KÜMELEME VE KURAL TABANLI ALGORİTMALAR

DOÇ. DR. MUSTAFA AGÂH TEKİNDAL
PROF. DR. FERHAN ELMALI
ÖĞR.GÖR. DR. BERHAN ÇOBAN

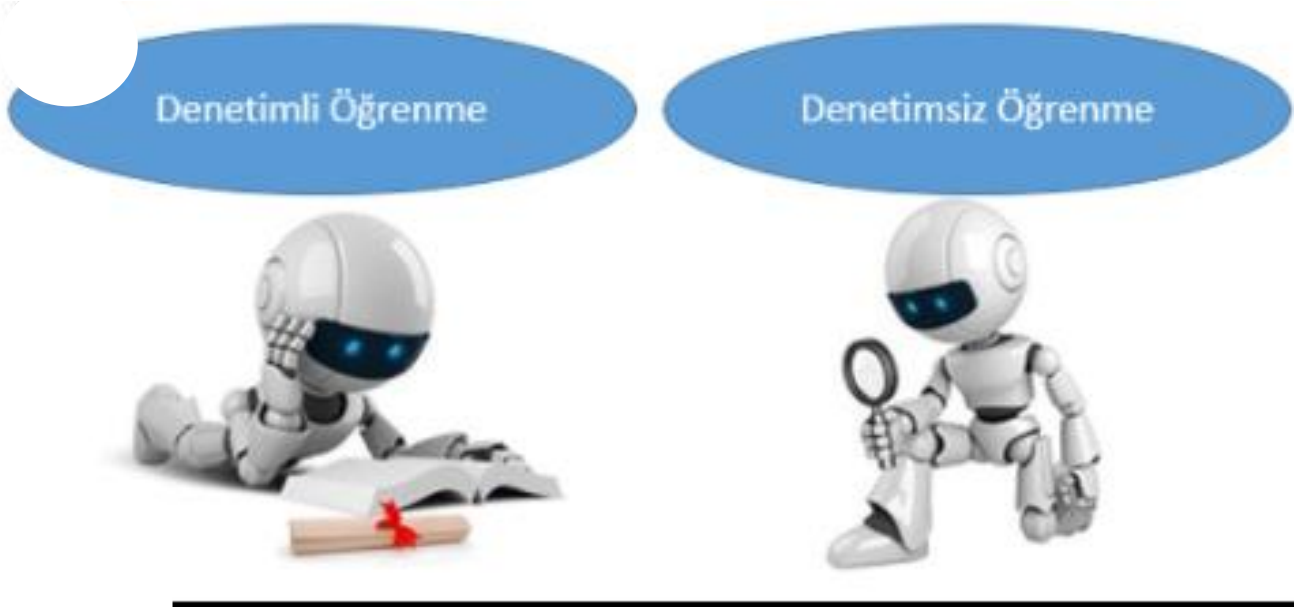
İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

TÜBİTAK

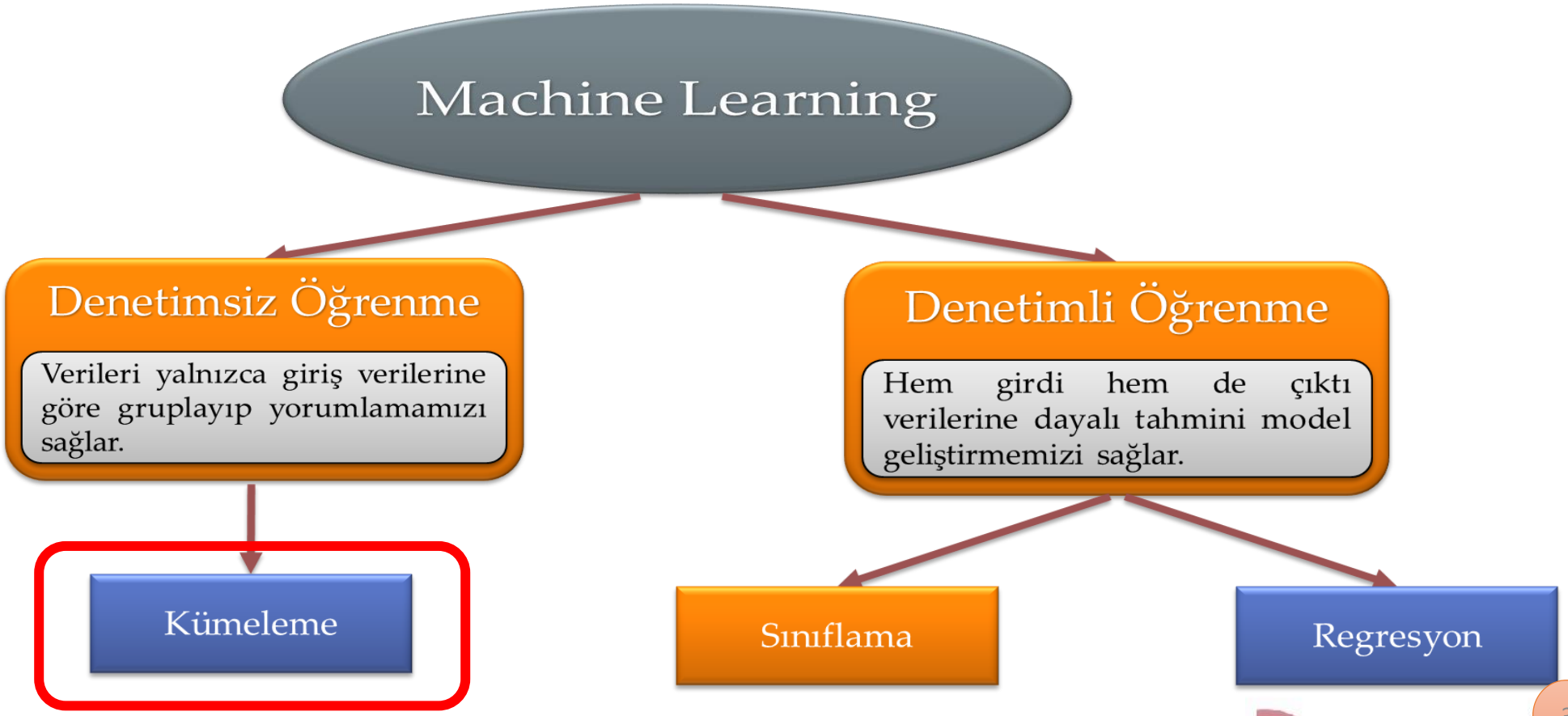
4005 TÜBİTAK

GİRİŞ

- Kümeleme, makine öğrenmesi konseptlerinden biri olan Denetimsiz Öğrenme için önemli bir kavramdır.
- Kümeleme algoritmaları basitçe veri kümesindeki elemanları kendi arasında gruplamaya çalışır.
- Burada kaç grup olacağı bizim inisiyatifimizde olan bir bilgi de olabilir veya en uygun küme sayısını algoritmanın kendisi de belirleyebilir.
- Kümeleme kavramının derinlerine inmeden önce denetimli ve denetimsiz öğrenme yöntemlerinden biraz bahsetmekte fayda var.



ŞEKİL 1. MAKİNE ÖĞRENMESİ METOTLARI

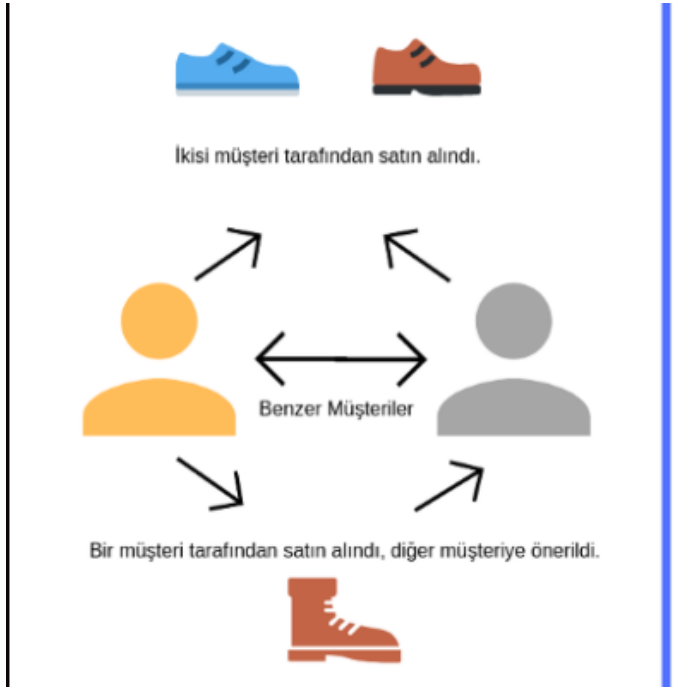


DENETİMSİZ ÖĞRENME NEDİR?

- Bir veri kümesinin temel yapısını nasıl buluyorsunuz?
- Bunu nasıl özetleyip en yararlı şekilde gruplandırıyorsunuz?
- Verileri sıkıştırılmış bir biçimde nasıl gösterirsiniz?
- Bunlar, “etiketsiz” olarak adlandırılan denetimsiz öğrenmenin hedefleridir, çünkü etiketsiz verilerle başlarsınız.
- Yani belli girdi değişkenleri kullanarak bu değişkenleri en iyi şekilde sınıflandırmaya çalışırız.

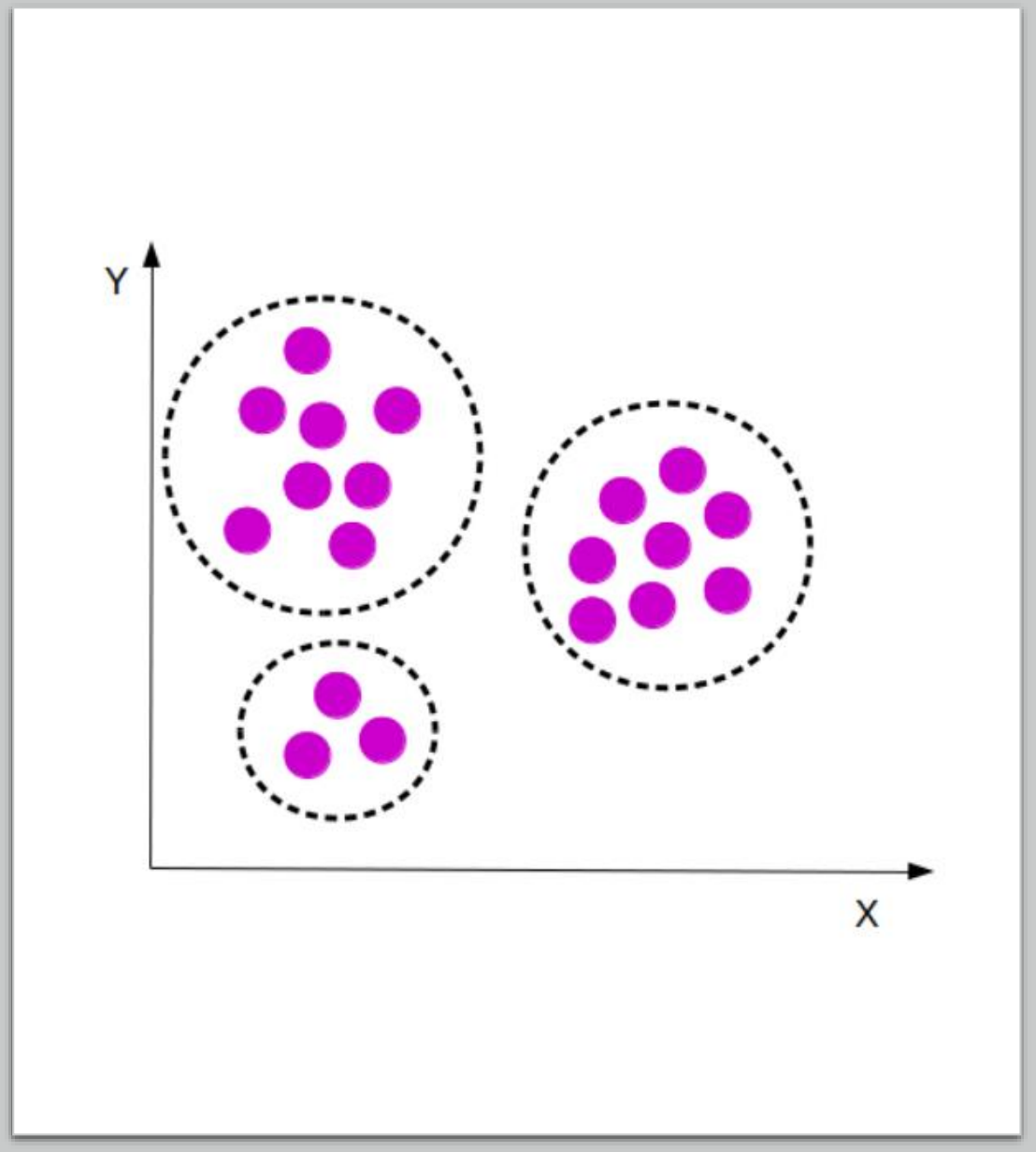
Bazı Denetimsiz Öğrenme Uygulama Örnekleri:

- YouTube Video Öneri Sistemi
- Facebook Arkadaş, Grup veya Sayfa Öneri Sistemi.
- E-ticaret Web Sitesi ile ilgili Ürünler Öneri Sistemi.



Kümeleme

- Genel olarak, kategorize edilmemiş verilerden oluşan bir veri setinde bir yapı veya model bulma ile ilgilenir.
- Kümeleme algoritmaları verilerinizi işler ve verilerde varsa doğal kümeleri (grupları) bulur.
- Ayrıca algoritmalarınızın kaç kümeyi tanımlaması gerektiğini de değiştirebilirsiniz.
- Bu grupların ayrıntı düzeyini ayarlamanıza olanak tanır.

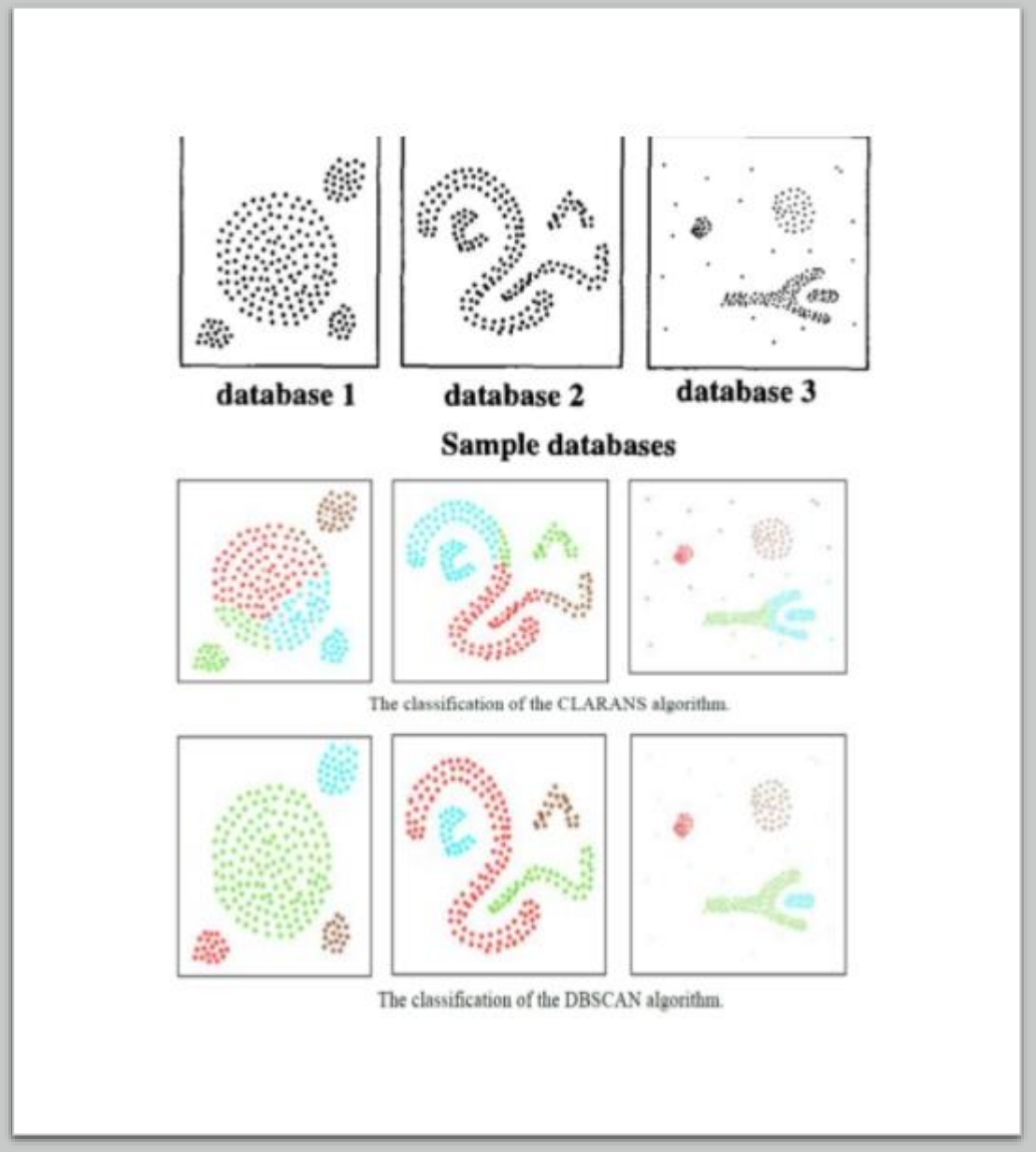


KÜMELEME ALGORİTMALARI

- Density Based Clustering (Yoğunluk Tabanlı Kümeleme)
- Fuzzy c-means Clustering (Bulanık c-ortalama Kümeleme)
- Hierarchical Clustering (Hiyerarşik kümeleme)
- K-means Clustering (K-ortalama Kümeleme)
- Random Forest Clustering (Rastgele Orman Kümelemesi)

Density Based Clustering (Yoğunluk Tabanlı Kümeleme)

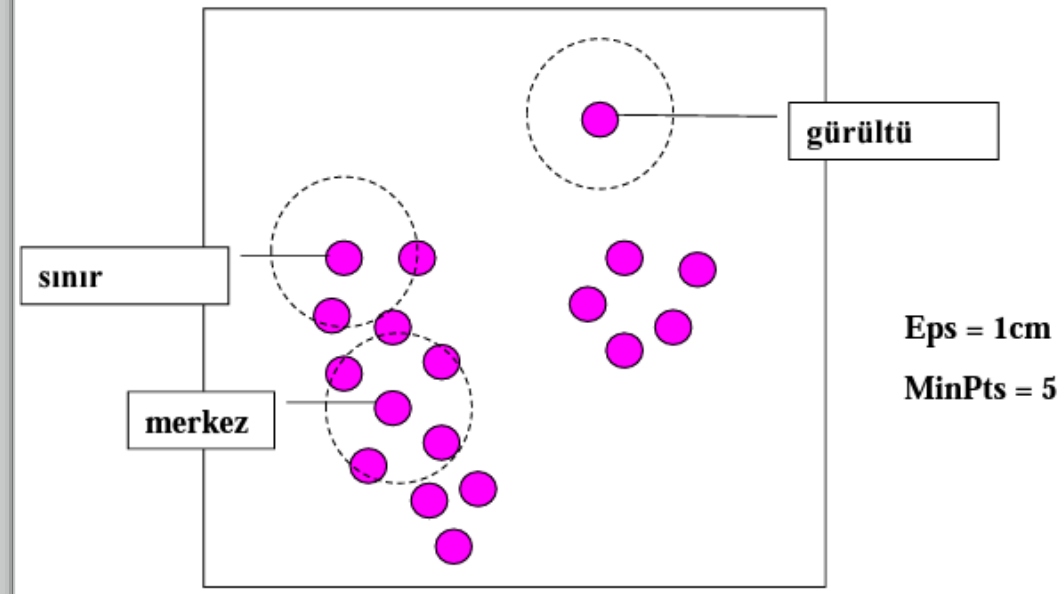
- Bu teknikte uzaklığa değil verilerin yoğunluğuna göre bir kümeleme işlemi yapılmaktadır.
- Yoğunluğa dayalı kümeleme tekniklerinde, veri tabanındaki daha yüksek yoğunluklu olan alanlarda kümeler oluşur.
- Küme yoğunluklarının seyrek olduğu alanlarda ise ya gürültülü veriler ya da küme sınırını oluşturan veriler bulunmaktadır.
- Bu algoritmayı büyük veri tabanları ve gürültülü verisi çok olan yapılarda kullanmak oldukça uygundur.
- Farklı büyüklüklerde ve şekillerdeki kümelerin oluşturulmasında bu algoritma kullanılabilir.



Density Based Clustering (Yoğunluk Tabanlı Kümeleme) (DEVAM)

- Yoğunluk: Verilen bir yarıçap (Eps) içerisinde olan nokta sayısıdır.
- Eğer bir noktanın Eps yarıçapında verilen minimum nokta sayısından (MinPts) daha fazla nokta varsa o noktaya **merkez nokta** denir. Bu noktalar kümenin iç bölgelerinde bulunan noktalardır.
- Eğer bir noktanın Eps yarıçapında verilen minimum nokta sayısından (MinPts) daha az nokta varsa ve o nokta bir merkez noktanın komşuluğunda ise noktaya **sınır nokta** denir.
- Ne merkez nokta ne de sınır nokta olan noktaya **gürültü nokta** denir.

Gürültü, belirli bir veri setinde veya formülünde bulunan açıklanamayan çeşitliliği veya rastgeleliği ifade eden bir terimdir.



Fuzzy c-means Clustering
(Bulanık c-ortalama Kümeleme)

- Bireylerin iki veya daha fazla kümeye ait olabilmesine izin verir.
- Tüm bireyler kümelerin her birine [0,1] arasında değişen birer üyelik değeri ile atanır. Bir bireyin tüm sınıflara olan üyelik değerleri toplamı “1” olmalıdır.
- Birey hangi küme merkezine yakın ise o kümeye ait olma ihtimali daha yüksek olacaktır.
- Üyelik matrisi belirsiz durumların tanımlanmasını kolaylaştırır. Ayrıca üyelik dereceleri düşük olduğundan sıra dışı verilerin etkisi azdır.
- Esnek bir yapıya sahiptir.
- Örtüşen kümeleri bulma kabiliyeti diğer bölünmeli algoritmalara göre daha fazladır.
- Bu algoritmasının bazı dezavantajları da vardır. Üyelik fonksiyonu işlemsel karmaşıklığı arttırdığı için zaman açından maliyetli bir bölünmeli kümeleme algoritmasıdır.

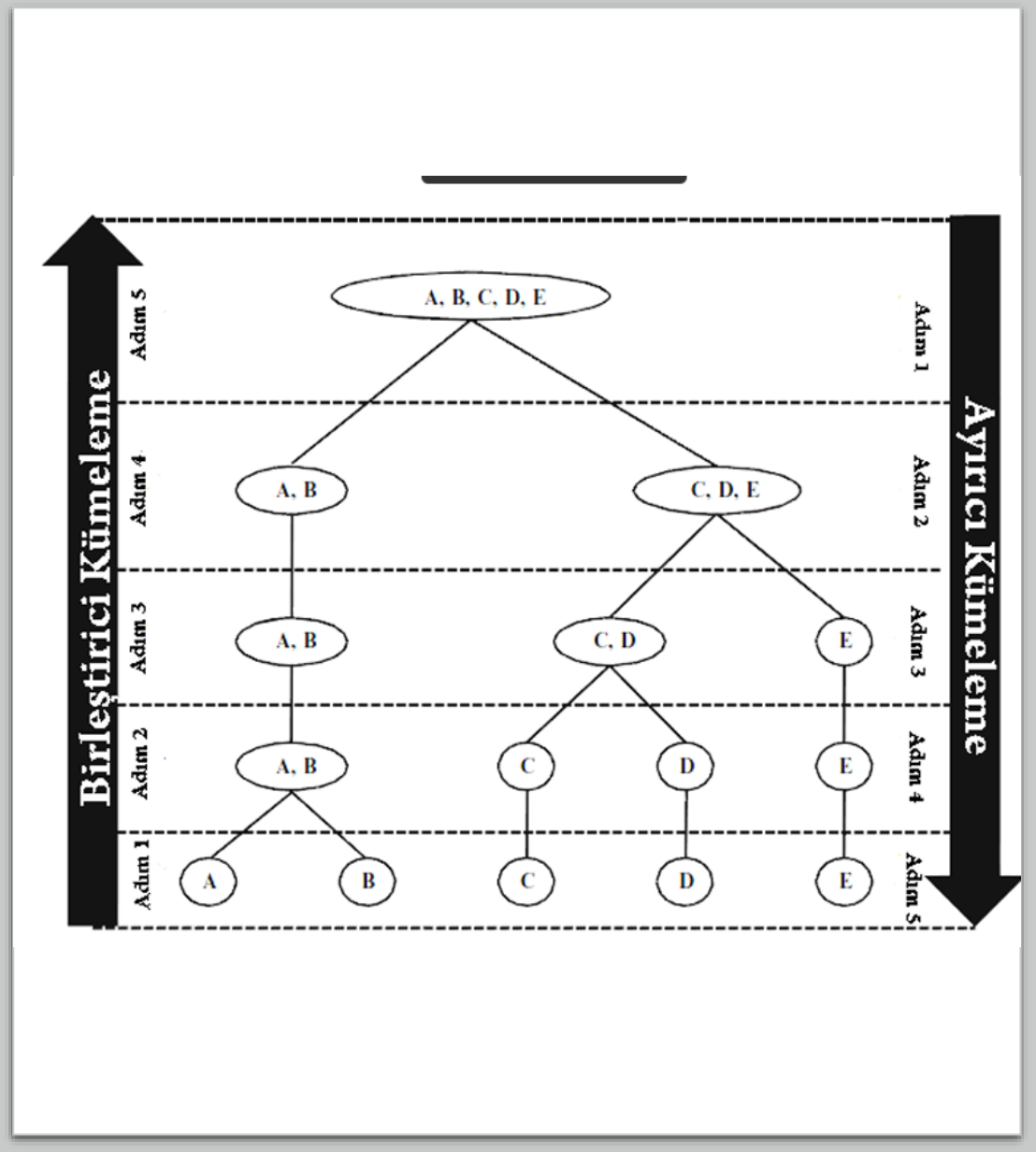
Hierarchical Clustering (Hiyerarşik kümeleme)

- Hiyerarşik yöntemler, genellikle tek birim içeren kümelerden başlayarak, tüm birimler bir kümede toplanana kadar birleştirme işlemi yaparak küme dizileri üretir. Bu yöntemler "**Birleştirici Hiyerarşik Kümeleme Yöntemi**" olarak adlandırılır.

- Diğer yöntemler ise; tek bir kümeyle başlayarak, birimleri art arda ayırarak, tek birim içeren kümeler oluşturana dek devam eder. Bu yöntemlere ise "**Ayırıcı Hiyerarşik Kümeleme Yöntemleri**" denir.

- Hiyerarşik sınıflamalar "dendrogram" olarak bilinen iki boyutlu şemalarla ifade edilebilir. Dendrogramlar birbirini izleyen her bir aşamadaki birleşmeleri ya da bölünmeleri gösterir.

- Birleştirici ve ayırıcı kümelemeye ilişkin örnek dendrogram yandaki şekilde verilmiştir.



Birleştirici hiyerarşik kümeleme yöntemleri

- Birleştirici hiyerarşik kümeleme yöntemleri Öklid uzaklığını kullanır.

- Tüm birimler tek başına birer grup kabul edilerek başlanır ve bu gruplar kendilerine “yakın” olan kümelerle birleştirilir.

- Bu yakınlığı tanımlamak için farklı yollar mevcuttur.

- En çok kullanılan birleştirici hiyerarşik kümeleme yöntemleri şunlardır:
    - Tek Bağlantı Yöntemi (Single Linkage Method, Nearest Neighbours Method)
    - Tam Bağlantı Yöntemi (Complete Linkage Method, Furthest Neighbours Method)
    - Ortalama Bağlantı Yöntemi (Average Linkage Method)
    - Küresel Ortalama Bağlantı Yöntemi (Centroid Method)
    - Ward Hiyerarşik Yöntemi (Ward’s Hierarchical Clustering)

- **Tek bağlantı yöntemi:** En yakın komşuluk tekniği olarak da bilinir. En kısa mesafe esasına dayanır. Bu yöntemde uzaklıklar matrisi kullanılarak birbirine en yakın kümeler birleştirilmek suretiyle birleştirmeler art arda tekrarlanmaktadır.

- **Tam bağlantı yöntemi:** En uzak komşuluk tekniği olarak da bilinir. Tek bağlantı yöntemine çok benzer ancak burada iki küme arsında uzaklık olarak her kümedeki eleman çiftlerinin arasındaki uzaklığın en büyüğü alınır.

- **Ortalama bağlantı yöntemi:** Bu yöntemde ayrı gruplarda yer alan gözlem çiftleri arasındaki ortalama uzaklık iki küme arasındaki uzaklık olarak alınır.

- **Küresel Ortalama Bağlantı Yöntemi:** Bu yaklaşımda ilk olarak her kümenin geometrik merkezi (centroid) hesaplanır. İki küme arasındaki uzaklık bu iki centroid arasındaki uzaklığa eşittir.

- **Ward's bağlantı yöntemi:** Bir kümenin ortasına düşen gözlemin, aynı kümenin içinde bulunan gözlemlerden ortalama uzaklığını esas alır.

Ayırıcı hiyerarşik kümeleme yöntemleri

- Ayırıcı hiyerarşik kümeleme yöntemleri tüm birimleri tek bir küme elemanı olarak kabul edip art arda daha küçük kümelere bölerek devam eder.

- Ayırıcı hiyerarşik kümeleme yöntemleri yukarıdan-aşağıya stratejisini kullanır.

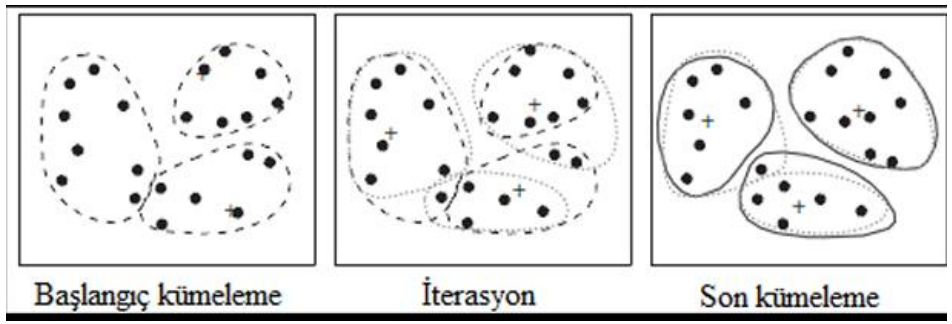
- Bu strateji tüm birimlerin tek bir kümede yer almasıyla başlar, bu küme hiyerarşinin kökünü oluşturur.

- Daha sonra kök, kendi içinde daha küçük kümelere, o kümeler de yinelemeli olarak kendinden daha küçük alt kümelere bölünür.

- Bu bölme işlemi her küme kendi içinde tutarlı en alt kümeye bölünene kadar devam eder

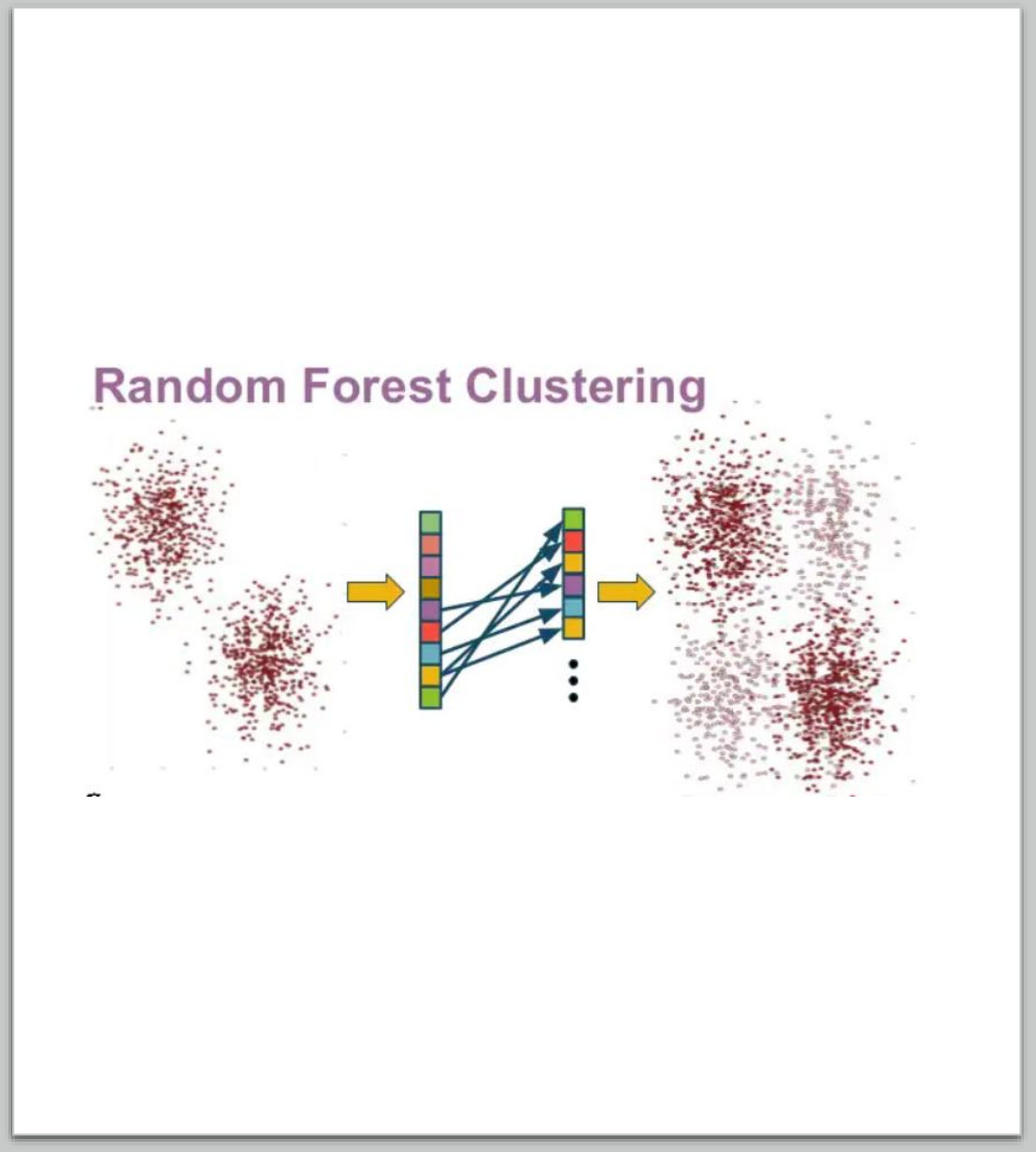
K-means Clustering
(K-ortalama Kümeleme)

- Bu yöntemde önce araştırmacının ön bilgisine ve tecrübesine dayanarak küme sayısı belirlenir.
- Sonra her kümenin tipik bir gözlemi seçilir, benzer gözlemler tipik gözlemin etrafında birer birer kümelendirilir.
- Burada bazı istatistiksel testler kullanılarak her kümeyi oluşturan gözlemlerin değişkenlere göre ortalamalarına bakılır.
- Güvenilir olması en belirgin üstünlüğüdür. Buna karşılık yorumlaması zordur.
- K-ortalama kümeleme yönteminde **Hartigan-Wong, Lloyd Forgy ve MacQueen** algoritmaları kullanılır.



Random Forest Clustering (Rastgele Orman Kümelemesi)

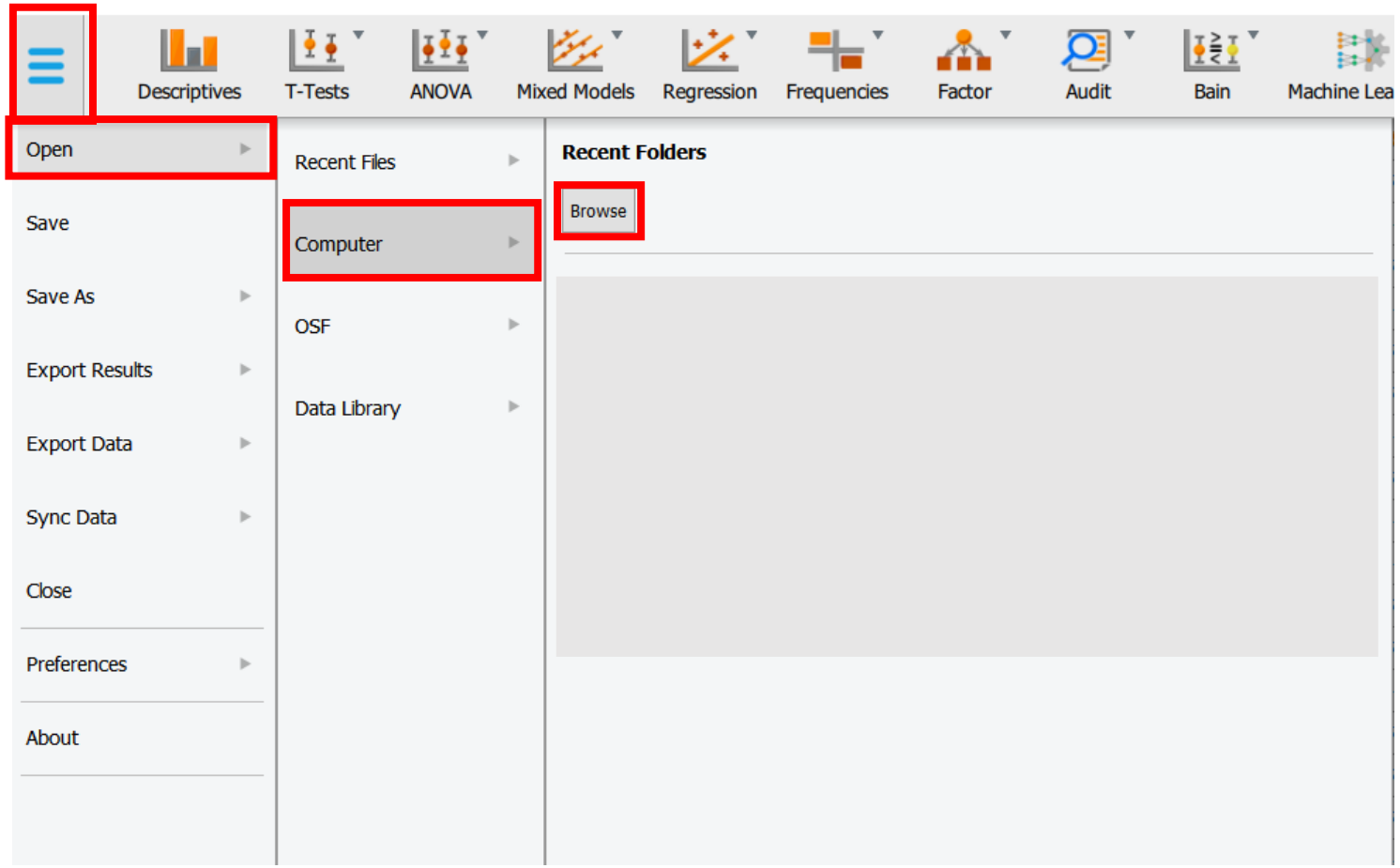
- Random Forest kümeleme, verileri her gözlemin yalnızca bir gruba ait olduğu birkaç kümeye bölmeyi amaçlayan katı bir bölümleme algoritmasıdır.
- Bu kümeleme yöntemi, Random Forest algoritmasını denetimsiz bir şekilde kullanır ve sonuç değişkeni (y) kullanılmaz.
- Rastgele Orman algoritması, aynı yaprak düğümünde biten gözlemlerin sıklığına dayalı olarak gözlemler arasındaki mesafenin bir tahminini veren bir yakınlık matrisi üretir.

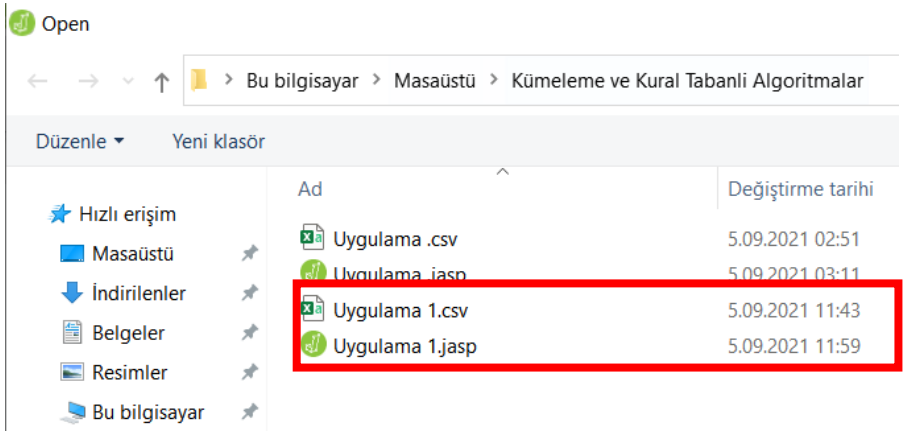


UYGULAMA

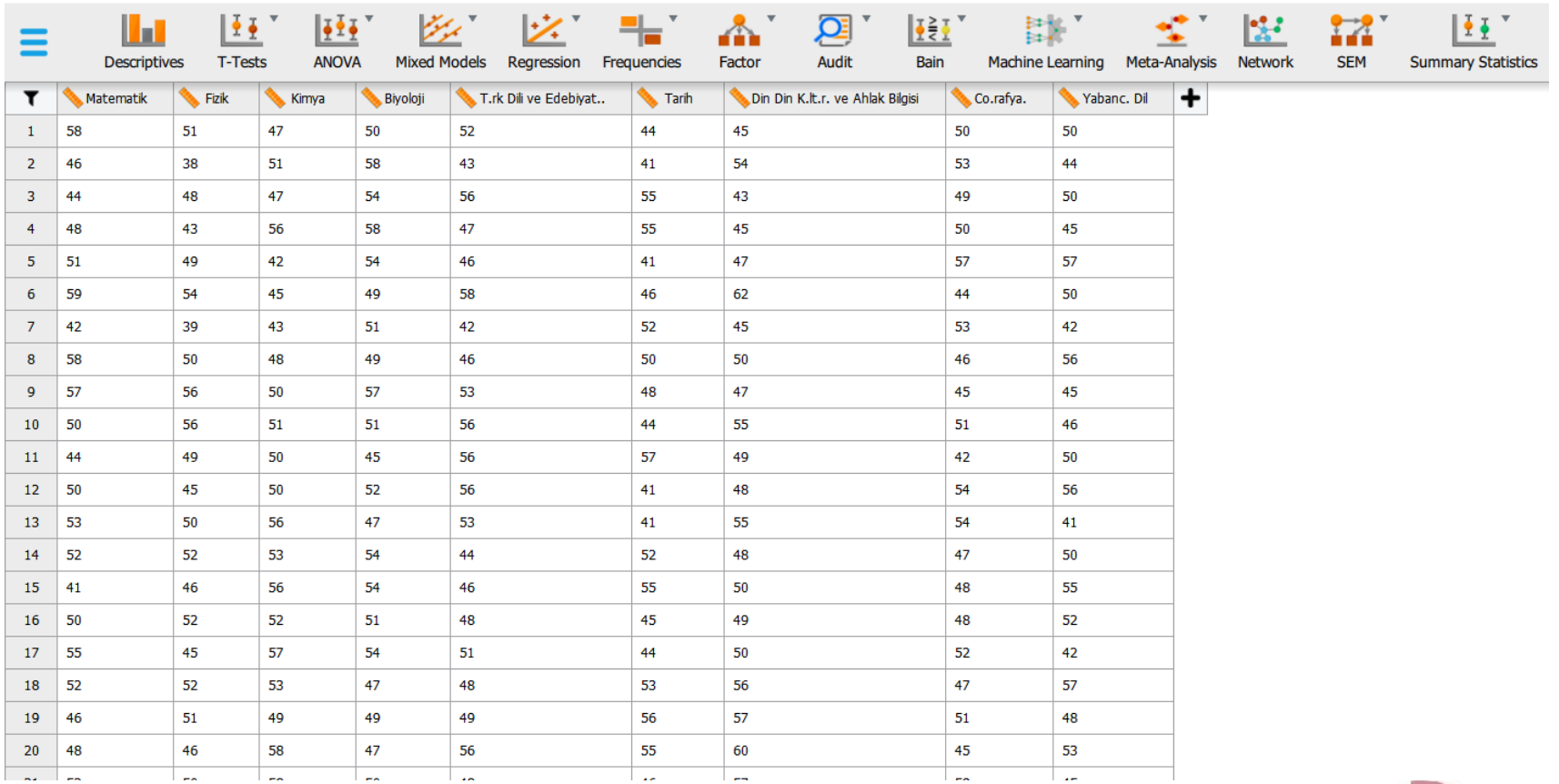
- 900 öğrenci üzerinden yapılan bir araştırmada öğrencilerin ders notlarına göre başarı durumları belirlenmek isteniyor. Bunun için öğrencilerin 9 dersten aldıkları notlar değerlendiriliyor.

VERİ GİRİŞİ


Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Audit
Bain
Machine Lea
Open
Save
Save As
Export Results
Export Data
Sync Data
Close
Preferences
About
Recent Files
Computer
OSF
Data Library
Recent Folders
Browse

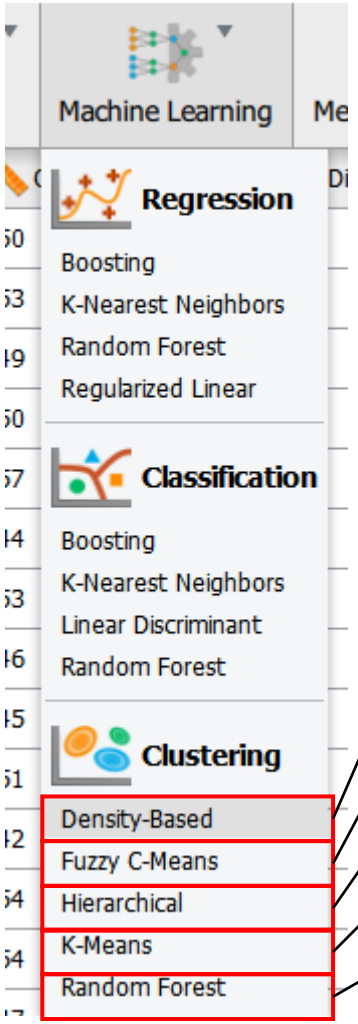

Open
Bu bilgisayar
Masaüstü
Kümeleme ve Kural Tabanli Algoritmalar
Düzenle
Yeni klasör
Hızlı erişim
Masaüstü
İndirilenler
Belgeler
Resimler
Bu bilgisayar
Ad
Değiştirme tarihi
Uygulama .csv
5.09.2021 02:51
Uygulama 1.csv
5.09.2021 11:43
Uygulama 1.jasp
5.09.2021 11:59

VERİ GİRİŞİ



	Matematik	Fizik	Kimya	Biyoloji	T.rk Dili ve Edebiyat..	Tarih	Din Din K.lt.r. ve Ahlak Bilgisi	Co.rafya.	Yabanc. Dil
1	58	51	47	50	52	44	45	50	50
2	46	38	51	58	43	41	54	53	44
3	44	48	47	54	56	55	43	49	50
4	48	43	56	58	47	55	45	50	45
5	51	49	42	54	46	41	47	57	57
6	59	54	45	49	58	46	62	44	50
7	42	39	43	51	42	52	45	53	42
8	58	50	48	49	46	50	50	46	56
9	57	56	50	57	53	48	47	45	45
10	50	56	51	51	56	44	55	51	46
11	44	49	50	45	56	57	49	42	50
12	50	45	50	52	56	41	48	54	56
13	53	50	56	47	53	41	55	54	41
14	52	52	53	54	44	52	48	47	50
15	41	46	56	54	46	55	50	48	55
16	50	52	52	51	48	45	49	48	52
17	55	45	57	54	51	44	50	52	42
18	52	52	53	47	48	53	56	47	57
19	46	51	49	49	49	56	57	51	48
20	48	46	58	47	56	55	60	45	53

VERİ GİRİŞİ



Density Based Clustering
(Yoğunluk Tabanlı Kümeleme)

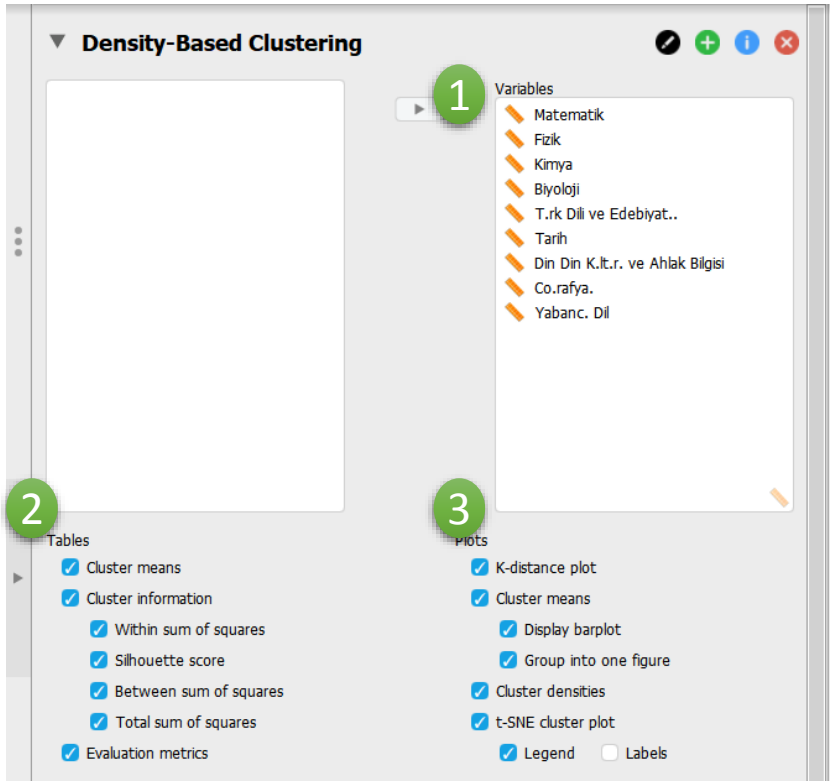
Fuzzy c-means Clustering
(Bulanık c-ortalama Kümeleme)

Hierarchical Clustering
(Hiyerarşik kümeleme)

K-means Clustering
(K-ortalama Kümeleme)

Random Forest Clustering
(Rastgele Orman Kümelemesi)

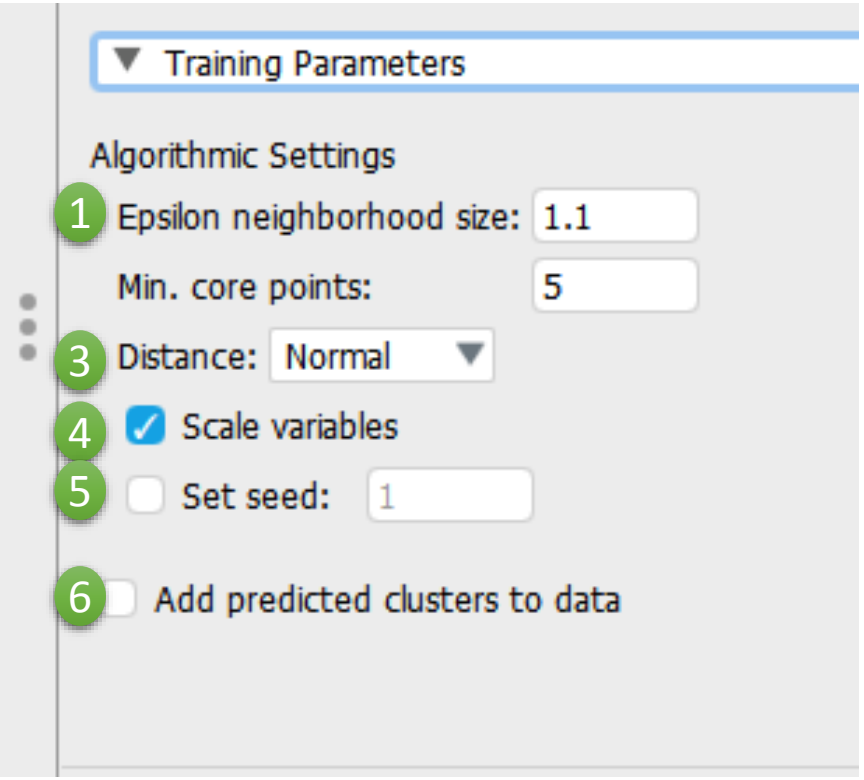
Density Based Clustering



1. **Veriables:** Makine öğrenmesi kümeleme yöntemi için değişken (dersler) bu bölüme atanır.
2. **Tables:** Oluşturmak istediğimiz tabloları belirlediğimiz bölümdür.
   - **Cluster means:** Değişkenlerim tüm kümelerde standardize ortalamaları
   - **Cluster Information:** Kümelerin toplam, grupiçi ve gruplar arası kareler toplamı ve tüm kümeler için siluet değerlerini incelediğimiz tablodur
   - **Evaluation Metrics Table:** Yöntemin performansını değerlendirdiğimiz tabloyu oluşturur.
3. **Plots:** Oluşturmak istediğimiz grafikleri belirlediğimiz bölümdür.
   - **K-distance plot:** Bu çizim, optimum Epsilon değerini belirlemek için kullanılabilir. Grafiğin bükülmeyi gösterdiği değer, optimum Epsilon değerini temsil eder.
   - **Cluster means:** Her tahmin değişkeni için, her kümenin ortalamasını ve %95 güven aralığını gösteren bir grafik oluşturur.
   - **Cluster densities:** Her tahmin değişkeni için kümeler için örtüşen yoğunlukları gösteren bir grafik oluşturur.
   - **t-SNE cluster plot:** Veri gözlemleri arasındaki göreceli mesafeleri göstermeyi amaçlayan iki boyutlu düşük boyutlu bir uzayda yüksek boyutlu verileri görselleştirmek için kullanılır.

$$\text{Standardizasyon} = \frac{\text{Değer} - \text{Değişken Ortalaması}}{\text{Değişken Standard Sapması}}$$

Density Based Clustering



1. **Epsilon neighborhood size:** Oluşturulacak kümelerin yarıçap boyutunu gösterir bu değer dışında kalan bireyler başka kümeye ait veya gürültü verisi olarakta adlandırılabilir.
2. **Min. core points:** Noktaların bir küme oluşturmasına izin vermek için Epsilon komşuluğunda olması gereken minimum nokta miktarını yansıtır. Eps ve MinPts parametreleri yoğunluk seviyesini belirler ve bir küme oluşturmak için belirli bir eşiği aşmak için belirli bir yarıçapta kaç nokta olması gerektiğini düzenler.
3. **Distance:** Küme oluşturmak için kullanılan ölçütü gösterir.
   - **Normal:** Noktalar arasındaki öklit uzaklıklarına göre kümeleme yapar.
   - **Correlated:** Noktalar arasındaki ilişkilere (korelasyon) göre kümeleme yapar.
4. **Scale variables:** Z-skor standardizasyonunu yapıp yapmayacağımızı belirleyen bölümdür. Standardizasyon yapmak verileri deki aykırı, uç veya gürültülü bireylerin sonuçları etkilememesini sağlar.
5. **Set seed:** Makine öğrenmesi yöntemleri algoritma tabanlı yöntemler olduğu için her çalıştırma sonucunda çıkan sonuçlar farklılık gösterebilir. Bunun için bu bölüm yardımı ile sonuçlar sabitlenir ve her program açıldığında aynı sonuçları görmemiz sağlanır.
6. **Add Predicted Clusters to Data:** Yapılan küme atamalarını veri setinde bir değişken sütunu olarak görmemizi sağlar.

TABLO DEĞERLENDİRMELERİ 1

Density-Based Clustering

1 Clusters	2 N	3 R^2	4 AIC	5 BIC	6 Silhouette
3	900	0.786	1177.160	1306.820	0.200

1 Elde edilen küme sayısını gösterir. Sonuçlarımıza göre değişkenler üç kümeden oluşmaktadır

2 Toplam birey sayısını gösterir. Veri setimizde 900 birey yer almaktadır.

3 9 dersin kümeleri açıklama oranını gösterir. Buna göre derslerin başarı durumunun %78,6'sını açıkladığı söylenebilir.

4 Model karşılaştırmalarında her zaman en düşük AIC değerini veren model tercih edilir. Burada en düşük AIC değeri 3 küme oluşturulması durumu için elde edilmiştir.

5 AIC'de olduğu gibi mevcut modeller arasında en küçük değerli BIC değerine sahip model, uygun model olarak seçilir. Burada en düşük BIC değeri 3 küme oluşturulması durumu için elde edilmiştir.

6 -1 ve 1 arasında değerler üretmektedir. 1'e en yakın K değeri en uygun olarak belirlenmektedir. Burada en yüksek siluet değeri için küme sayısı 3 olduğu bulunmuştur.

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

TÜBİTAK

4005 TÜBİTAK

TABLO DEĞERLENDİRMELERİ 2

Cluster Information ▼

Cluster	Noisepoints	1	2	3
Size	267	182	209	242
Explained proportion within-cluster heterogeneity	0.000	0.286	0.345	0.369
Within sum of squares	0.000	320.775	387.555	414.827
Silhouette score	0.000	0.485	0.413	0.499

Note. The Between Sum of Squares of the 3 cluster model is 4136.68

Note. The Total Sum of Squares of the 3 cluster model is 5259.84

Size: Her kümedeki birey sayısını gösterir.
Explained proportion within-cluster heterogeneity: Her kümenin heterojenlik varyasyon oranlarını verir
Total, Between, Within (Sum of Square): Verideki ortalamaya göre toplam değişimi, verinin tüm gruplardaki toplam değişimini ve verilerin gruplar içindeki toplam değişimini verir.
Silhouette Scoure: Tum gruplardaki Siluet skorlarını verir. Bu değerler bire ne kadar yakın ise o kadar iyi atama yapıldığını ifade eder.

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ
TÜBİTAK
4005 TÜBİTAK

TABLO DEĞERLENDİRMELERİ 3

Evaluation Metrics

	Value
Maximum diameter	9.560
Minimum separation	0.789
Pearson's γ	0.404
Dunn index	0.083
Entropy	1.376
Calinski-Harabasz index	347.973

Note. All metrics are based on the *euclidean* distance.

Evaluation Metrics: Bu tablo kümeleme algoritmasının uyum durumunu (performans metriklerini) verir.
Maximum diameter: Kümelerdeki iki nokta arasındaki maksimum uzaklığı verir.
Minimum separation: İki küme arasındaki münimum mesafeyi verir.
Pearson's γ: Burada iki küme arasındaki ilişkiyi gösterir. Bu değerin sıfıra yaklaşması aynı küme 1 e yaklaşması da farklı küme olduğunu gösterir.
Dunn endeksi: İki kümeyi ayrıştırmamızı sağlayan bir indekstir. Bu değer. Ayrıca bu değer gürültüler için uygun bölgeyi elde etmemizi sağlar.
Entropy: Entropy bireylerin oluşturduğu noktanın düzeni ile ilgili bir değerdir. Düzensizlik sistemin entropisinin artmasına sebep olur.
Calinski-Harabasz index: Varyans oranı kriteri olarak da bilinir. Bir nesnenin diğer kümelere kıyasla kendi kümesine ne kadar uyumlu olduğunun bir ölçüsüdür. Burada uyum, bir kümedeki veri noktalarından küme merkezine olan mesafelere göre tahmin edilir ve ayırma, küme merkezlerinin küresel merkeze olan mesafesine dayanır.

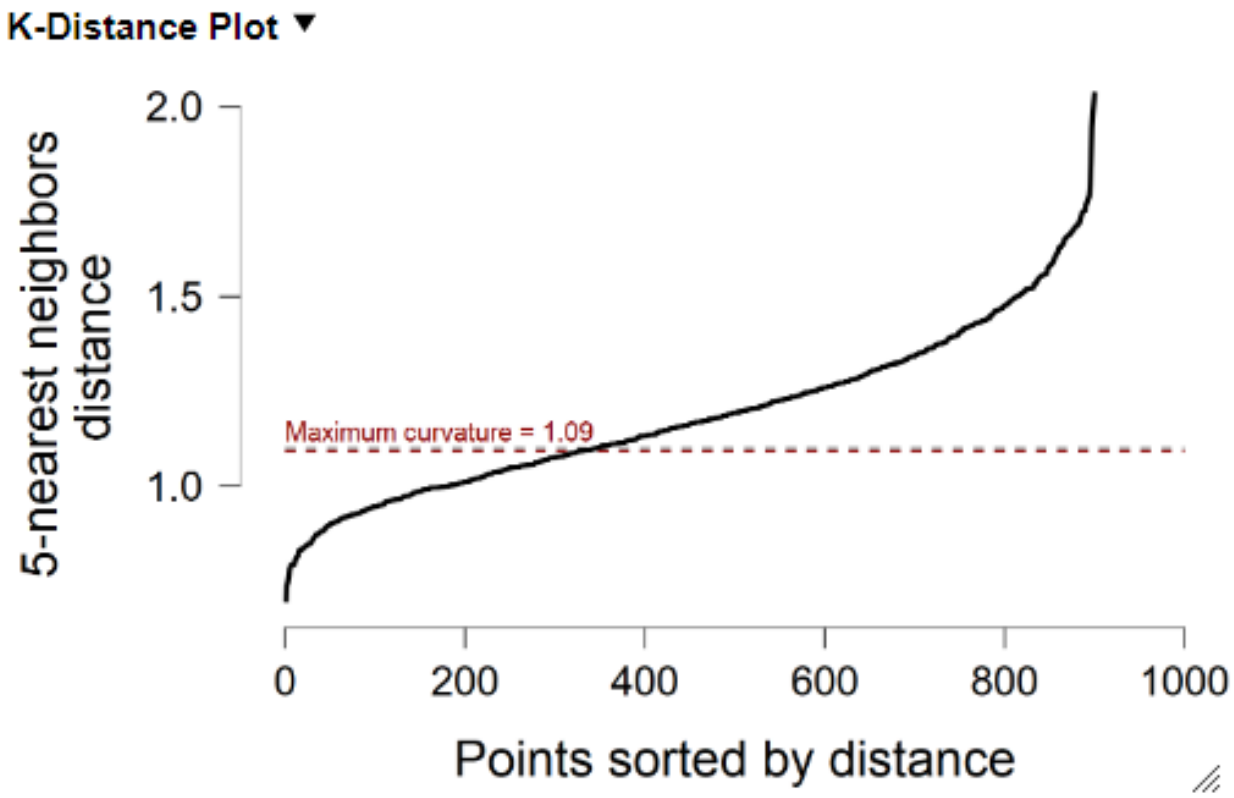
TABLO DEĞERLENDİRMELERİ 4

Cluster Means ▼

	Matematik	Fizik	Kimya	Biyoloji	T.rk Dili ve Edebiyat..	Tarih	Din Din K.lt.r. ve Ahlak Bilgisi	Co.rafya.	Yabanc. Dil
Cluster 0	−0.213	−0.237	−0.179	−0.272	−0.260	−0.294	−0.235	−0.224	−0.325
Cluster 1	−1.039	−1.050	−1.096	−0.997	−1.024	−1.000	−1.041	−1.094	−1.002
Cluster 2	−0.036	−0.009	0.030	0.001	−0.018	0.043	−0.011	−0.056	0.045
Cluster 3	1.047	1.059	0.996	1.049	1.072	1.040	1.052	1.118	1.073

Burada tüm kümelerde değişkenlerin (ders) standardize edilmiş değerlerinin ortalama farklarının toplamına yer verilmiştir.

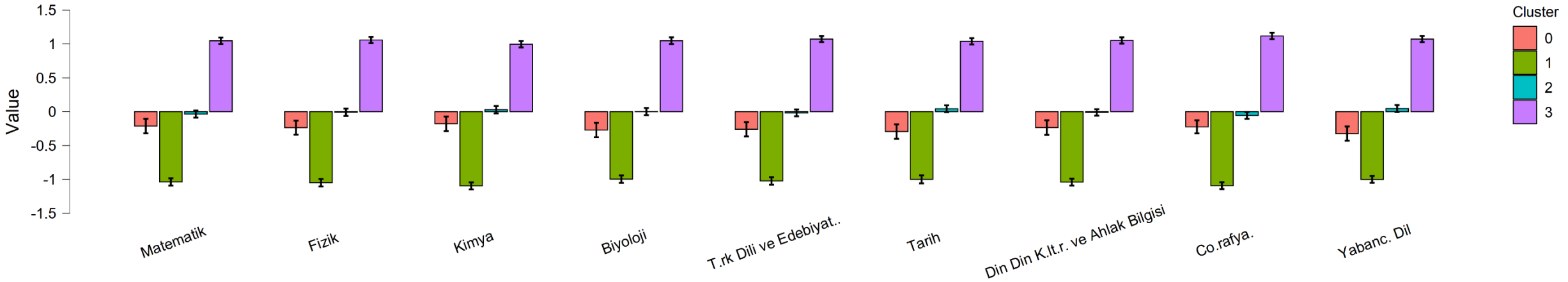
TABLO DEĞERLENDİRMELERİ 5



Y ekseninde en yakın komşu mesafesine ve x ekseninde mesafeye göre sıralanmış noktalara sahip bir grafik oluşturur.

Bu çizim, optimum Epsilon değerini belirlemek için kullanılabilir. Grafiğin bükülmeyi gösterdiği değer, optimum Epsilon değerini temsil eder.

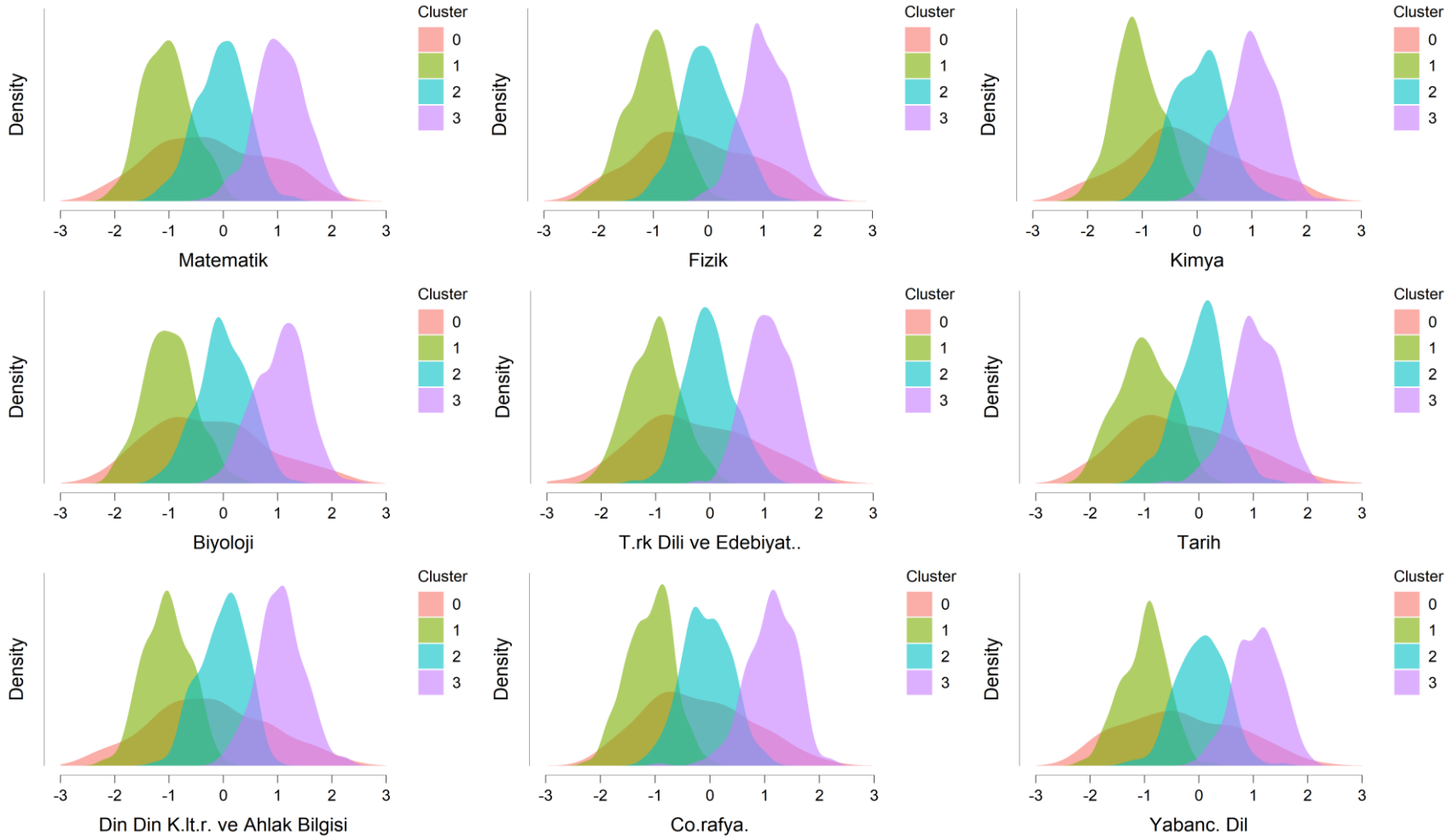
TABLO DEĞERLENDİRMELERİ 6



Standardize edilmiş değerlerin %95 güven aralığı ile grafiksel gösterimini gösterir.

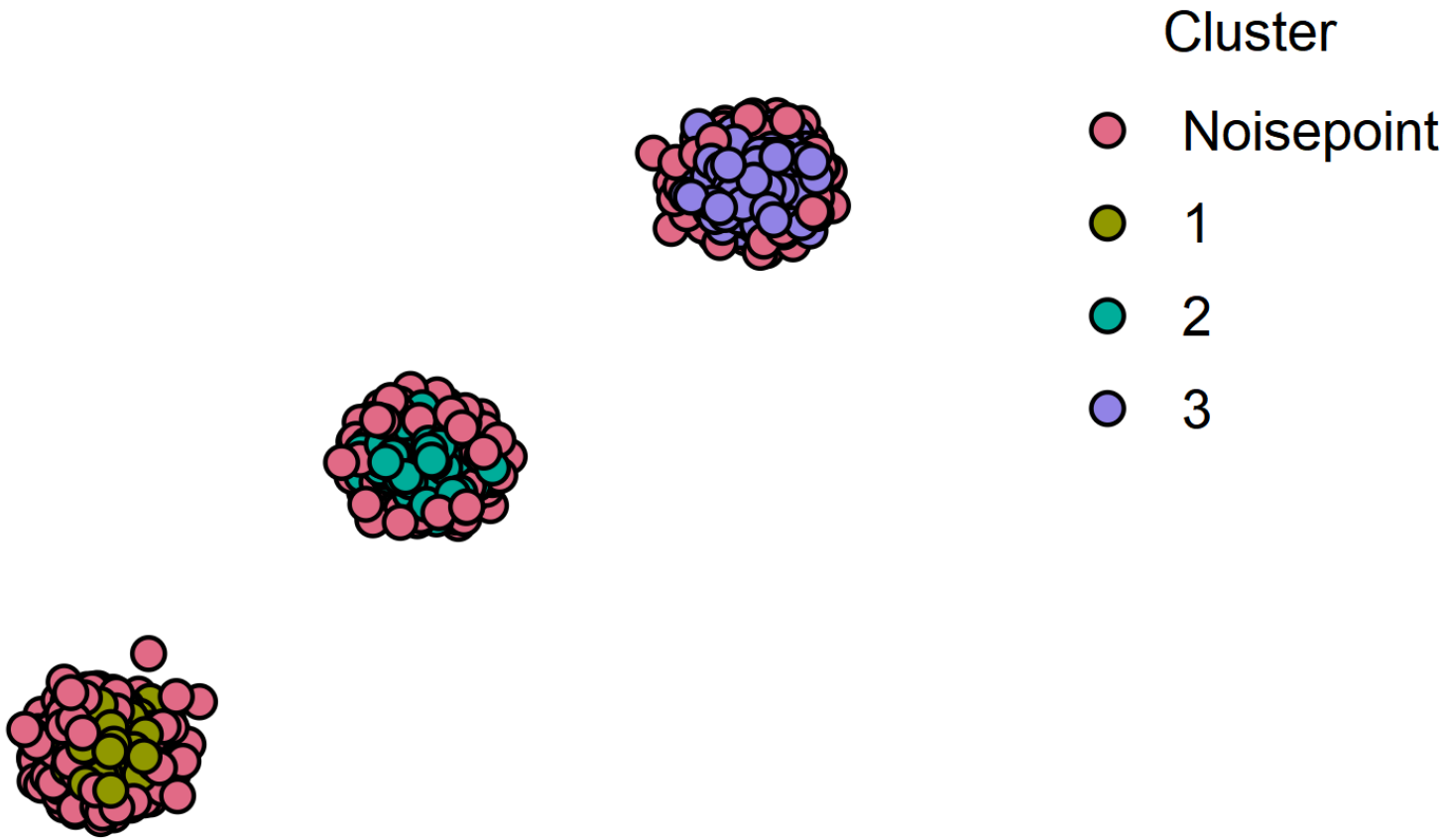
Güven aralığı, istatistik biliminde bir değerin olabileceği optimum sınır olarak belirtilebilir.

TABLO DEĞERLENDİRMELERİ 7



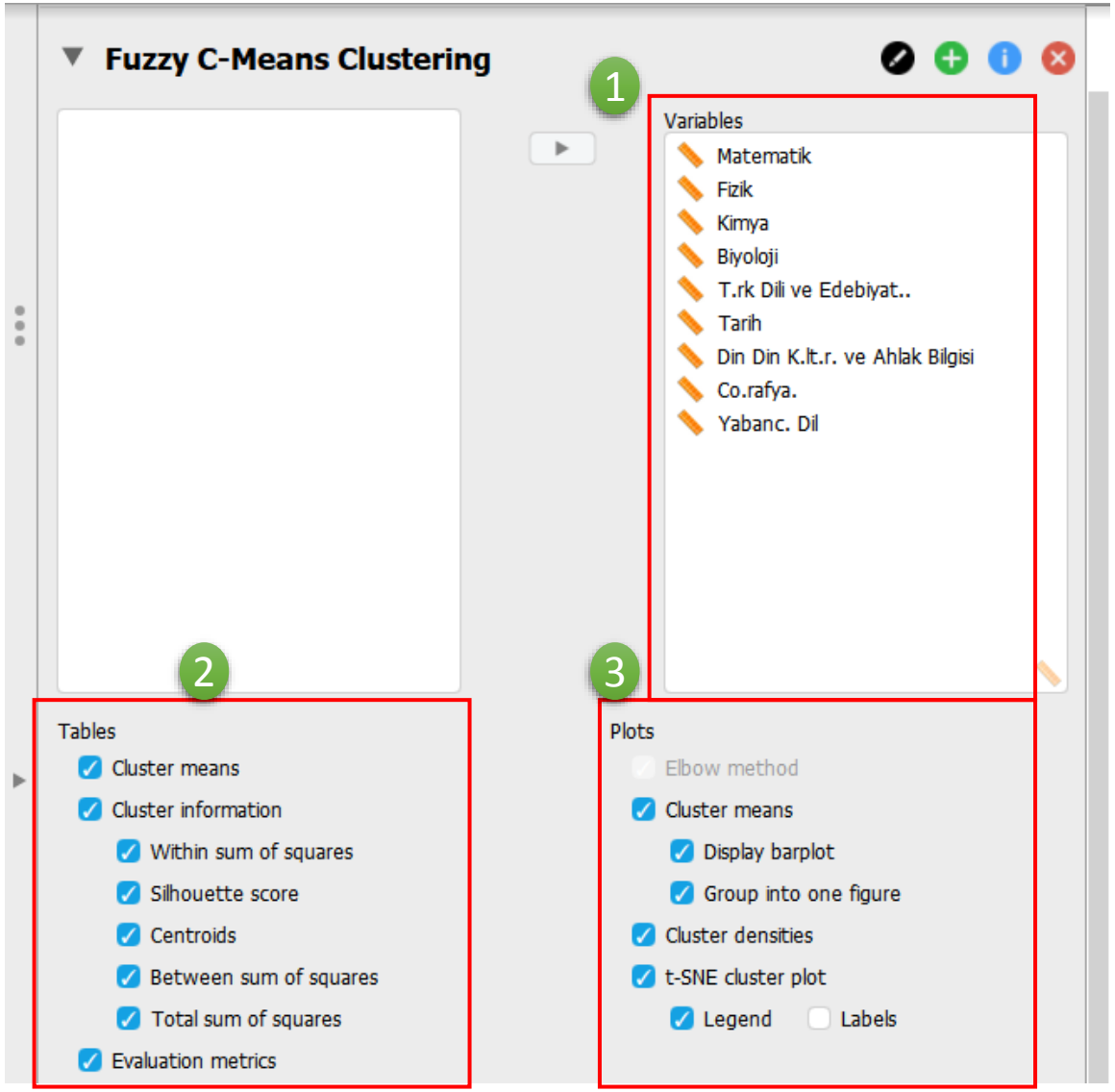
Tüm gruplarda yer alan standardize değerlerin dağılımlarını gösterir.

TABLO DEĞERLENDİRMELERİ 8



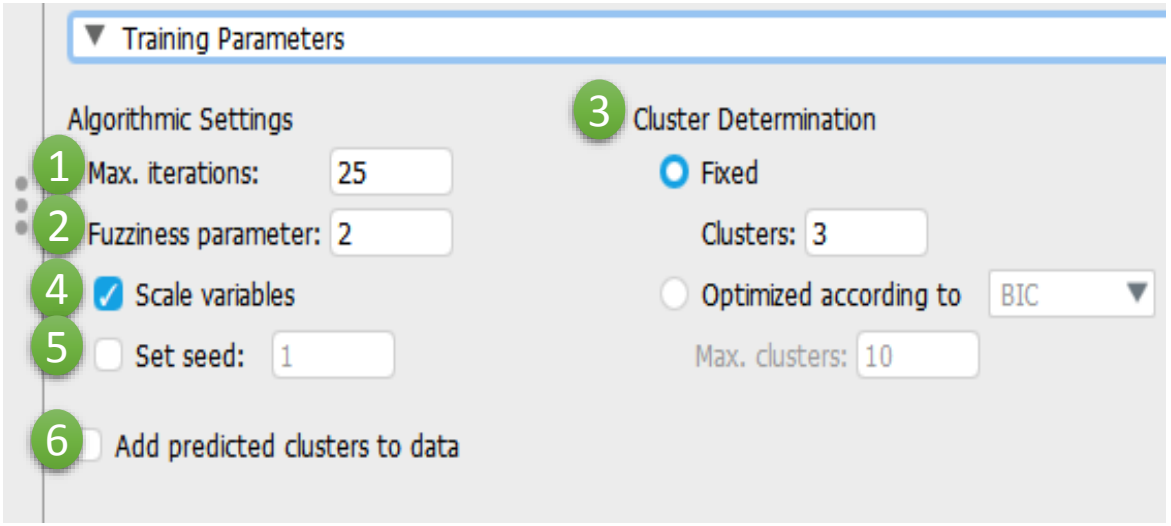
Tüm gruplarda yer alan standardize değerlerin kümelenmesini gösterir.

Fuzzy c-means Clustering



1. **Veriables:** Makine öğrenmesi kümeleme yöntemi için değişken (dersler) bu bölüme atanır.
2. **Tables:** Oluşturmak istediğimiz tabloları belirlediğimiz bölümdür.
   - **Cluster means:** Değişkenlerim tüm kümelerde standardize ortalamaları
   - **Cluster Information:** Kümelerin toplam, grupiçi ve gruplar arası kareler toplamı ve tüm kümeler için siluet değerlerini incelediğimiz tablodur
   - **Evaluation Metrics Table:** Yöntemin performansını değerlendirdiğimiz tabloyu oluşturur.
3. **Plots:** Oluşturmak istediğimiz grafikleri belirlediğimiz bölümdür.
   - **Cluster means:** Her tahmin değişkeni için, her kümenin ortalamasını ve %95 güven aralığını gösteren bir grafik oluşturur.
   - **Cluster densities:** Her tahmin değişkeni için kümeler için örtüşen yoğunlukları gösteren bir grafik oluşturur.
   - **t-SNE cluster plot:** Veri gözlemleri arasındaki göreceli mesafeleri göstermeyi amaçlayan iki boyutlu düşük boyutlu bir uzayda yüksek boyutlu verileri görselleştirmek için kullanılır.

Fuzzy c-means Clustering



1. **Max. iterations:** Maksimum yineleme sayısını ayarlar. Maksimum yineleme sayısı, algoritmanın optimal kümeleme çözümünü bulmak için yinelediği olası örnek sayısını yansıtır. Varsayılan olarak, bu 25 olarak ayarlanmıştır.
2. **Fuzziness parameter:** değeri artarsa (> 1) farklı kümelere üyeliklerin belirsizliği de artar. Başka bir deyişle, bulanıklık parametresi 1'e yaklaşırsa, bulanık kümelemenin sonucu, bir sabit kümeleme yöntemine benzer ve parametre artarsa, kümeleme sonucu daha bulanık hale gelir.
3. **Cluster Determination:** Küme sayısını belirlediğimiz bölümdür.
   - **Fixed:** Sabit miktarda küme oluşturmanıza olanak tanır. Bu, kendi belirtilen sayıda kümenizi oluşturmanıza ve böylece manuel olarak optimize etmenize olanak tanır.
   - **Max. clusters:** Bir optimizasyon yöntemi (AIC,BIC,Silhouette) seçerek bu yönteme göre en uygun küme sayısını belirlememizi sağlar. "Max clusters" bölümü oluşturulacak maksimum küme sayısını belirlememizi sağlar.
4. **Scale variables:** Z-skor standardizasyonunu yapıp yapmayacağımızı belirleyen bölümdür. Standardizasyon yapmak verileri deki aykırı, uç veya gürültülü bireylerin sonuçları etkilememesini sağlar.
5. **Set seed:** Makine öğrenmesi yöntemleri algoritma tabanlı yöntemler olduğu için her çalıştırma sonucunda çıkan sonuçlar farklılık gösterebilir. Bunun için bu bölüm yardımı ile sonuçlar sabitlenir ve her program açıldığında aynı sonuçları görmemiz sağlanır.
6. **Add Predicted Clusters to Data:** Yapılan küme atamalarını veri setinde bir değişken sütunu olarak görmemizi sağlar.

TABLO DEĞERLENDİRMELERİ 1

Fuzzy C-Means Clustering

1 Clusters	2 N	3 R^2	4 AIC	5 BIC	6 Silhouette
3	900	0.720	2123.810	2253.480	0.420

1 Elde edilen küme sayısını gösterir. Sonuçlarımıza göre değişkenler üç kümeden oluşmaktadır

2 Toplam birey sayısını gösterir. Veri setimizde 900 birey yer almaktadır.

3 9 dersin kümeleri açıklama oranını gösterir. Buna göre derslerin başarı durumunun %78,6'sını açıkladığı söylenebilir.

4 Model karşılaştırmalarında her zaman en düşük AIC değerini veren model tercih edilir. Burada en düşük AIC değeri 3 küme oluşturulması durumu için elde edilmiştir.

5 AIC'de olduğu gibi mevcut modeller arasında en küçük değerli BIC değerine sahip model, uygun model olarak seçilir. Burada en düşük BIC değeri 3 küme oluşturulması durumu için elde edilmiştir.

6 -1 ve 1 arasında değerler üretmektedir. 1'e en yakın K değeri en uygun olarak belirlenmektedir. Burada en yüksek siluet değeri için küme sayısı 3 olduğu bulunmuştur.

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

TÜBİTAK

4005 TÜBİTAK

TABLO DEĞERLENDİRMELERİ 2

Cluster Information ▼

Cluster	1	2	3
Size	299	301	300
Explained proportion within-cluster heterogeneity	0.340	0.362	0.299
Within sum of squares	703.074	748.393	618.345
Silhouette score	0.382	0.423	0.468
Centroid Matematik	−0.056	−0.813	0.636
Centroid Fizik	0.380	−0.381	1.007
Centroid Kimya	−0.131	−0.935	1.093
Centroid Biyoloji	−0.046	−0.894	1.267
Centroid T.rk Dili ve Edebiyat..	0.740	−1.259	0.693
Centroid Tarih	−0.027	−1.157	1.359
Centroid Din Din K.lt.r. ve Ahlak Bilgisi	−0.045	−0.969	0.857
Centroid Co.rafya.	0.129	−0.848	0.687
Centroid Yabanc. Dil	−0.164	−1.325	0.953

Note. The Between Sum of Squares of the 3 cluster model is 5325.82
Note. The Total Sum of Squares of the 3 cluster model is 7395.64

Size: Her kümedeki birey sayısını gösterir.
Explained proportion within-cluster heterogeneity: Her kümenin heterojenlik varyasyon oranlarını verir
Total, Between, Within (Sum of Square): Verideki ortalamaya göre toplam değişimi, verinin tüm gruplardaki toplam değişimini ve verilerin gruplar içindeki toplam değişimini verir.
Silhouette Scoure: Tum gruplardaki Siluet skorlarını verir. Bu değerler bire ne kadar yakın ise o kadar iyi atama yapıldığını ifade eder.
Centroidler: Değişken başına her kümenin standardize edilmiş ortalama değerini (ağırlık merkezi) gösterir.

TABLO DEĞERLENDİRMELERİ 3

Cluster Means

	Matematik	Fizik	Kimya	Biyoloji	T.rk Dili ve Edebiyat..	Tarih	Din Din K.lt.r. ve Ahlak Bilgisi	Co.rafya.	Yabanc. Dil
Cluster 1	0.013	−0.016	0.031	−0.015	−0.023	0.023	0.001	−0.048	0.040
Cluster 2	−1.058	−1.047	−1.068	−1.028	−1.044	−1.063	−1.057	−1.035	−1.081
Cluster 3	1.048	1.067	1.040	1.046	1.070	1.044	1.059	1.086	1.044

Burada tüm kümelerde değişkenlerin (ders) standardize edilmiş değerlerinin ortalama farklarının toplamına yer verilmiştir.

TABLO DEĞERLENDİRMELERİ 4

Evaluation Metrics ▼

	Value
Maximum diameter	4.997
Minimum separation	1.154
Pearson's γ	0.701
Dunn index	0.231
Entropy	1.099
Calinski-Harabasz index	1304.709

Note. All metrics are based on the *euclidean* distance.

Evaluation Metrics: Bu tablo kümeleme algoritmasının uyum durumunu (performans metriklerini) verir.

Maximum diameter: Kümelerdeki iki nokta arasındaki maksimum uzaklığı verir.

Minimum separation: İki küme arasındaki münimum mesafeyi verir.

Pearson's γ: Burada iki küme arasındaki ilişkiyi gösterir. Bu değerin sıfıra yaklaşması aynı küme 1 e yaklaşması da farklı küme olduğunu gösterir.

Dunn endeksi: İki kümeyi ayrıştırmamızı sağlayan bir indekstir. Bu değer. Ayrıca bu değer gürültüler için uygun bölgeyi elde etmemizi sağlar.

Entropy: Entropy bireylerin oluşturduğu noktanın düzeni ile ilgili bir değerdir. Düzensizlik sistemin entropisinin artmasına sebep olur.

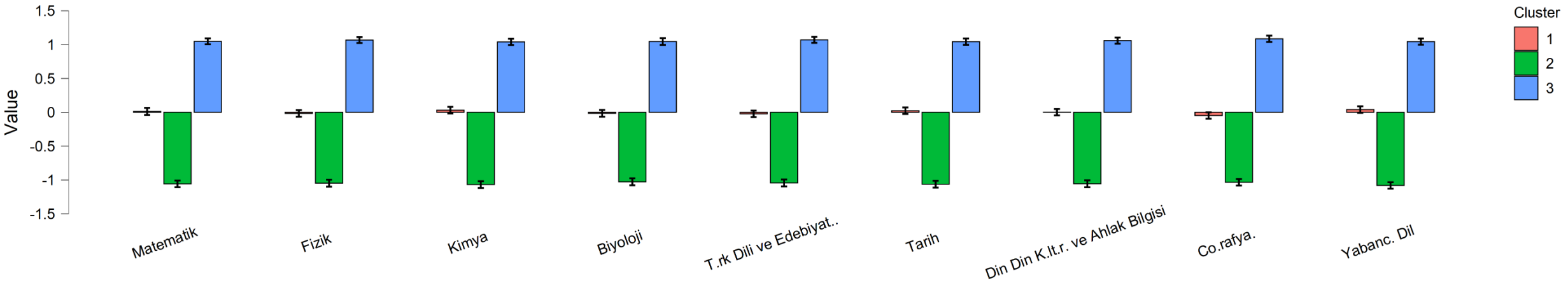
Calinski-Harabasz index: Varyans oranı kriteri olarak da bilinir. Bir nesnenin diğer kümelere kıyasla kendi kümesine ne kadar uyumlu olduğunun bir ölçüsüdür. Burada uyum, bir kümedeki veri noktalarından küme merkezine olan mesafelere göre tahmin edilir ve ayırma, küme merkezlerinin küresel merkeze olan mesafesine dayanır.

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

TÜBİTAK

4005 TÜBİTAK

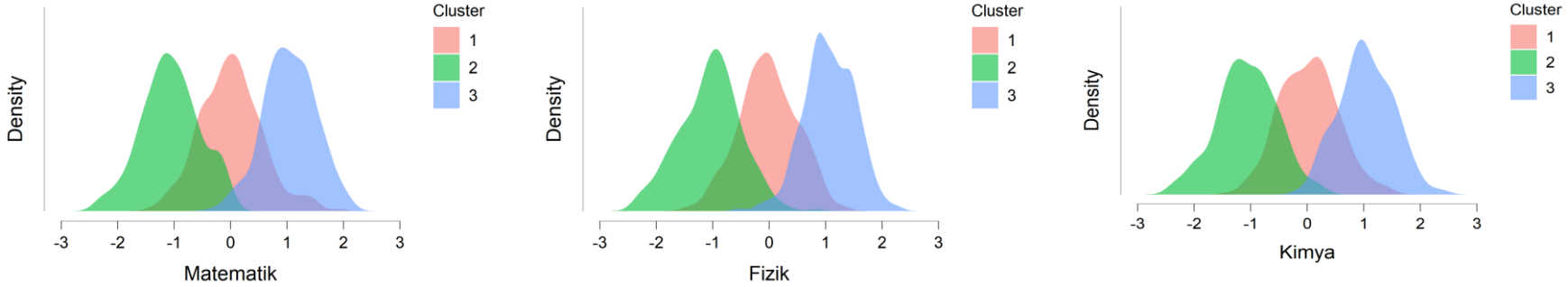
TABLO DEĞERLENDİRMELERİ 5

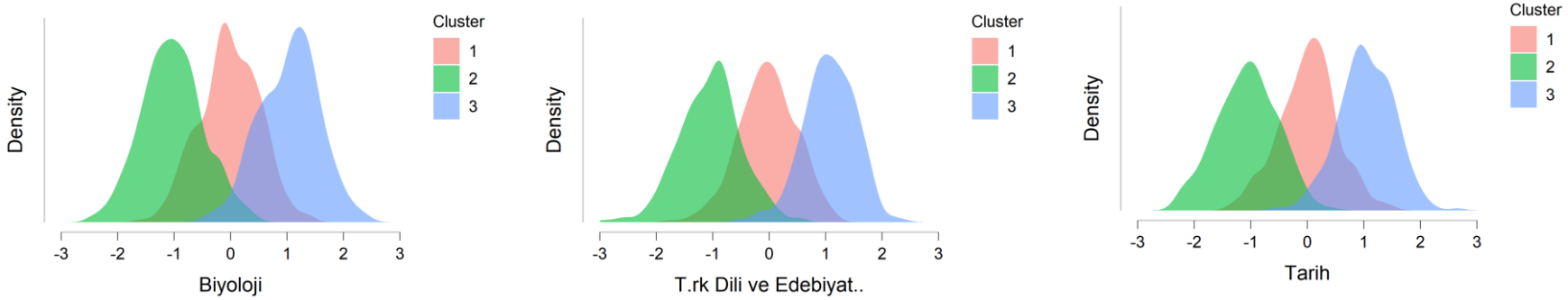


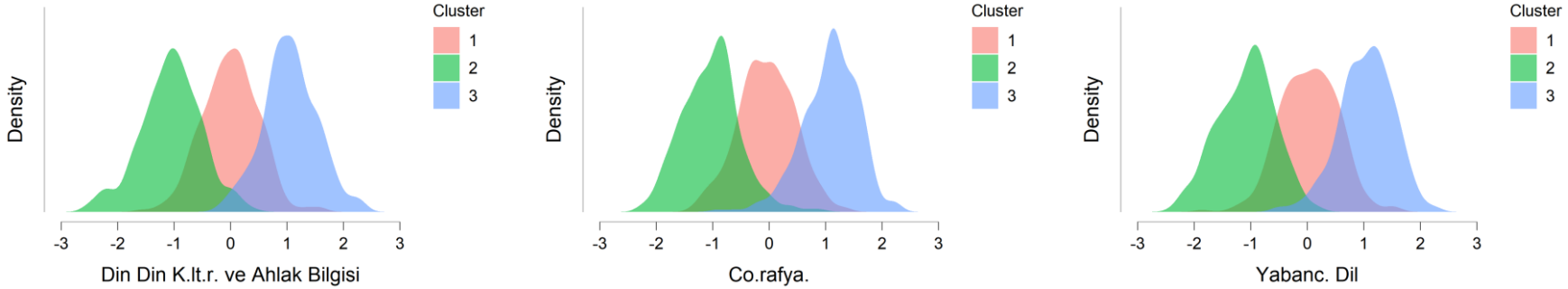
Standardize edilmiş değerlerin %95 güven aralığı ile grafiksel gösterimini gösterir.

Güven aralığı, istatistik biliminde bir değerin olabileceği optimum sınır olarak belirtilebilir.

TABLO DEĞERLENDİRMELERİ 6

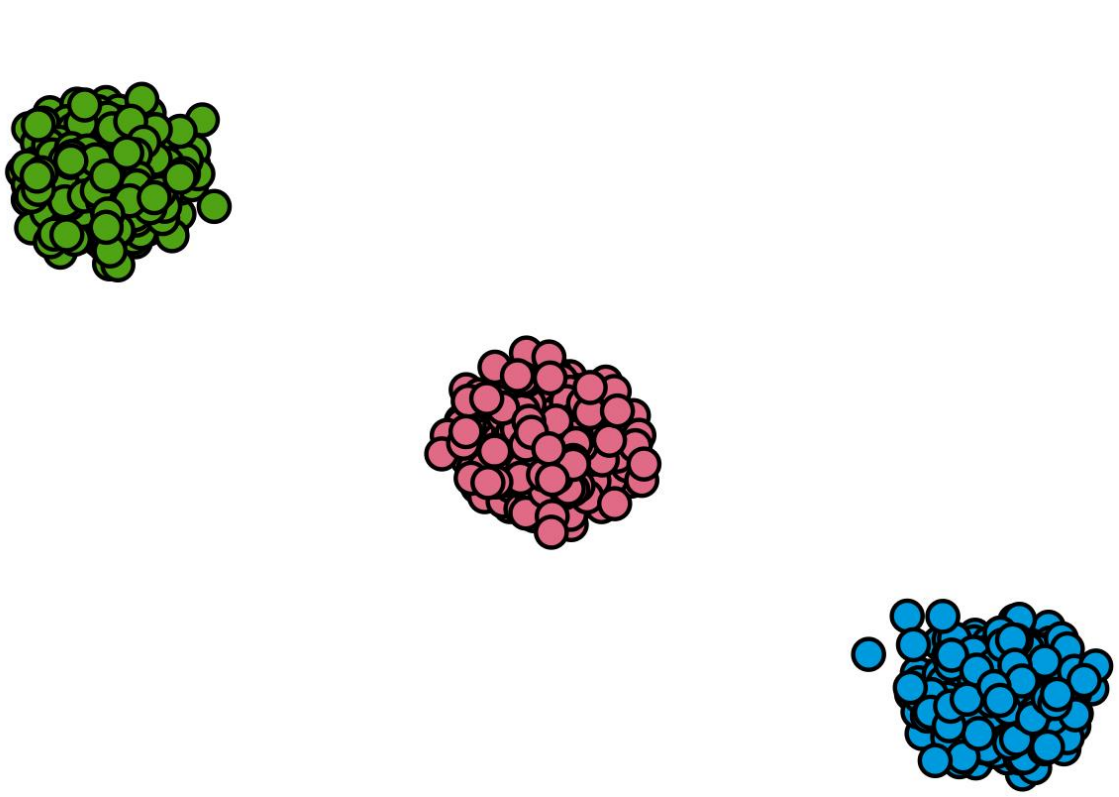






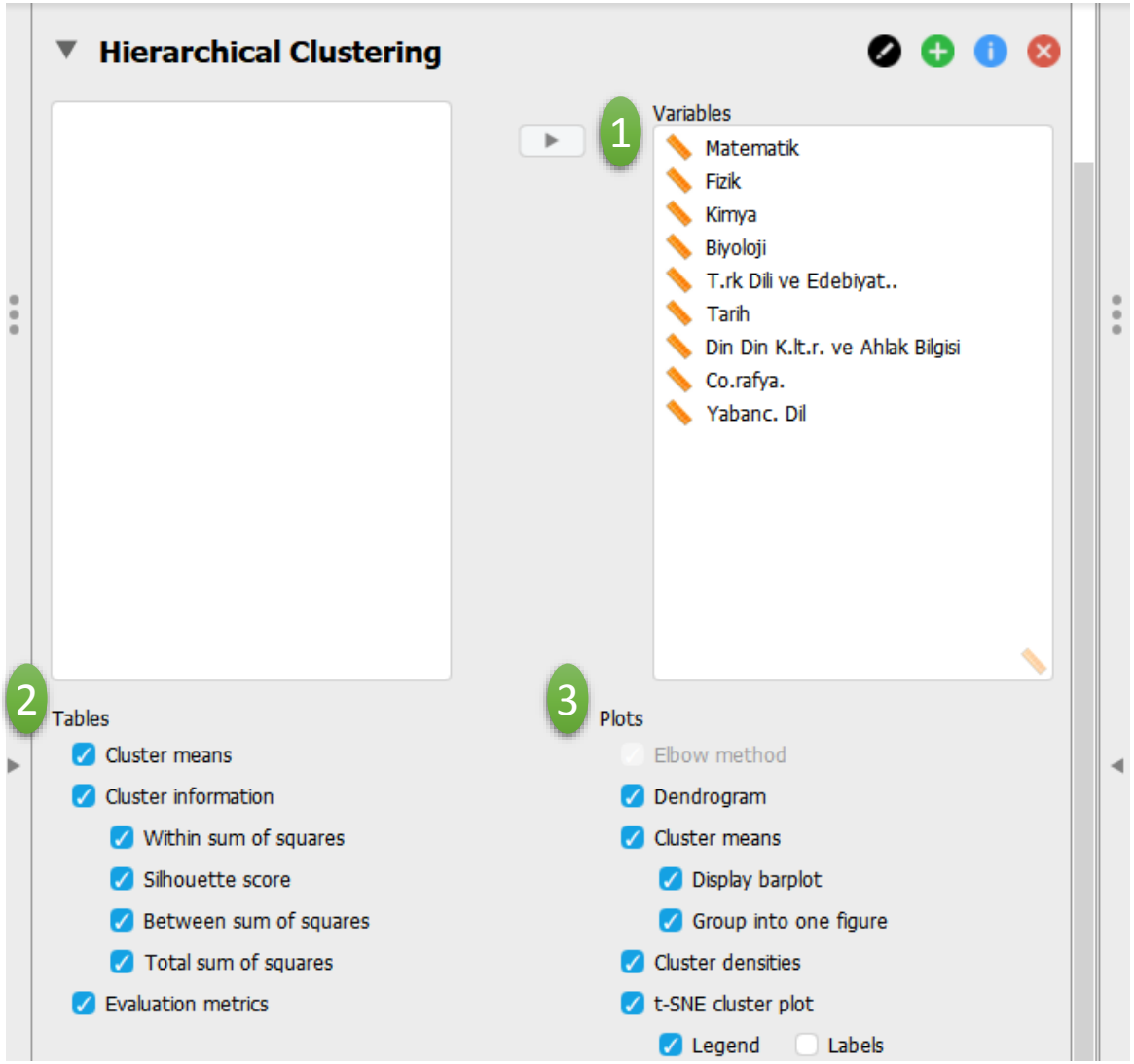
Tüm gruplarda yer alan standardize değerlerin dağılımlarını gösterir.

TABLO DEĞERLENDİRMELERİ 7



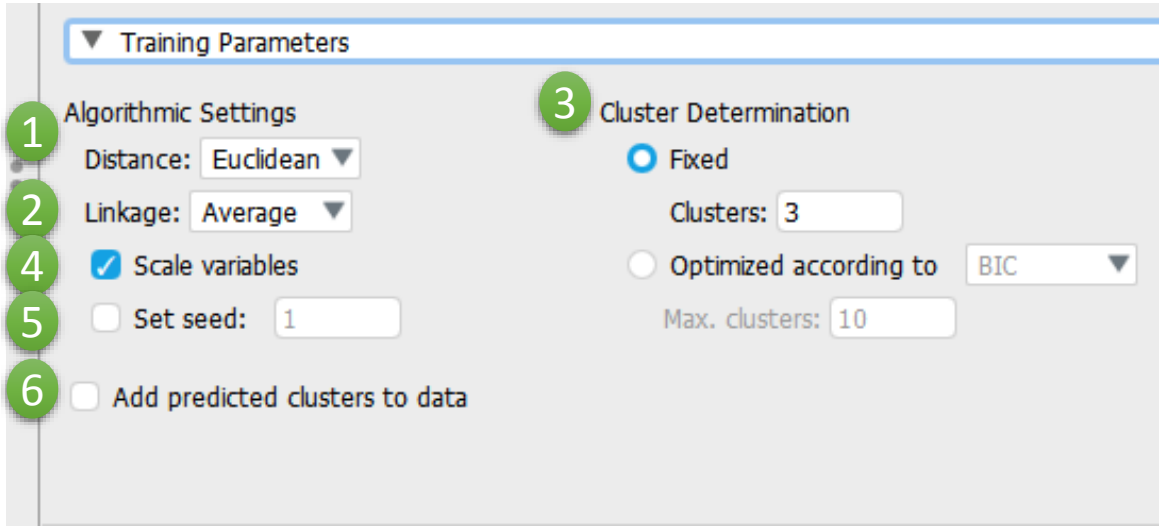
Tüm gruplarda yer alan standardize değerlerin kümelenmesini gösterir.

Hierarchical Clustering



1. **Veriables:** Makine öğrenmesi kümeleme yöntemi için değişken (dersler) bu bölüme atanır.
2. **Tables:** Oluşturmak istediğimiz tabloları belirlediğimiz bölümdür.
   - **Cluster means:** Değişkenlerim tüm kümelerde standardize ortalamaları
   - **Cluster Information:** Kümelerin toplam, grupiçi ve gruplar arası kareler toplamı ve tüm kümeler için siluet değerlerini incelediğimiz tablodur
   - **Evaluation Metrics Table:** Yöntemin performansını değerlendirdiğimiz tabloyu oluşturur.
3. **Plots:** Oluşturmak istediğimiz grafikleri belirlediğimiz bölümdür.
   - **Dendrogram:** Kümeleme çıktısının bir dendrogramını oluşturur. (Çıktı bölümünde detaylı olarak anlatılacaktır).
   - **Cluster means:** Her tahmin değişkeni için, her kümenin ortalamasını ve %95 güven aralığını gösteren bir grafik oluşturur.
   - **Cluster densities:** Her tahmin değişkeni için kümeler için örtüşen yoğunlukları gösteren bir grafik oluşturur.
   - **t-SNE cluster plot:** Veri gözlemleri arasındaki göreceli mesafeleri göstermeyi amaçlayan iki boyutlu düşük boyutlu bir uzayda yüksek boyutlu verileri görselleştirmek için kullanılır.

Hierarchical Clustering



1. **Distance:** Küme oluşturmak için kullanılan ölçütü gösterir.
   - **Euclidean:** Noktalar arasındaki öklit uzaklıklarına göre kümeleme yapar.
   - **Pearson:** Noktalar arasındaki ilişkilere (korelasyon) göre kümeleme yapar.
2. **Linkage:** Kullanılan bağlantı ölçüsünü belirtin.
3. **Cluster Determination:** Küme sayısını belirlediğimiz bölümdür.
   - **Fixed:** Sabit miktarda küme oluşturmanıza olanak tanır. Bu, kendi belirtilen sayıda kümenizi oluşturmanıza ve böylece manuel olarak optimize etmenize olanak tanır.
   - **Max. clusters:** Bir optimizasyon yöntemi (AIC,BIC,Silhouette) seçerek bu yönteme göre en uygun küme sayısını belirlememizi sağlar. "Max clusters" bölümü oluşturulacak maksimum küme sayısını belirlememizi sağlar.
4. **Scale variables:** Z-skor standardizasyonunu yapıp yapmayacağımızı belirleyen bölümdür. Standardizasyon yapmak verileri deki aykırı, uç veya gürültülü bireylerin sonuçları etkilememesini sağlar.
5. **Set seed:** Makine öğrenmesi yöntemleri algoritma tabanlı yöntemler olduğu için her çalıştırma sonucunda çıkan sonuçlar farklılık gösterebilir. Bunun için bu bölüm yardımı ile sonuçlar sabitlenir ve her program açıldığında aynı sonuçları görmemiz sağlanır.
6. **Add Predicted Clusters to Data:** Yapılan küme atamalarını veri setinde bir değişken sütunu olarak görmemizi sağlar.

EK (Slayt 13)- **Linkage Bölümünde Kullanılan Bağlantılar**

- **Tek bağlantı yöntemi:** En yakın komşuluk tekniği olarak da bilinir. En kısa mesafe esasına dayanır. Bu yöntemde uzaklıklar matrisi kullanılarak birbirine en yakın kümeler birleştirilmek suretiyle birleştirmeler art arda tekrarlanmaktadır.

- **Tam bağlantı yöntemi:** En uzak komşuluk tekniği olarak da bilinir. Tek bağlantı yöntemine çok benzer ancak burada iki küme arsında uzaklık olarak her kümedeki eleman çiftlerinin arasındaki uzaklığın en büyüğü alınır.

- **Ortalama bağlantı yöntemi:** Bu yöntemde ayrı gruplarda yer alan gözlem çiftleri arasındaki ortalama uzaklık iki küme arasındaki uzaklık olarak alınır.

- **Küresel Ortalama Bağlantı Yöntemi:** Bu yaklaşımda ilk olarak her kümenin geometrik merkezi (centroid) hesaplanır. İki küme arasındaki uzaklık bu iki centroid arasındaki uzaklığa eşittir.

- **Ward's bağlantı yöntemi:** Bir kümenin ortasına düşen gözlemin, aynı kümenin içinde bulunan gözlemlerden ortalama uzaklığını esas alır.

TABLO DEĞERLENDİRMELERİ 1

Hierarchical Clustering

1 Clusters	2 N	3 R^2	4 AIC	5 BIC	6 Silhouette
3	900	0.744	2123.470	2253.140	0.420

1 Elde edilen küme sayısını gösterir. Sonuçlarımıza göre değişkenler üç kümeden oluşmaktadır

2 Toplam birey sayısını gösterir. Veri setimizde 900 birey yer almaktadır.

3 9 dersin kümeleri açıklama oranını gösterir. Buna göre derslerin başarı durumunun %78,6'sını açıkladığı söylenebilir.

4 Model karşılaştırmalarında her zaman en düşük AIC değerini veren model tercih edilir. Burada en düşük AIC değeri 3 küme oluşturulması durumu için elde edilmiştir.

5 AIC'de olduğu gibi mevcut modeller arasında en küçük değerli BIC değerine sahip model, uygun model olarak seçilir. Burada en düşük BIC değeri 3 küme oluşturulması durumu için elde edilmiştir.

6 -1 ve 1 arasında değerler üretmektedir. 1'e en yakın K değeri en uygun olarak belirlenmektedir. Burada en yüksek siluet değeri için küme sayısı 3 olduğu bulunmuştur.

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

TÜBİTAK

4005 TÜBİTAK

TABLO DEĞERLENDİRMELERİ 2

Cluster Information

Cluster	1	2	3
Size	300	301	299
Explained proportion within-cluster heterogeneity	0.358	0.346	0.296
Within sum of squares	741.013	716.613	611.845
Silhouette score	0.426	0.378	0.470

Note. The Between Sum of Squares of the 3 cluster model is 6021.53

Note. The Total Sum of Squares of the 3 cluster model is 8091

Size: Her kümedeki birey sayısını gösterir.
Explained proportion within-cluster heterogeneity: Her kümenin heterojenlik varyasyon oranlarını verir
Total, Between, Within (Sum of Square): Verideki ortalamaya göre toplam değişimi, verinin tüm gruplardaki toplam değişimini ve verilerin gruplar içindeki toplam değişimini verir.
Silhouette Scoure: Tum gruplardaki Siluet skorlarını verir. Bu değerler bire ne kadar yakın ise o kadar iyi atama yapıldığını ifade eder.

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ
TÜBİTAK
4005 TÜBİTAK

TABLO DEĞERLENDİRMELERİ 3

Cluster Means ▼

	Matematik	Fizik	Kimya	Biyoloji	T.rk Dili ve Edebiyat..	Tarih	Din Din K.lt.r. ve Ahlak Bilgisi	Co.rafya.	Yabanc. Dil
Cluster 1	−1.058	−1.052	−1.073	−1.027	−1.044	−1.067	−1.060	−1.038	−1.082
Cluster 2	0.011	−0.012	0.034	−0.014	−0.026	0.029	0.001	−0.048	0.037
Cluster 3	1.050	1.067	1.041	1.044	1.074	1.041	1.063	1.090	1.048

Burada tüm kümelerde değişkenlerin (ders) standardize edilmiş değerlerinin ortalama farklarının toplamına yer verilmiştir.

TABLO DEĞERLENDİRMELERİ 4

Evaluation Metrics

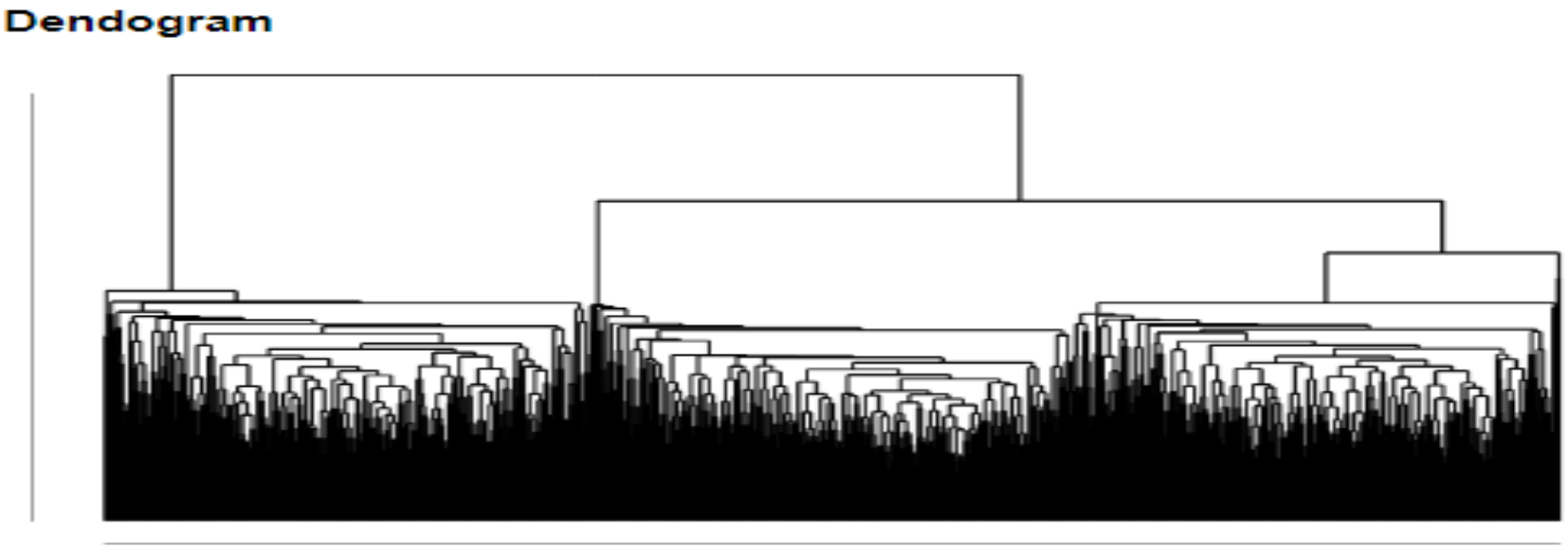
	Value
Maximum diameter	4.997
Minimum separation	1.154
Pearson's γ	0.701
Dunn index	0.231
Entropy	1.099
Calinski-Harabasz index	1304.999

Note. All metrics are based on the *euclidean* distance.

Evaluation Metrics: Bu tablo kümeleme algoritmasının uyum durumunu (performans metriklerini) verir.
Maximum diameter: Kümelerdeki iki nokta arasındaki maksimum uzaklığı verir.
Minimum separation: İki küme arasındaki münimum mesafeyi verir.
Pearson's γ: Burada iki küme arasındaki ilişkiyi gösterir. Bu değerin sıfıra yaklaşması aynı küme 1 e yaklaşması da farklı küme olduğunu gösterir.
Dunn endeksi: İki kümeyi ayrıştırmamızı sağlayan bir indekstir. Bu değer. Ayrıca bu değer gürültüler için uygun bölgeyi elde etmemizi sağlar.
Entropy: Entropy bireylerin oluşturduğu noktanın düzeni ile ilgili bir değerdir. Düzensizlik sistemin entropisinin artmasına sebep olur.
Calinski-Harabasz index: Varyans oranı kriteri olarak da bilinir. Bir nesnenin diğer kümelere kıyasla kendi kümesine ne kadar uyumlu olduğunun bir ölçüsüdür. Burada uyum, bir kümedeki veri noktalarından küme merkezine olan mesafelere göre tahmin edilir ve ayırma, küme merkezlerinin küresel merkeze olan mesafesine dayanır.

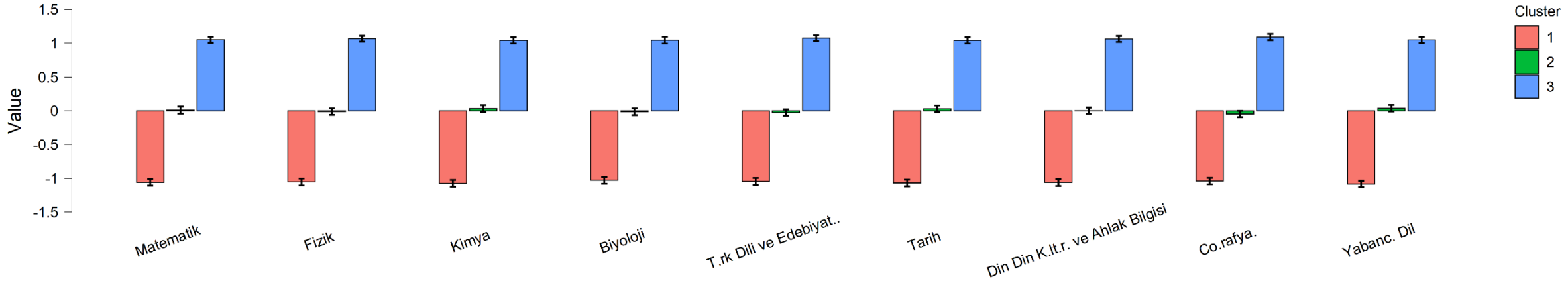
İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ
TÜBİTAK
4005 TÜBİTAK

TABLO DEĞERLENDİRMELERİ 5



Bir dendrogram, baş aşağı bir ağaç olarak yorumlanabilir. Altta yapraklar vardır ve ağaç yukarı çıkarken yapraklar dallar halinde birleşir (yani gözlem grupları). Yapraklar bir gözleme karşılık gelir. Ayrıca, yaprakların bir araya geldiği yükseklik, gözlemler veya gözlem grupları arasındaki (farklılık) durumu gösterir. Dendrogram, dendrogram içinde bir kesim oluşturarak kümeler oluşturmak için kullanılır. Kümeleme yapısını çıkarmak için bir dendrogram kullanılabilir.

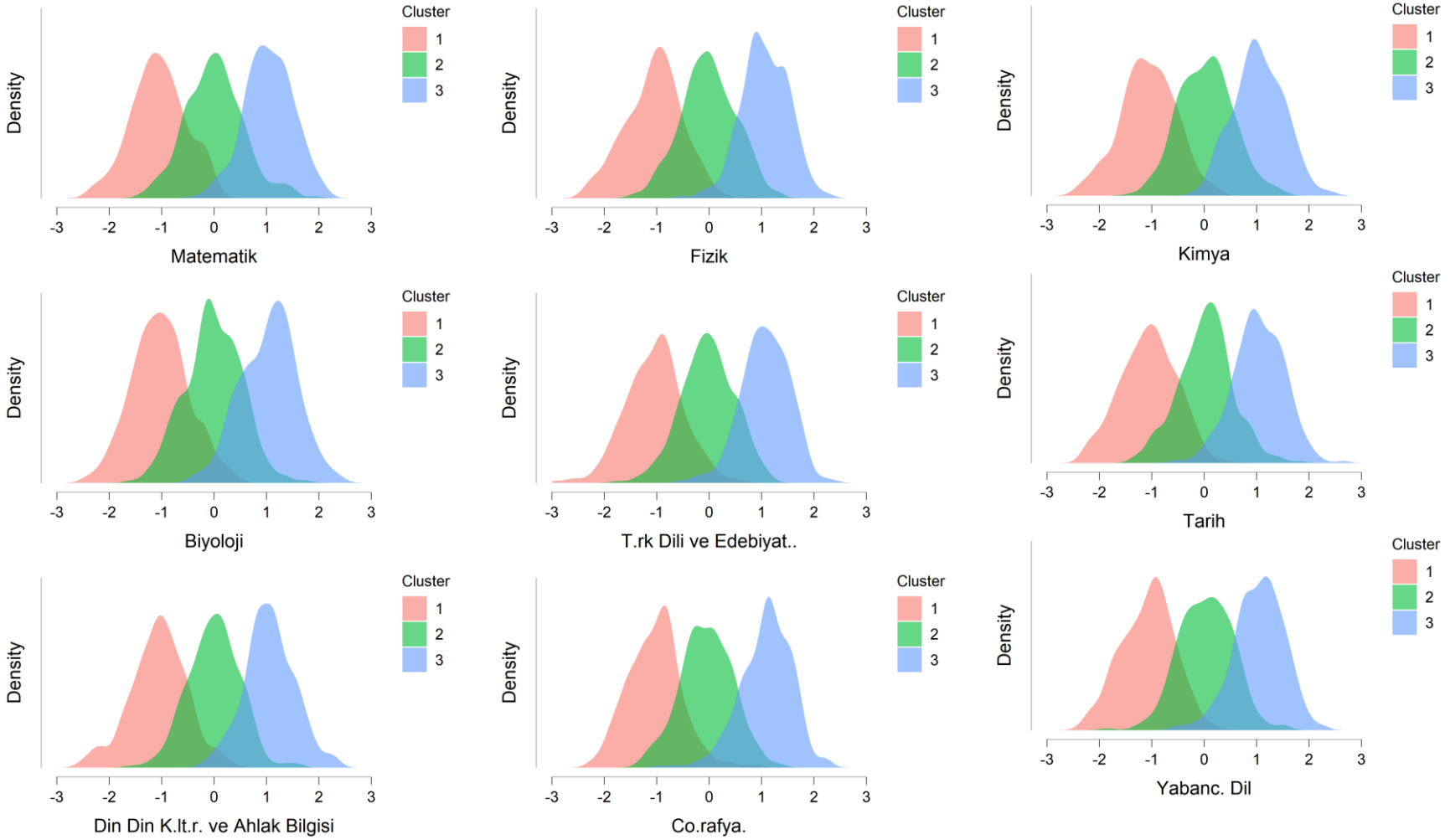
TABLO DEĞERLENDİRMELERİ 6



Standardize edilmiş değerlerin %95 güven aralığı ile grafiksel gösterimini gösterir.

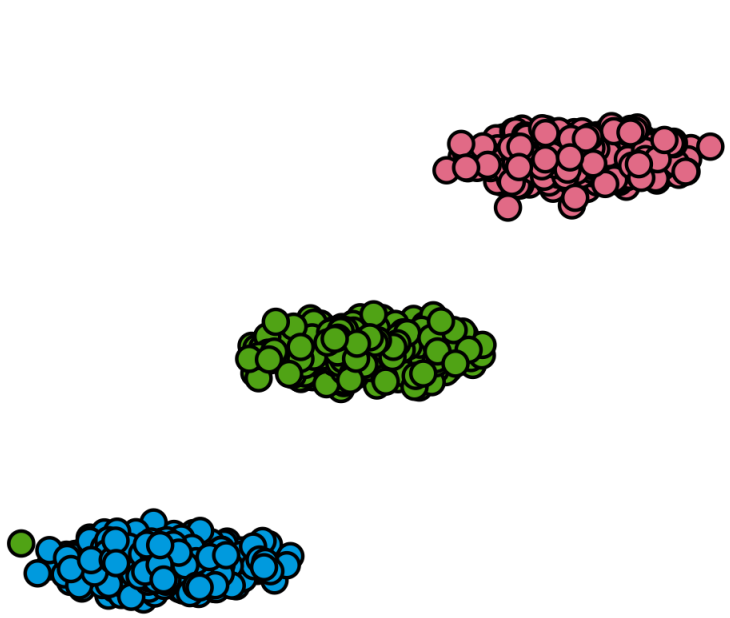
Güven aralığı, istatistik biliminde bir değerin olabileceği optimum sınır olarak belirtilebilir.

TABLO DEĞERLENDİRMELERİ 7



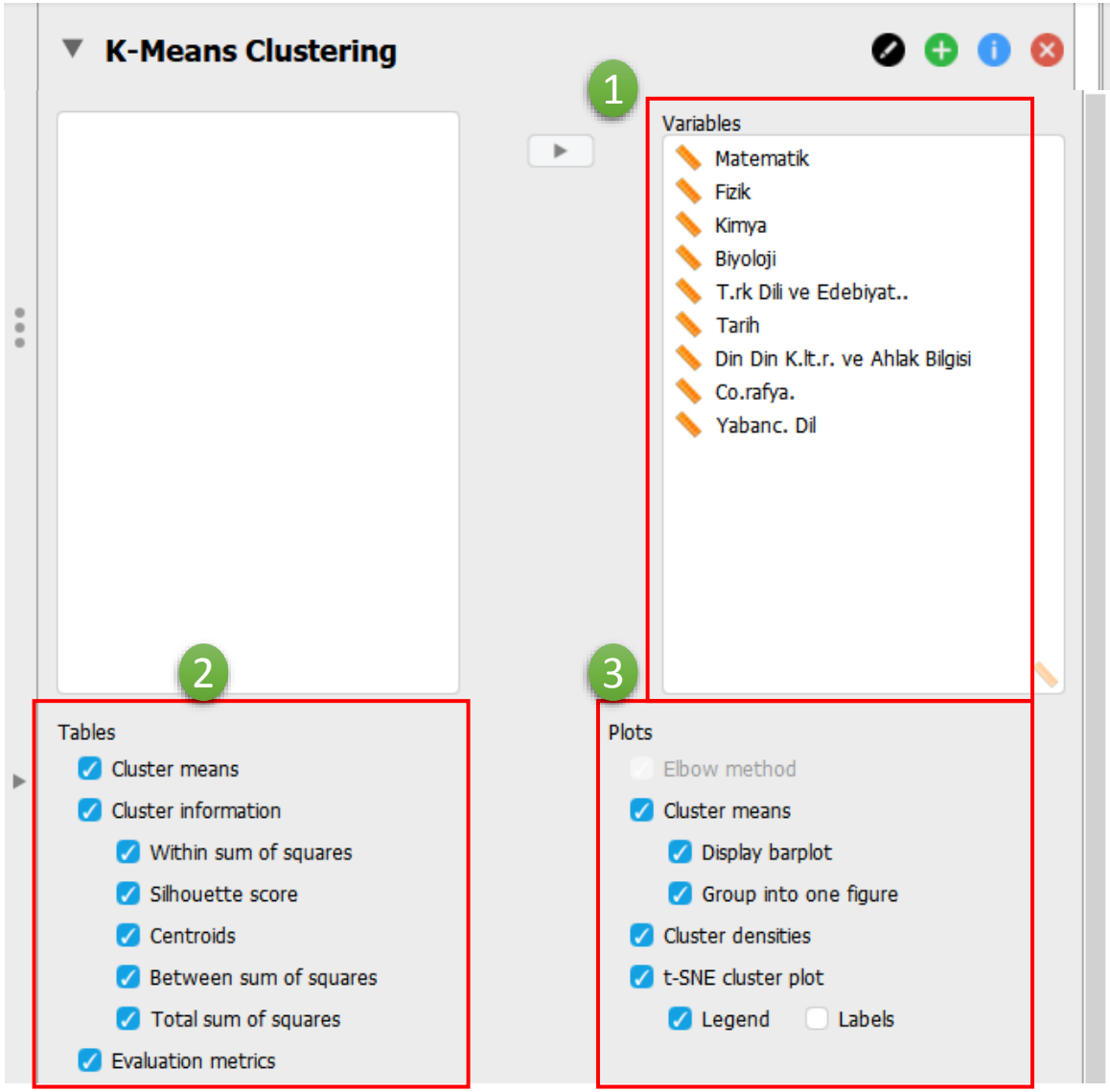
Tüm gruplarda yer alan standardize değerlerin dağılımlarını gösterir.

TABLO DEĞERLENDİRMELERİ 8



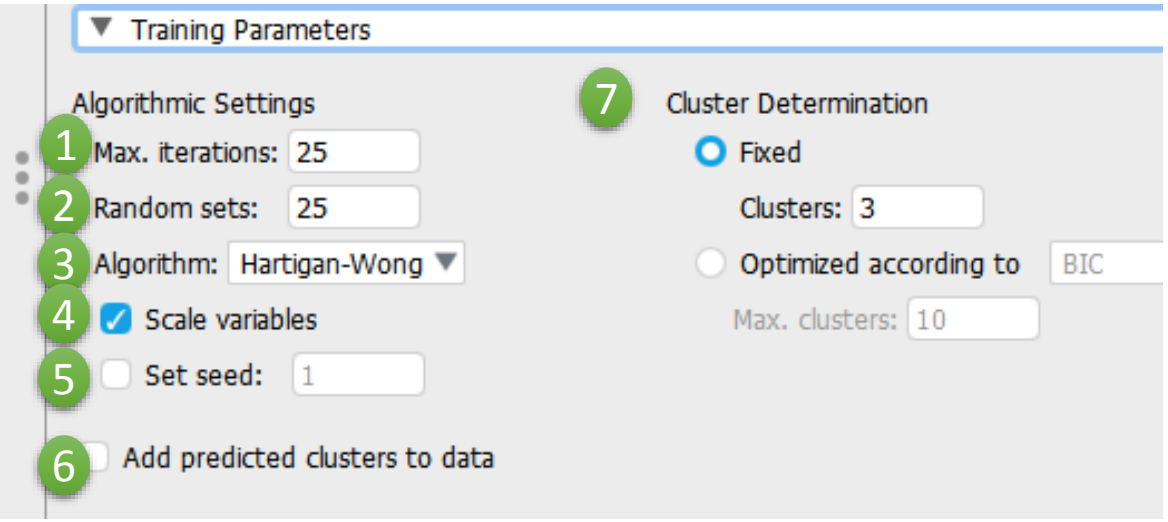
Tüm gruplarda yer alan standardize değerlerin kümelenmesini gösterir.

K-means Clustering



1. **Veriables:** Makine öğrenmesi kümeleme yöntemi için değişken (dersler) bu bölüme atanır.
2. **Tables:** Oluşturmak istediğimiz tabloları belirlediğimiz bölümdür.
   - **Cluster means:** Değişkenlerim tüm kümelerde standardize ortalamaları
   - **Cluster Information:** Kümelerin toplam, grupiçi ve gruplar arası kareler toplamı ve tüm kümeler için siluet değerlerini incelediğimiz tablodur
   - **Evaluation Metrics Table:** Yöntemin performansını değerlendirdiğimiz tabloyu oluşturur.
3. **Plots:** Oluşturmak istediğimiz grafikleri belirlediğimiz bölümdür.
   - **Cluster means:** Her tahmin değişkeni için, her kümenin ortalamasını ve %95 güven aralığını gösteren bir grafik oluşturur.
   - **Cluster densities:** Her tahmin değişkeni için kümeler için örtüşen yoğunlukları gösteren bir grafik oluşturur.
   - **t-SNE cluster plot:** Veri gözlemleri arasındaki göreceli mesafeleri göstermeyi amaçlayan iki boyutlu düşük boyutlu bir uzayda yüksek boyutlu verileri görselleştirmek için kullanılır.

Fuzzy c-means Clustering



1. **Max. iterations:** Maksimum yineleme sayısını ayarlar. Maksimum yineleme sayısı, algoritmanın optimal kümeleme çözümünü bulmak için yinelediği olası örnek sayısını yansıtır. Varsayılan olarak, bu 25 olarak ayarlanmıştır.
2. **Random sets:** Kullanılan maksimum olası rastgele küme sayısını ayarlar. Rastgele kümelerin sayısı, rasgele seçilen kaç tane ilk küme atamasının kullanıldığını yansıtır. Varsayılan olarak, bu 25 olarak ayarlanmıştır.
3. **Algorithm:** Kullanmak istediğiniz algoritmayı seçin. Varsayılan olarak bu, 'Hartigan-Wong' algoritmasına ayarlanmıştır.
4. **Scale variables:** Z-skor standardizasyonunu yapıp yapmayacağımızı belirleyen bölümdür. Standardizasyon yapmak verileri deki aykırı, uç veya gürültülü bireylerin sonuçları etkilememesini sağlar.
5. **Set seed:** Makine öğrenmesi yöntemleri algoritma tabanlı yöntemler olduğu için her çalıştırma sonucunda çıkan sonuçlar farklılık gösterebilir. Bunun için bu bölüm yardımı ile sonuçlar sabitlenir ve her program açıldığında aynı sonuçları görmemiz sağlanır.
6. **Add Predicted Clusters to Data:** Yapılan küme atamalarını veri setinde bir değişken sütunu olarak görmemizi sağlar.
7. **Cluster Determination:** Küme sayısını belirlediğimiz bölümdür.
   - **Fixed:** Sabit miktarda küme oluşturmanıza olanak tanır. Bu, kendi belirtilen sayıda kümenizi oluşturmanıza ve böylece manuel olarak optimize etmenize olanak tanır.
   - **Max. clusters:** Bir optimizasyon yöntemi (AIC,BIC,Silhouette) seçerek bu yönteme göre en uygun küme sayısını belirlememizi sağlar. "Max clusters" bölümü oluşturulacak maksimum küme sayısını belirlememizi sağlar.

TABLO DEĞERLENDİRMELERİ 1

K-Means Clustering

1 Clusters	2 N	3 R^2	4 AIC	5 BIC	6 Silhouette
3	900	0.745	2121.200	2250.860	0.430

1 Elde edilen küme sayısını gösterir. Sonuçlarımıza göre değişkenler üç kümeden oluşmaktadır

2 Toplam birey sayısını gösterir. Veri setimizde 900 birey yer almaktadır.

3 9 dersin kümeleri açıklama oranını gösterir. Buna göre derslerin başarı durumunun %78,6'sını açıkladığı söylenebilir.

4 Model karşılaştırmalarında her zaman en düşük AIC değerini veren model tercih edilir. Burada en düşük AIC değeri 3 küme oluşturulması durumu için elde edilmiştir.

5 AIC'de olduğu gibi mevcut modeller arasında en küçük değerli BIC değerine sahip model, uygun model olarak seçilir. Burada en düşük BIC değeri 3 küme oluşturulması durumu için elde edilmiştir.

6 -1 ve 1 arasında değerler üretmektedir. 1'e en yakın K değeri en uygun olarak belirlenmektedir. Burada en yüksek siluet değeri için küme sayısı 3 olduğu bulunmuştur.

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

TÜBİTAK

4005 TÜBİTAK

TABLO DEĞERLENDİRMELERİ 2

Cluster Information ▼

Cluster	1	2	3
Size	300	300	300
Explained proportion within-cluster heterogeneity	0.358	0.342	0.299
Within sum of squares	741.013	707.839	618.345
Silhouette score	0.425	0.381	0.469
Centroid Matematik	−1.058	0.009	1.048
Centroid Fizik	−1.052	−0.016	1.067
Centroid Kimya	−1.073	0.032	1.040
Centroid Biyoloji	−1.027	−0.019	1.046
Centroid T.rk Dili ve Edebiyat..	−1.044	−0.026	1.070
Centroid Tarih	−1.067	0.023	1.044
Centroid Din Din K.lt.r. ve Ahlak Bilgisi	−1.060	0.001	1.059
Centroid Co.rafya.	−1.038	−0.047	1.086
Centroid Yabanc. Dil	−1.082	0.037	1.044

Note. The Between Sum of Squares of the 3 cluster model is 6023.8

Note. The Total Sum of Squares of the 3 cluster model is 8091

TABLO DEĞERLENDİRMELERİ 3

Cluster Means ▼

	Matematik	Fizik	Kimya	Biyoloji	T.rk Dili ve Edebiyat..	Tarih	Din Din K.lt.r. ve Ahlak Bilgisi	Co.rafya.	Yabanc. Dil
Cluster 1	−1.058	−1.052	−1.073	−1.027	−1.044	−1.067	−1.060	−1.038	−1.082
Cluster 2	0.009	−0.016	0.032	−0.019	−0.026	0.023	0.001	−0.047	0.037
Cluster 3	1.048	1.067	1.040	1.046	1.070	1.044	1.059	1.086	1.044

Burada tüm kümelerde değişkenlerin (ders) standardize edilmiş değerlerinin ortalama farklarının toplamına yer verilmiştir.

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

TÜBİTAK

4005 TÜBİTAK

TABLO DEĞERLENDİRMELERİ 4

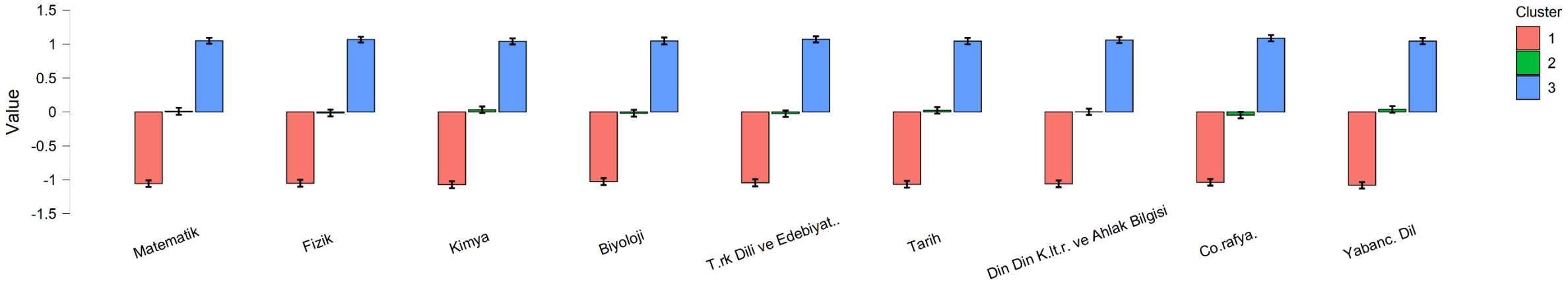
Evaluation Metrics

	Value
Maximum diameter	4.997
Minimum separation	1.154
Pearson's γ	0.701
Dunn index	0.231
Entropy	1.099
Calinski-Harabasz index	1306.927

Note. All metrics are based on the *euclidean* distance.

Evaluation Metrics: Bu tablo kümeleme algoritmasının uyum durumunu (performans metriklerini) verir.
Maximum diameter: Kümelerdeki iki nokta arasındaki maksimum uzaklığı verir.
Minimum separation: İki küme arasındaki münimum mesafeyi verir.
Pearson's γ: Burada iki küme arasındaki ilişkiyi gösterir. Bu değerin sıfıra yaklaşması aynı küme 1 e yaklaşması da farklı küme olduğunu gösterir.
Dunn endeksi: İki kümeyi ayrıştırmamızı sağlayan bir indekstir. Bu değer. Ayrıca bu değer gürültüler için uygun bölgeyi elde etmemizi sağlar.
Entropy: Entropy bireylerin oluşturduğu noktanın düzeni ile ilgili bir değerdir. Düzensizlik sistemin entropisinin artmasına sebep olur.
Calinski-Harabasz index: Varyans oranı kriteri olarak da bilinir. Bir nesnenin diğer kümelere kıyasla kendi kümesine ne kadar uyumlu olduğunun bir ölçüsüdür. Burada uyum, bir kümedeki veri noktalarından küme merkezine olan mesafelere göre tahmin edilir ve ayırma, küme merkezlerinin küresel merkeze olan mesafesine dayanır.

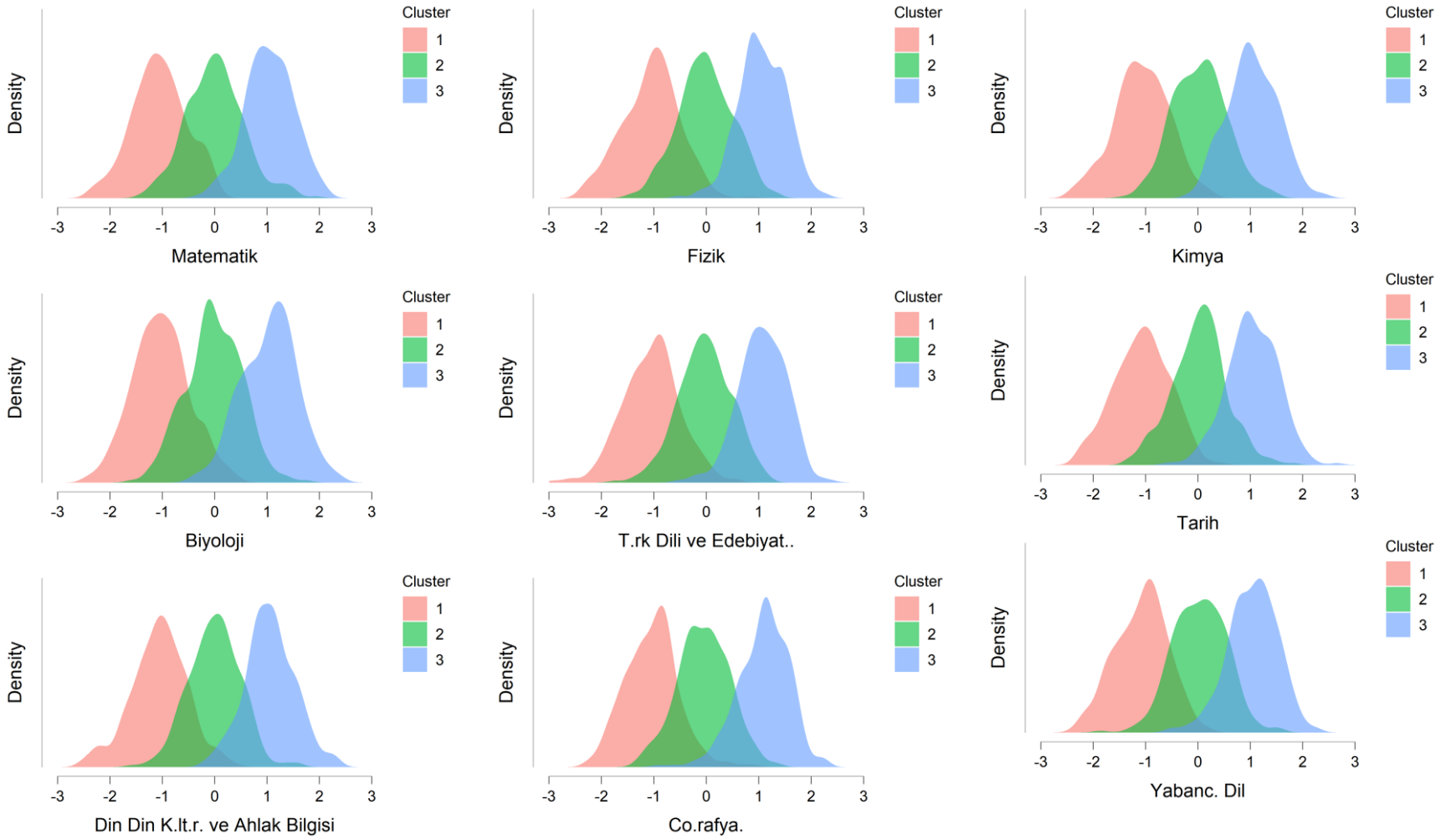
TABLO DEĞERLENDİRMELERİ 5



Standardize edilmiş değerlerin %95 güven aralığı ile grafiksel gösterimini gösterir.

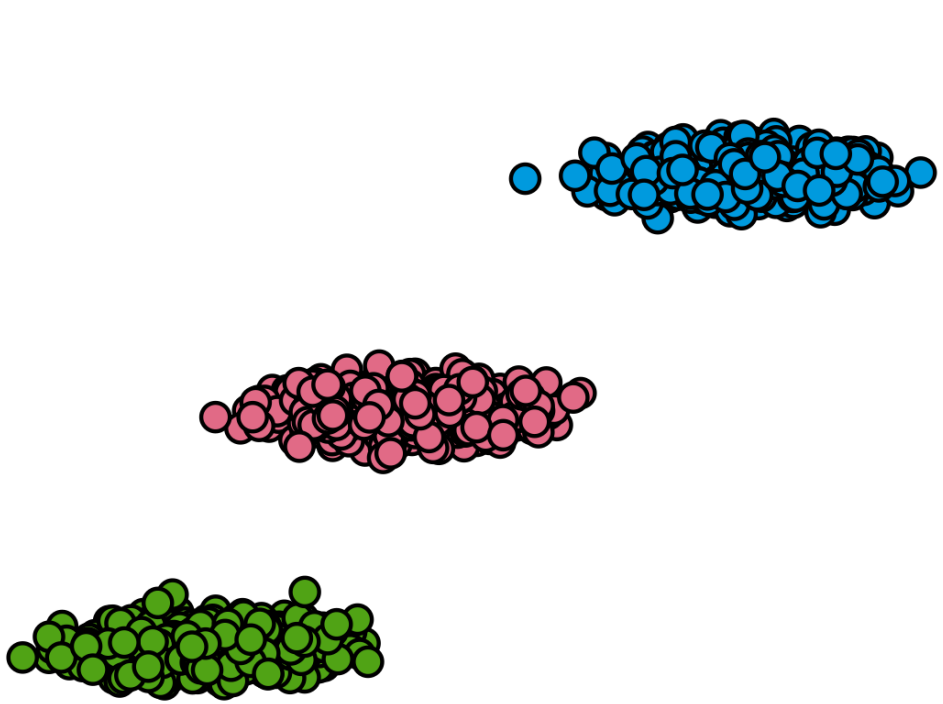
Güven aralığı, istatistik biliminde bir değerin olabileceği optimum sınır olarak belirtilebilir.

TABLO DEĞERLENDİRMELERİ 6



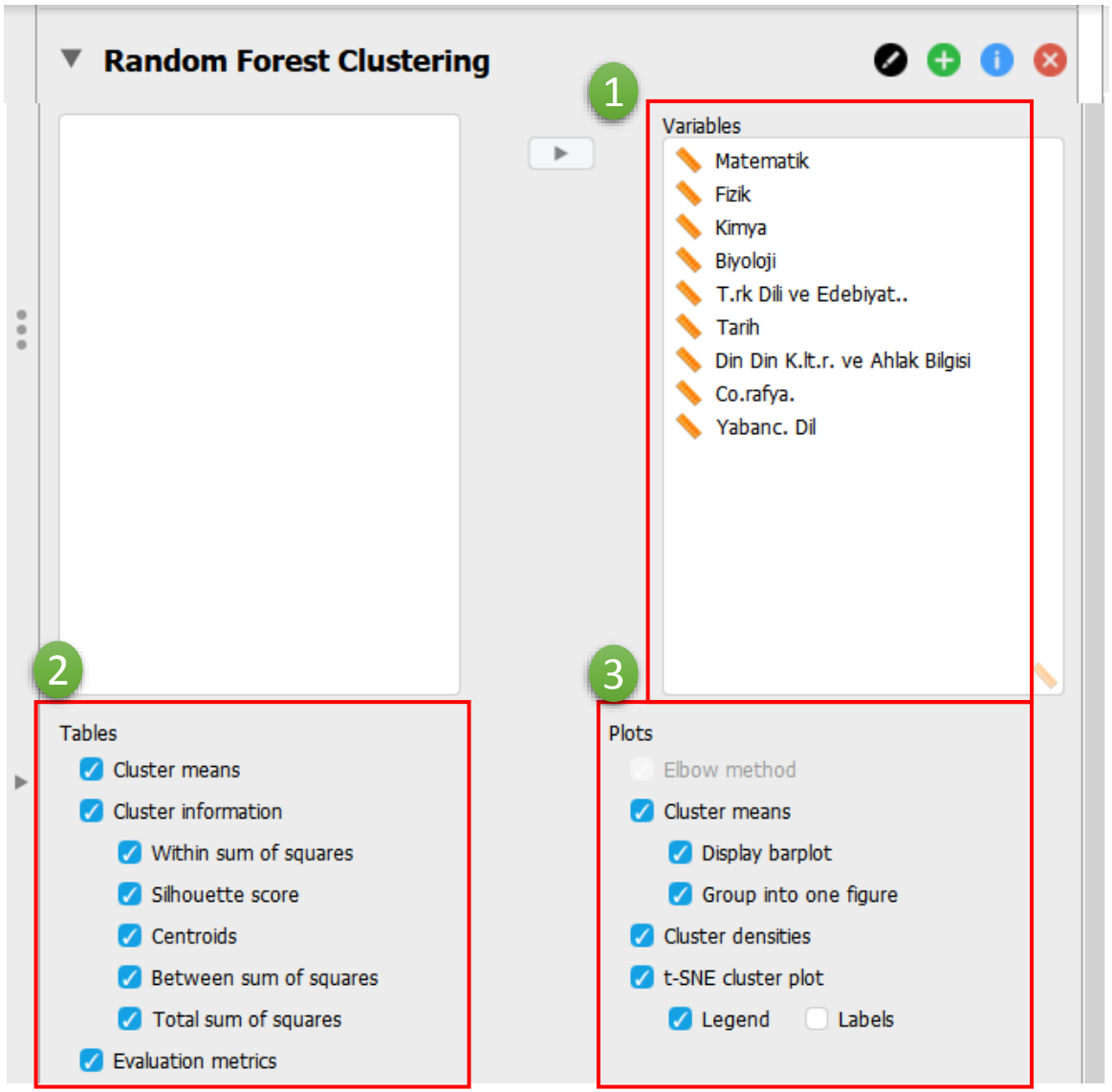
Tüm gruplarda yer alan standardize değerlerin dağılımlarını gösterir.

TABLO DEĞERLENDİRMELERİ 7



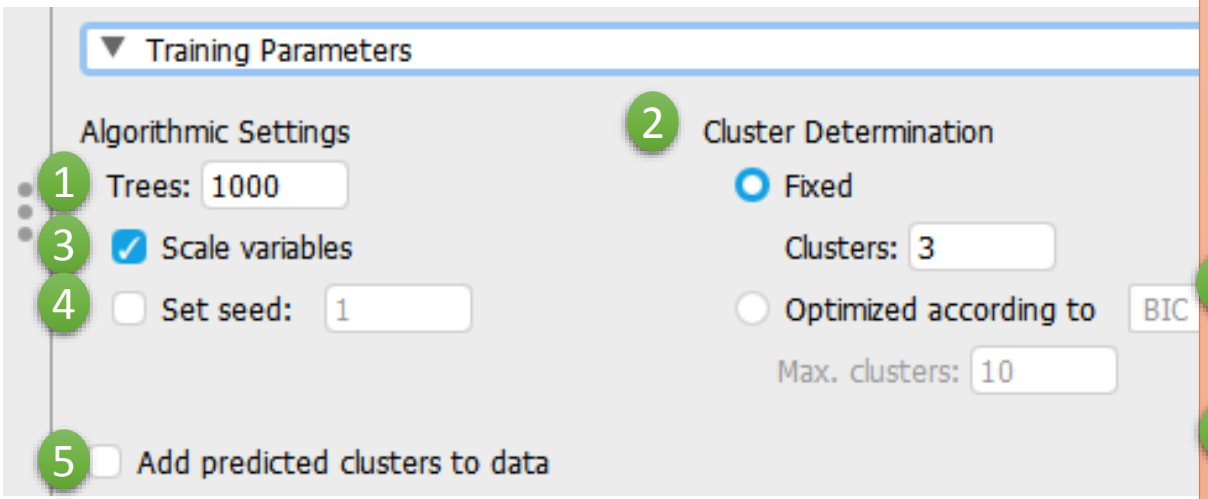
Tüm gruplarda yer alan standardize değerlerin kümelenmesini gösterir.

Random Forest Clustering



1. **Veriables:** Makine öğrenmesi kümeleme yöntemi için değişken (dersler) bu bölüme atanır.
2. **Tables:** Oluşturmak istediğimiz tabloları belirlediğimiz bölümdür.
   - **Cluster means:** Değişkenlerim tüm kümelerde standardize ortalamaları
   - **Cluster Information:** Kümelerin toplam, grupiçi ve gruplar arası kareler toplamı ve tüm kümeler için siluet değerlerini incelediğimiz tablodur
   - **Evaluation Metrics Table:** Yöntemin performansını değerlendirdiğimiz tabloyu oluşturur.
3. **Plots:** Oluşturmak istediğimiz grafikleri belirlediğimiz bölümdür.
   - **Cluster means:** Her tahmin değişkeni için, her kümenin ortalamasını ve %95 güven aralığını gösteren bir grafik oluşturur.
   - **Cluster densities:** Her tahmin değişkeni için kümeler için örtüşen yoğunlukları gösteren bir grafik oluşturur.
   - **t-SNE cluster plot:** Veri gözlemleri arasındaki göreceli mesafeleri göstermeyi amaçlayan iki boyutlu düşük boyutlu bir uzayda yüksek boyutlu verileri görselleştirmek için kullanılır.

Random Forest Clustering



1. **Trees:** Oluşturulacak maksimum ağaç sayısını gösterir
2. **Cluster Determination:** Küme sayısını belirlediğimiz bölümdür.
   Fixed: Sabit miktarda küme oluşturmanıza olanak tanır. Bu, kendi belirtilen sayıda kümenizi oluşturmanıza ve böylece manuel olarak optimize etmenize olanak tanır.
   Max. clusters: Bir optimizasyon yöntemi (AIC,BIC,Silhouette) seçerek bu yönteme göre en uygun küme sayısını belirlememizi sağlar. "Max clusters" bölümü oluşturulacak maksimum küme sayısını belirlememizi sağlar.
3. **Scale variables:** Z-skor standardizasyonunu yapıp yapmayacağımızı belirleyen bölümdür. Standardizasyon yapmak verileri deki aykırı, uç veya gürültülü bireylerin sonuçları etkilememesini sağlar.
4. **Set seed:** Makine öğrenmesi yöntemleri algoritma tabanlı yöntemler olduğu için her çalıştırma sonucunda çıkan sonuçlar farklılık gösterebilir. Bunun için bu bölüm yardımı ile sonuçlar sabitlenir ve her program açıldığında aynı sonuçları görmemiz sağlanır.
5. **Add Predicted Clusters to Data:** Yapılan küme atamalarını veri setinde bir değişken sütunu olarak görmemizi sağlar.

TABLO DEĞERLENDİRMELERİ 1

Random Forest Clustering

1 Clusters	2 N	3 R^2	4 AIC	5 BIC	6 Silhouette
3	900	0.742	2138.800	2268.460	0.420

1 Elde edilen küme sayısını gösterir. Sonuçlarımıza göre değişkenler üç kümeden oluşmaktadır

2 Toplam birey sayısını gösterir. Veri setimizde 900 birey yer almaktadır.

3 9 dersin kümeleri açıklama oranını gösterir. Buna göre derslerin başarı durumunun %78,6'sını açıkladığı söylenebilir.

4 Model karşılaştırmalarında her zaman en düşük AIC değerini veren model tercih edilir. Burada en düşük AIC değeri 3 küme oluşturulması durumu için elde edilmiştir.

5 AIC'de olduğu gibi mevcut modeller arasında en küçük değerli BIC değerine sahip model, uygun model olarak seçilir. Burada en düşük BIC değeri 3 küme oluşturulması durumu için elde edilmiştir.

6 -1 ve 1 arasında değerler üretmektedir. 1'e en yakın K değeri en uygun olarak belirlenmektedir. Burada en yüksek siluet değeri için küme sayısı 3 olduğu bulunmuştur.

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

TÜBİTAK

4005 TÜBİTAK

TABLO DEĞERLENDİRMELERİ 2

Cluster Information ▼

Cluster	1	2	3
Size	300	302	298
Explained proportion within-cluster heterogeneity	0.356	0.354	0.291
Within sum of squares	741.315	737.822	605.660
Silhouette score	0.426	0.372	0.473

Note. The Between Sum of Squares of the 3 cluster model is 6006.2

Note. The Total Sum of Squares of the 3 cluster model is 8091

Size: Her kümedeki birey sayısını gösterir.
Explained proportion within-cluster heterogeneity: Her kümenin heterojenlik varyasyon oranlarını verir
Total, Between, Within (Sum of Square): Verideki ortalamaya göre toplam değişimi, verinin tüm gruplardaki toplam değişimini ve verilerin gruplar içindeki toplam değişimini verir.
Silhouette Scoure: Tum gruplardaki Siluet skorlarını verir. Bu değerler bire ne kadar yakın ise o kadar iyi atama yapıldığını ifade eder.

TABLO DEĞERLENDİRMELERİ 3

Cluster Means

	Matematik	Fizik	Kimya	Biyoloji	T.rk Dili ve Edebiyat..	Tarih	Din Din K.lt.r. ve Ahlak Bilgisi	Co.rafya.	Yabanc. Dil
Cluster 1	−1.061	−1.054	−1.067	−1.027	−1.044	−1.064	−1.052	−1.035	−1.079
Cluster 2	0.019	−0.006	0.029	−0.008	−0.026	0.027	−0.008	−0.043	0.035
Cluster 3	1.049	1.067	1.044	1.043	1.077	1.044	1.066	1.086	1.051

Burada tüm kümelerde değişkenlerin (ders) standardize edilmiş değerlerinin ortalama farklarının toplamına yer verilmiştir.

TABLO DEĞERLENDİRMELERİ 3

Evaluation Metrics

	Value
Maximum diameter	5.675
Minimum separation	1.154
Pearson's γ	0.699
Dunn index	0.203
Entropy	1.099
Calinski-Harabasz index	1292.107

Note. All metrics are based on the *euclidean* distance.

Evaluation Metrics: Bu tablo kümeleme algoritmasının uyum durumunu (performans metriklerini) verir.

Maximum diameter: Kümelerdeki iki nokta arasındaki maksimum uzaklığı verir.

Minimum separation: İki küme arasındaki münimum mesafeyi verir.

Pearson's γ: Burada iki küme arasındaki ilişkiyi gösterir. Bu değerin sıfıra yaklaşması aynı küme 1 e yaklaşması da farklı küme olduğunu gösterir.

Dunn endeksi: İki kümeyi ayrıştırmamızı sağlayan bir indekstir. Bu değer. Ayrıca bu değer gürültüler için uygun bölgeyi elde etmemizi sağlar.

Entropy: Entropy bireylerin oluşturduğu noktanın düzeni ile ilgili bir değerdir. Düzensizlik sistemin entropisinin artmasına sebep olur.

Calinski-Harabasz index: Varyans oranı kriteri olarak da bilinir. Bir nesnenin diğer kümelere kıyasla kendi kümesine ne kadar uyumlu olduğunun bir ölçüsüdür. Burada uyum, bir kümedeki veri noktalarından küme merkezine olan mesafelere göre tahmin edilir ve ayırma, küme merkezlerinin küresel merkeze olan mesafesine dayanır.

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

TÜBİTAK

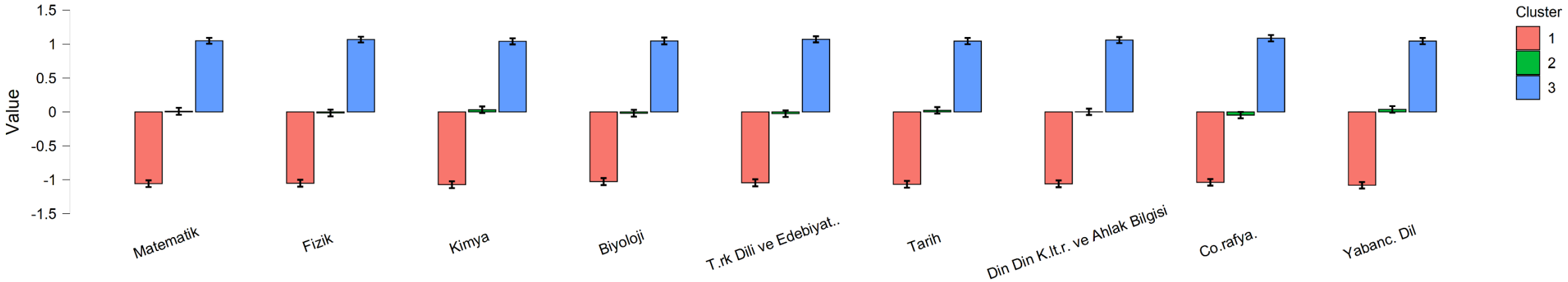
4005 TÜBİTAK

TABLO DEĞERLENDİRMELERİ 4

Variable Importance

	Mean decrease in Gini Index
Din Din K.lt.r. ve Ahlak Bilgisi	107.491
T.rk Dili ve Edebiyat..	102.923
Yabanc. Dil	102.801
Kimya	102.214
Co.rafya.	101.242
Matematik	100.168
Tarih	96.762
Fizik	95.932
Biyoloji	89.909

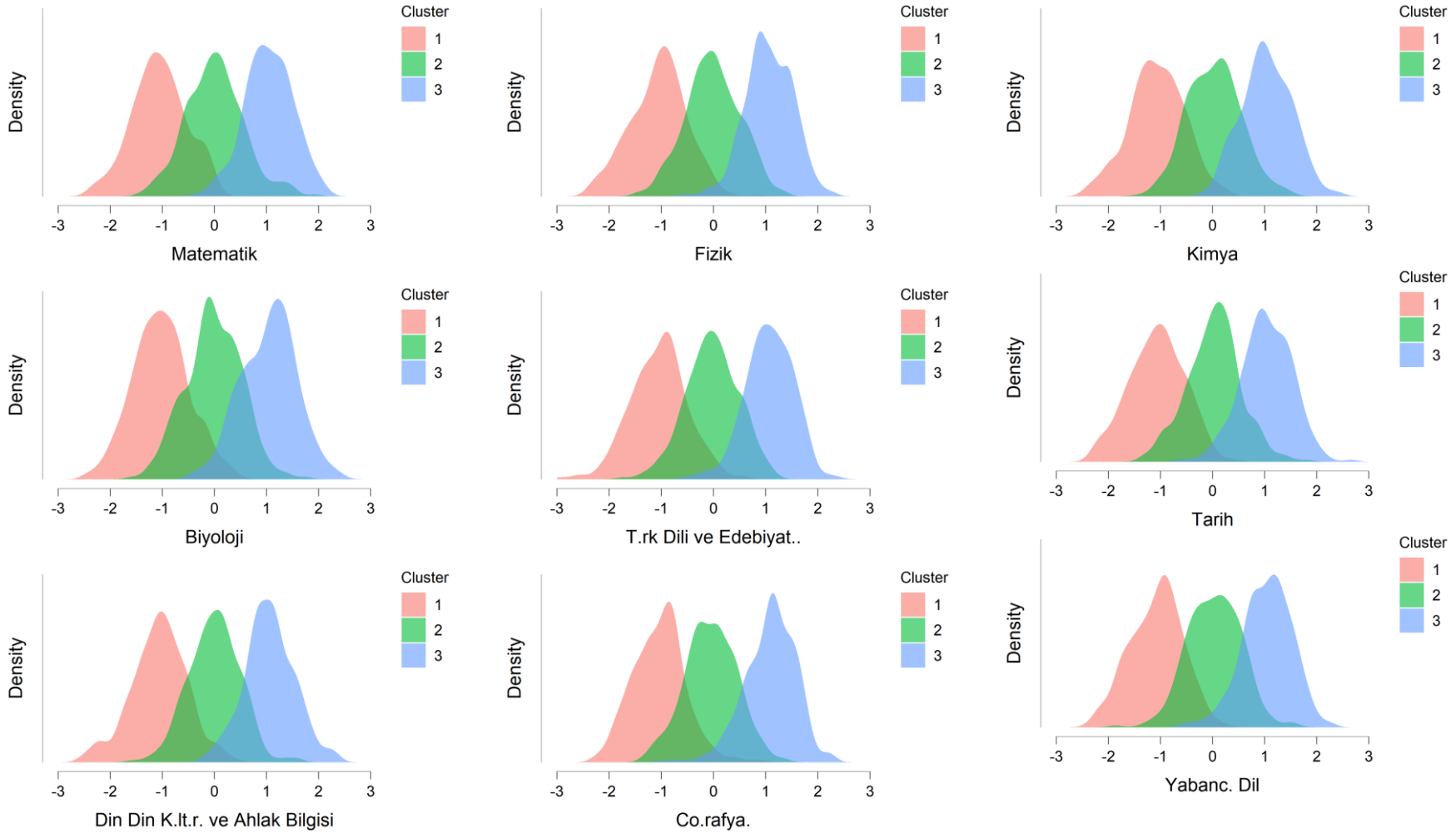
TABLO DEĞERLENDİRMELERİ 5



Standardize edilmiş değerlerin %95 güven aralığı ile grafiksel gösterimini gösterir.

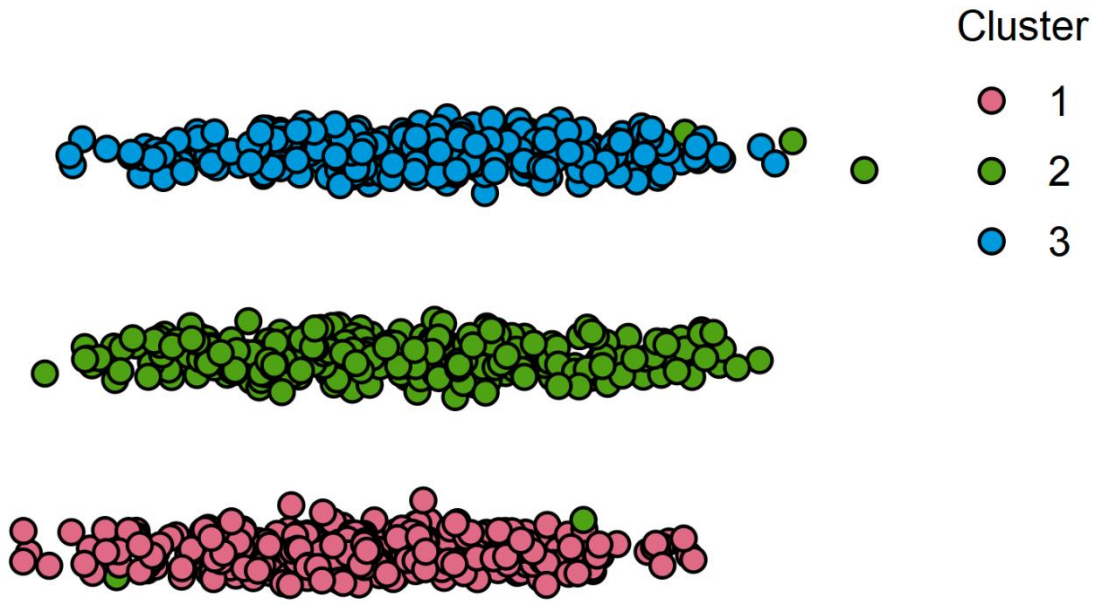
Güven aralığı, istatistik biliminde bir değerin olabileceği optimum sınır olarak belirtilebilir.

TABLO DEĞERLENDİRMELERİ 6



Tüm gruplarda yer alan standardize değerlerin dağılımlarını gösterir.

TABLO DEĞERLENDİRMELERİ 7



Tüm gruplarda yer alan standardize değerlerin kümelenmesini gösterir.

KÜMELEME VE KURAL TABANLI ALGORİTMALAR

DOÇ. DR. MUSTAFA AGÂH TEKİNDAL
PROF. DR. FERHAN ELMALI
ÖĞR.GÖR. DR. BERHAN ÇOBAN

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

TÜBİTAK

4005 TÜBİTAK