **yüksək hesablama xərci** mövcuddur, çünki bütün məsafə matrisinin hesablanması və saxlanması tələb olunur.
- Bu, $O(n^2)$ yaddaş və $O(n^3)$ hesablama mürəkkəbliyi ilə nəticələnə bilər.
- Böyük datasetlərdə bu çox vaxt aparır və yaddaş limitlərini aşır.
- **Həll yolları:**
 - Nümunə götürmə (**sampling**) üsulu ilə verilənlərin bir hissəsini istifadə etmək
 - Daha sürətli alternativ clustering metodlarından istifadə (məsələn, **K-Means**, **MiniBatch K-Means**)
 - Məsafə hesablamalarını optimallaşdıran **approximate nearest neighbor** üsulları

9.2 Noise və Outlier 🌀

- Verilənlərdəki **noise** və **outlier**-lər klasterlərin formalaşmasını poza bilər.
- Hierarchical clustering hər bir nöqtəni bir klaster kimi qəbul etdiyi üçün **səs-küyə həssasdır**.
- **Həll yolları:**
 - Əvvəlcədən **outlier deteksiyası** və təmizlənməsi (məsələn, Z-score, IQR metodu)
 - Gürültüyə qarşı davamlı olan metodlardan istifadə, məsələn **DBSCAN** və ya **HDBSCAN**

- Məsafə ölçüsünü dəyişmək (məsələn, **Manhattan** və ya **cosine distance**)

9.3 Yüksək Ölçülülük Problemi

- Ölçülərin çox olması (**high dimensionality**) məsafə hesablamalarını etibarsız edə bilər (distance concentration problemi).
- Bu halda nöqtələr arasındakı məsafə fərqləri azalır, nəticədə klasterlər aydın formalaşmır.
- **Həll yolları:**
 - **Ölçü azaldıcı metodlar** ilə əvvəlcə ölçünü azaltmaq:
 - **PCA (Principal Component Analysis)**
 - **t-SNE (t-distributed Stochastic Neighbor Embedding)**
 - **UMAP (Uniform Manifold Approximation and Projection)**
 - Ölçü azaldıldıqdan sonra hierarchical clustering tətbiq etmək

10. İrəli Səviyyə Mövzular

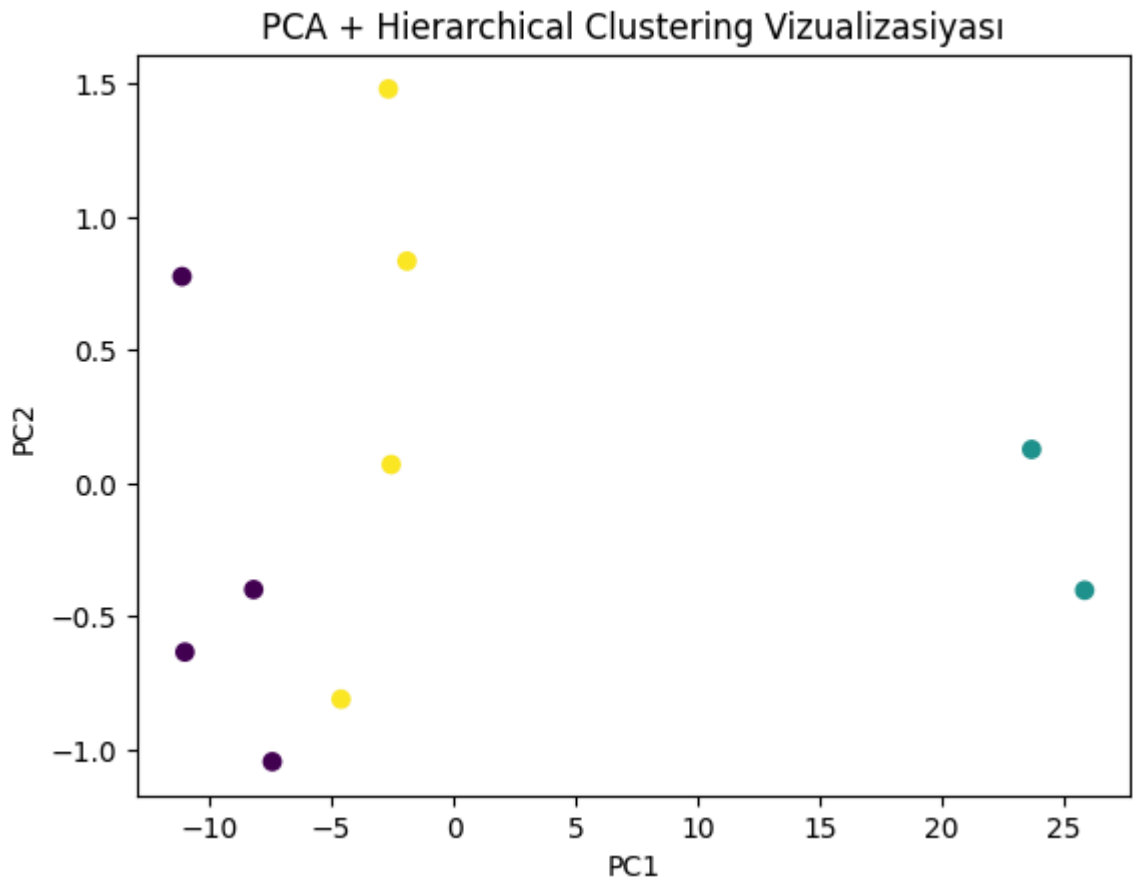
10.1 Ensemble Hierarchical Clustering

- **Ensemble clustering** — fərqli clustering nəticələrinin birləşdirilməsi ilə daha stabil və dəqiq nəticələr əldə etmək metodudur.
- Məqsəd — müxtəlif **linkage metodları** (single, complete, average) və ya fərqli məsafə ölçüləri ilə aparılmış hierarchical clustering nəticələrini birləşdirərək **sabitlik** və **davamlılıq** artırmaqdır.
- İcra addımları:
 1. Eyni dataset üzərində bir neçə fərqli hierarchical clustering modeli qurulur.
 2. Hər modeldən alınan klaster etiketləri **consensus matrix** və ya **co-association matrix** vasitəsilə birləşdirilir.
 3. Final klasterlər bu matris üzərində yenidən clustering tətbiq etməklə müəyyən edilir.
- Üstünlüklər:
 - **Sabit nəticələr** verir, çünki tək metodun təsadüfi səhvlərini azaldır.
 - Fərqli model güclərini birləşdirir.

10.2 Klaster Vizualizasiyası və Sonrası Analiz

- Hierarchical clustering nəticələrinin **PCA** və digər ölçü azaldıcı metodlarla (məsələn, **t-SNE**, **UMAP**) 2D və ya 3D müstəvidə təsviri nəticələrin interpretasiyasını asanlaşdırır.
- Klasterlər **rəng kodları** ilə göstərilir, hər klaster fərqli rəngdə olur.
- **Heatmap**-lər:
 - Dendrogram ilə birlikdə istifadə oluna bilər.
 - Klasterlərin xüsusiyyətləri üzrə ortalama dəyərlər vizual şəkildə göstərilir.
- Sonrası analiz addımları:
 - Klasterlərin statistik xülasəsini çıxarmaq.
 - Hər klasterdəki nümunələrin ortaq xüsusiyyətlərini tapmaq.
 - Potensial outlier-ləri müəyyənləşdirmək.

In [13]:



10.3 Digər ML Metodları ilə Kombinasiya 🚗

- Klaster nəticələri **supervised learning** (məsələn, classification və regression) modellərində **xüsusiyyət (feature)** kimi istifadə oluna bilər.
 - Məsələn: Müştəri segmentlərini tapmaq üçün hierarchical clustering, daha sonra bu segmentləri satış proqnoz modelinə daxil etmək.
- Ölçü azaldılması və clustering-in kombinasiyası:
 - **PCA + Hierarchical Clustering** tez-tez yüksək performans verir.
 - Bu yanaşma həm sürəti artırır, həm də overfitting riskini azaldır.
- Hibrid metodlar:
 - Əvvəlcə **K-Means** və ya **DBSCAN** ilə ilkin klasterlər yaradılır.
 - Daha sonra həmin klasterlər hierarchical yanaşma ilə daha incə bölünür.