T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

BENZERLİK ÖLÇÜLERİ ve KÜMELEME ANALİZİ

Fatih FIRAT

Matematik Anabilim Dalı

Matematik Programı

DANIŞMAN
Yrd. Doç. Dr. Mehmet CEVRİ

Haziran, 2016

İSTANBUL

Bu çalışma 30.06.2016 tarihinde aşağıdaki jüri tarafından Matematik Anabilim Dalı Matematik Programında Yüksek Lisans Tezi olarak kabul edilmiştir.

Tez Jürisi:

Yrd. Doç. Dr. Mehmet CEVRİ(Danışman)
İstanbul Üniversitesi
Fen Fakültesi

Prof. Dr. Kamuran SAYGILI
İstanbul Üniversitesi
Fen Fakültesi

Prof. Dr. İ. Müfit GİRESUNLU
İstanbul Üniversitesi
Fen Fakültesi

Prof. Dr. N. Füsun OYMAN SERTELLER
Marmara Üniversitesi
Teknoloji Fakültesi

Doç. Dr. Ayten Pekin
İstanbul Üniversitesi
Fen Fakültesi

ÖNSÖZ

Bu tez çalışması boyunca yönlendirmeleri ve bilgi birikimi ile her konuda bana destek olan danışmanım Yard. Doç. Dr. Mehmet CEVRİ'ye teşekkürlerimi sunarım.

Araştırmaları sonucu elde ettiği bilgileri paylaşan İstanbul Üniversitesi Eczacılık Fakültesi öğretim üyelerinden olan Yard. Doç. Dr. Gülay ECEVİT GENÇ'e desteğinden dolayı teşekkürlerimi borç bilirim.

Haziran, 2016 Fatih FIRAT

İÇİNDEKİLER

ŞEKİL LİSTESİ

TABLO LİSTESİ

SİMGE VE KISALTMA LİSTESİ

Simgeler		Açıklama
$\mathbb{F}$	:	Bütün eşleşmelerin toplam sayısı
d_{ij}	:	i ve j nesneleri arasındaki uzaklık ölçüsü
s_{ij}	:	i ve j nesneleri arasındaki benzerlik ölçüsü
r	:	Korelasyon katsayısı
w_{ij}	:	Kısmi korelasyon katsayısı
VIF_j	:	Varyans şişirme değeri
α_{ji}	:	j'nci değişkenin i'nci faktör yükü
F_i	:	i'nci faktör
ε_j	:	Artık faktörler
z_j	:	Gizli faktörler

Kısaltmalar		Açıklama
KMO	:	Kaiser-Meyer-Olkin örneklem yeterliliği ölçüsü
PCA	:	Temel bileşenler analizi
PCoA	:	Temel koordinatlar analizi
ESS	:	Hata kareler toplamı
UPGMA	:	Aritmetik ortalama kullanılarak ağırlıksız grup ortalaması metodu

ÖZET

YÜKSEK LİSANS TEZİ

BENZERLİK ÖLÇÜLERİ VE KÜMELEME ANALİZİ

Fatih FIRAT

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Matematik Anabilim Dalı

Danışman: Yard.Doç.Dr. Mehmet CEVRİ

Güncel hayatta karşılaşılan çoğu problem iki veya daha fazla değişken içermektedir. Bu değişkenler arasındaki ilişkinin tespiti çok önemlidir. Bunun için kümeleme analizi, faktör analizi ve temel bileşenler analizi gibi çok değişkenli istatistiksel veri analizi teknikleri kullanılmaktadır. Dolayısıyla, bu tezin amacı bahsedilen teknikleri detaylı olarak irdelemektir. İncelenen bu istatistiki yöntemler Teucrium türüne ait verilere uygulanarak, bu türler bazı karakterlere göre sınıflandırılmaya çalışılmıştır.

Bu çalışma yedi ana bölümden oluşmaktadır. Birinci bölümde tezin konusu ile ilgili alt yapı verilmiştir. İkinci bölümde literatürde sıkça kullanılan klasik benzerlik ölçüleri tanıtılmıştır. Üçüncü bölümde, faktör analizi hakkında kısaca bilgi verilmiştir. Dördüncü bölümde, kümeleme analizi irdelenmiştir. Ayrıca, kümeleme analizi sonucu oluşan kümelerin şematik olarak gösterilimi incelenmiştir. gösterilme yöntemleri incelenmiştir. Beşinci bölümde, teoriksel bilgilerin uygulaması yapılmadan önce, uygulama için gerekli olan malzeme ve yöntem hakkında kısa bir bilgi verilmiştir. Altıncı bölümde, benzerlik yöntemlerinin yapay ve gerçek dünya problemi verileri üzerine uygulaması yapılmıştır. Yedinci bölümde, uygulamadan elde edilen sonuçlar ve ileriye dönük çalışmalar hakkında bilgi verilmiştir.

Haziran 2016, 139 sayfa.

Anahtar kelimeler: Benzerlik/uzaklık, kümeleme analizi, faktör analizi, sınıflandırma, anlamsal ilişki.

SUMMARY

M. Sc. THESIS

SIMILARITY MEASURES AND CLUSTER ANALYSIS

Fatih FIRAT

İstanbul University

Institute of Graduate Studies in Science and Engineering

Department of Mathematics

Supervisor: Asst.Prof.Dr. Mehmet CEVRİ

Most problems encountered in daily life contain two or more variables. Determining the relationship between these variables is very important. For this, multivariate statistical data analysis techniques such as cluster analysis, factor analysis and principal component analysis are used. Therefore, the objective of this thesis is to analysis the mentioned techniques in detail. In addition, the examined these statistical techniques are applied data of Teucrium species that microbiological taxa and these species were classified based on their characters.

This study consists of seven main sections. In the first section, it is given to infrastructure related to the topic of the thesis. In the second section, classical similarity measure commonly used in the literature are introduced. In the third section, brief information about the factor analysis are given. In the fourth section, cluster analysis are examined. In addition, clusters formed as a result of cluster analysis was examined schematic illustration. In the fifth section, before making the application of theoretical information, a brief information has been given about the methods and materials required for the application. In the sixth section, the application of similarity methods on artification and real-world problems are reviewed. In the seventh section, it is given that information about prospective study and the results obtained from the application.

June 2016, 139 pages.

Keywords: Similarity/distance, cluster analysis, factor analysis, classification, meaning relate.

1. GİRİŞ

İlk çağlardan günümüze kadar olan süreçte insanoğlu bazı keşiflerde bulunmuştur. Buldukları bu keşifler sonucunda gün geçtikçe nesnelerin sayısı artmış ve bunları birbirinden farklı kılmak ya da ayırt etmek gibi ihtiyaçlar ortaya çıkmıştır. Bu ihtiyaçları ilk çağlarda gidermek pek zor olmamıştır. Ancak, günümüzde nesnelerin sayısının oldukça fazla ve karmaşık halde olmasından dolayı, nesneleri sınıflandırma ihtiyacı önemli bir durum haline gelmiş ve bununla birlikte bir takım sınıflandırma yöntemleri geliştirilmiştir [1]. Bu tezde, geliştirilen sınıflandırma yöntemleri içerisinde yer alan faktör analizi ve kümeleme analizine odaklanıldı.

Bunlardan birincisi olan faktör analizi, karmaşık bir vaziyette olan ve birbirleriyle ilişkili olan nesneleri daha basit ve birbiriyle ilişkili olmayan nesnelerle ifade etmeye yarayan çok değişkenli istatistiksel analiz yöntemlerinden biridir. Ayrıca, faktör analizinin diğer istatistiksel analiz yöntemlerine bazı kolaylıklar sağlamak gibi bir amacı da bulunmaktadır. İkinci yöntem ise, "kümeleme analizi"dir. Kümeleme analizi, veri tabanlarında bulunan nesneleri belli kriterler çerçevesinde veya araştırmacının amaçlarına göre seçilen benzerlik veya uzaklık ölçüleri doğrultusunda gruplayan istatistiksel analiz yöntemidir.

Bu iki yöntem günümüzde kendine pek çok uygulama alanı bulmuştur. Özellikle, bu yöntemler büyümekte olan veya büyük çaplı şirketlerin stratejik çalışmalarında önemli bir rol oynamaktadır. Aynı zamanda, bu yöntemler eczacılık, biyoloji, genetik, tıp, arkeoloji, ekonomi, bankacılık, kriptoloji, psikoloji gibi alanlarda da kullanılmaktadır [2]. Bu teknikler günümüzde yukarıda bahsedilen alanlarda kullanılsa da, aslında faktör analizi ve kümeleme analizinin tarihi çok eskilere dayanmaktadır. Bu tarihsel süreçlerle ile ilgili bilgiler aşağıda kısaca verilmiştir.

Önemli Yunan bilginlerinden Hipokrat hayvan türlerinin sayısını bulmakla ilgilenmiş ve daha sonrasında bu hayvanların sınıflandırılması üzerinde çalışmıştır. Diğer önemli bilginlerden biri olan Aristoteles ise, yaşam şartlarının sınıflandırılması üzerinde çalışmıştır.

1602 yılında böcek türlerinin sınıflandırılması üzerine Ulisse Aldrovandi ilk derlemesini ortaya koymuştur. Ulisse Aldrovandi'nin çalışmalarını ortaya koyduğu dönemler, hayvanlar üzerinde çalışmaların yapıldığı dönemlerdi ve bu dönemlerde öncelikle ailesel türlerin sınıflandırılmasına yönelmişlerdir. Bundan dolayı, organların benzerliği ile başlayan sınıflandırma anatomik benzerliklere dayandırılmıştır.

Daha sonraki dönemlerde "*Historia Plantarum*" adlı eseriyle bitkiler, hayvanlar ve din teorisi üzerinde ortaya önemli bir yapıt koyan John Ray, modern taksonominin gelişmesine bitkilerin sınıflandırılması ile katkı sağlamıştır [3].

20. yüzyılın başlarında Charles Spearman psikoloji alanında insanların davranışlarını ve yeteneklerini açıklamak için matematiksel model olarak faktör analizini kullanmıştır [4].

1963 yılında Sokal ve Sneath tarafından yayımlanan "*Principles of Numerical Taxonomy*" adlı kitapla kümeleme analizinin gelişmesinde önemli bir adım atılmıştır [2].

1974 yılında Everitt ilk kümelendirme çalışmasını zooloji ve biyoloji üzerinde yapmıştır [2]. Bundan bir yıl sonra ise Cox, 43 sert kış buğdayının genetik ilişkilerinin nasıl olduğunu kümeleme analizi vasıtasıyla ortaya koymuştur [2].

1975 yılında Child "*The Essentials of Factor Analysis*" adlı kitabında insan yetenekleri, psikoloji, tıp ve halk sağlığı ile ilgili sorunları faktör analizi ile incelemiştir. Bu çalışmaları sonucunda, faktör analizinin önemli ve etkili bir yol olduğunu ortaya çıkarmıştır. Ayrıca, faktör analizinin çok değişkenli araştırmalar için

açıklayıcılığının yeterli düzeyde olduğu bilgisine de ulaşmıştır [5].

1979 yılında Koç ve Ülkü, eğitim ve öğretim ile ilgili testlerin uygulamasında faktör analizinin boyut indirgeme özelliği üzerinde durmuşlardır [6].

1994 yılında Kline, sürekli sakinleştirici ilaç kullanan hastaların sakinleştirici ilaç kullanmalarının sebebinin ne olduğunu faktör analizi ile tespit etmiştir. Kline'nın bu araştırmasında, sakinleştirici ilaç kullanma sebepleri arasında hastaların kaygılarının olduğu bilgisine ulaşılmıştır [7].

1997 yılında He ve arkadaşları, Hawaii adalarında 1991 ile 1994 yılları arasındaki dip trolü ile yapmış oldukları çalışmada elde ettikleri verilere kümeleme analizi uygulamış ve birim çabada harcanan güce göre trolden çıkan balık türlerini sınıflandırmışlardır [8].

1998 yılında Maes ve arkadaşları, Belçika'nın Doel nükleer santralinin suyunu bıraktığı bölgede 1994 ile 1995 yılları arasında 1 yıl boyunca yapmış oldukları çalışmada 2 karides, 55 balık 4 yengeç türü yakalamışlardır. Bu türlerin mevsimsel değişimlerini incelemek için temel bileşenler analizini uygulamışlardır. Temel bileşenler analizi sonucunda türleri beş grupta sınıflandırmışlar ve buradaki balıkların 36 tanesinin deniz balığı, 16 tanesinin tatlı su balığı olduğunu tespit etmişlerdir [8].

2000 yılında Araujo ve Santos, Brezilya'nın Lajes rezervuarlarında bulunan 15 balık türü tespit edip dağılımlarını incelemişlerdir. Kümeleme analizlerini kullanarak gruplamışlar ve temel bileşenler analizi ile en çok çıkan türün Loricariichthys Spixii olduğu ve toplam biomasın %80'ni oluşturduğunu tespit etmişlerdir [8].

2001 yılında Albertelli ve arkadaşları, makrobentozdaki sıcaklık dalgalanmasını aylık olarak tespit etmişler ve kümeleme analizini kullanarak gruplara ayırmışlardır [8].

2004 yılında Tekin; 1987 ile 1996 yılları arasında illerin, bir ülkenin ekonomik

büyüklüğünün ölçütlerinden biri olan gayri safi yurt içi hasılasına katkısını faktörlerle ayırmıştır. Tekin çalışmaları doğrultusunda hizmet, sanayi, tarım faktörü olmak üzere üç faktör elde etmiştir. Hizmet, sanayi, tarım faktörlerine göre sırasıyla en çok katkı sağlayan illerin İstanbul, Kocaeli, Çanakkale olduğu bilgisine ulaşmıştır. Bununla birlikte ise, en az katkı sırasıyla Tunceli, Aydın, İstanbul illerinin olduğunu ortaya koymuştur [9].

2005 yılında Hernandez, Kanarya adalarının farklı bölgelerinden toplanan bal örneklerinin özellikleri ve sınıflandırılmasında çeşitli çok değişkenli istatistik tekniklerini kullanmıştır. Temel bileşen ve kümeleme analizi ile alınan örneklerin Na, K, Sr, Mg, Ca ve Cu içerikleri açısından karakterize edilebildiğini açıklamıştır [8].

2005 yılında Voncina ve arkadaşları, yapmış oldukları çalışmada yağ asit içeriklerinin karakterize edilmesi ile bitkisel yağların sınıflandırılmasında ayırma ve temel bileşenler analizinden yararlanmışlardır. Çalışmada 7 element için 132 yağ örneğini incelemiş ve ilk 2 temel bileşen ile toplam varyansın $\%97,8$'inin açıklandığını ortaya çıkarmışlardır [8].

2006 yılında Pierce ve arkadaşları, yapmış oldukları araştırmalarında inceledikleri bitki örneklerine ait kimyasal farklılıkları ortaya koymak amacıyla temel bileşenler analizinden yararlanmışlardır. Analizi sonucunda incelenen kimyasal özelliklere ait toplam varyasyonun $\%78,6$'sının ilk 2 bileşen tarafından açıklandığını ortaya koymuşlardır [8].

2007 yılında Shanmugan ve Johnson, çalışmalarında 45 ülkedeki çevresel faktörlerin kansere ne oranda etkili olduğunu test etmek için temel bileşenler ve çevresel veri analizini uygulamışlardır. Temel bileşenler analizi sonunda birinci bileşenin toplam varyasyonun $\%95.53$'ünü açıkladığını tespit etmişlerdir [8].

2010 yılında Karabayır ve Doğanay, kümeleme analizi ile portföy seçimi üzerine çalışmışlardır. Bu çalışmalarında hisse senedi alan yatırımcıların kümeleme analizi vasıtasıyla nasıl rasyonel yatırım yapacaklarını göstermişlerdir [10].

2012 yılında Akın ve Eren, OECD ülkelerinin eğitim göstergelerine kümeleme analizi uygulayarak Türkiye'nin OECD ülkeleri arasında geride kaldığı bilgisine ulaşmışlardır. Ayrıca, kümeleme analizi ile OECD ülkeleri eğitim göstergeleri bakımından üç kümeye ayrılmış ve Türkiye'nin tek başına bir küme oluşturduğu görülmüştür [11].

2012 yılında yayımlanan "Principal Component Analysis-Multidisciplinary Applications" adlı kitapta pek çok alanda yapılan faktör analizi ve temel bileşenler analizi ile ilgili çalışmalardan bahsedilmiştir. Bunlar arasından en dikkat çekici çalışmalar; Eric Belasco, Billy U. Philips, Jr. ve Gordon Gong'un temel bileşenler analizi ile ertelenmiş kanserleri algılama çalışması, Érica C. M. Nascimento ve João B. L. Martins'in temel bileşenler analizi ile yeni potansiyel ilaçların tasarımı çalışmasıdır [12].

2012 yılında Çelik, Türkiye'de illerin bitkisel ürün üretimi bakımından durumlarını faktör analizi ile incelemiş ve İç Anadolu bölgesinin tahıl, baklagil ve şeker pancarı üretiminde, Akdeniz bölgesinin sebze ve meyve üretiminde daha zengin olduğu ortaya çıkmıştır. Ayrıca, ülke genelinde üretim bakımından gelişmiş bazı illerin Antalya, Mersin, Muğla, Konya, Ankara, İzmir, Çanakkale, Diyarbakır, Malatya olduğu bilgisine ulaşılmıştır. En az üretim yapılan illerin ise Osmaniye, Düzce, Bilecik, Yozgat, Niğde, Kilis olduğu ortaya çıkmıştır [13].

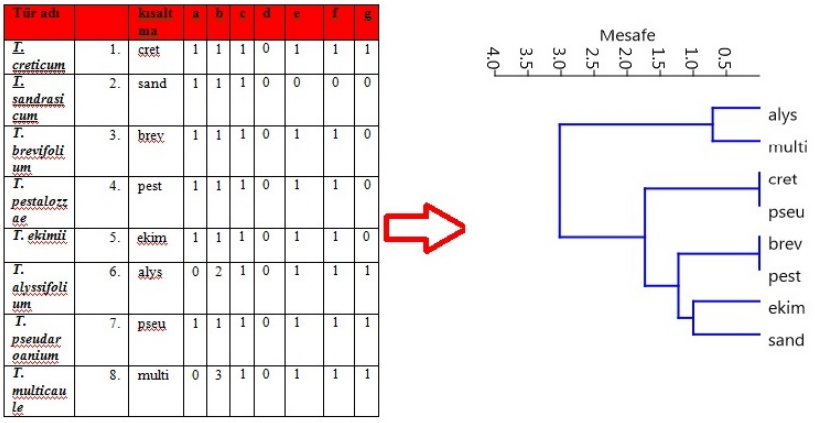
2013 yılında Kaya, 23 maddelik sürdürülebilir kalkınmaya yönelik tutum ölçeği ile uzmanların görüşlerini almıştır. Faktör analizi doğrultusunda sosyal, çevresel ve ekonomik olmak üzere uzmanların görüşlerini üç faktör olarak elde etmiştir. Daha sonra, ölçek için güvenirlik katsayısı hesaplanmıştır. Böylelikle, üç faktörlü ölçeğin yapı geçerliliği faktör analizi ile doğrulamıştır [14].

2015 yılında, Yıldırım, fen bilimlerindeki öğrenme kaygı ölçeği ile 15 öğrenciyle görüşmüştür. Bu görüşme sonuçlarına faktör analizi uygulanmış ve daha sonrasında ölçek için güvenirlik katsayısı hesaplanmıştır. Elde edilen bilgiler sonucunda ölçeğin güvenilir

olduğu bilgisine ulaşılmıştır [15].

2016 yılında Ross ve arkadaşları, kümeleme analizini kullanarak erken doğan çok düşük kilolu bebeklerin gelişimi üzerine çalışma yapmışlardır. Bu çalışmanın sonucunda elde edilen bilgilerle de ileriye dönük tahminler yapacaklardır [16].

Faktör analizi ve kümeleme analizi ile ilgili kaynak çok nadir bulunmaktadır. Bulunan kaynaklar arasında da her detay verilmemektedir. Bu tezde, faktör analizi ve kümeleme analizi yöntemleri ile ilgili detaylı bir matematiksel inceleme yaparak araştırmacılara bir kaynak oluşturulması amaçlanmaktadır. Bunun için öncelikle benzerlik ya da uzaklık ölçüleri, faktör analizi ve kümeleme analizi ile ilgili teoriksel bilgiler verilmiştir. Diğer yandan, bu çalışmanın diğer bir amacı ise, Şekil 1.1'de görüldüğü gibi Teucrium türüne ait mikrobiyolojik canlılar pek çok özelliğe sahiptir, bu canlıların özellikleri göz önüne alındığında verilerin karmaşık bir yapı oluşturduğu ve gözle bakıldığında net bir fikir elde edilemediğinden dolayı, "Hangi türler birbirleriyle aynı gruptadır?" sorusunun cevabını vermek de mümkün olmamaktadır. Dolayısıyla bu tezin diğer bir amacı, gözle incelenemeyen yaklaşık 32 tane Teucrium türüne ait mikrobiyolojik canlının sahip olduğu belli başlı özellikleri göz önünde bulundurarak hem matematiksel hem de istatistiksel bir yaklaşım sergilemek ve bu türler arasındaki akrabalık ilişkilerini kümeleme analizi ve faktör analizi ile ortaya koymaktır.



Tür adı		kısaltma	a	b	c	d	e	f	g
T. creticum	1.	cret	1	1	1	0	1	1	1
T. sandrasicum	2.	sand	1	1	1	0	0	0	0
T. brevifolium	3.	brev	1	1	1	0	1	1	0
T. pestalozzae	4.	pest	1	1	1	0	1	1	0
T. ekimii	5.	ekim	1	1	1	0	1	1	0
T. alyssifolium	6.	alys	0	2	1	0	1	1	1
T. pseudaroanium	7.	pseu	1	1	1	0	1	1	1
T. multicaule	8.	multi	0	3	1	0	1	1	1

Şekil 1.1: Canlı türlerinin gruplandırması.

2. GENEL KISIMLAR

Bu bölümde, tez boyunca sıkça kullanılan konularla ilgili bir altyapı verilecektir ve ayrıca literatürde sıkça kullanılan yöntemler izah edilecektir.

2.1. BENZERLİK ÖLÇÜSÜ

Benzerlik ve uzaklık kavramları pek çok bilim dalı için önemli bir yere sahiptir. Genellikle kümeleme, sınıflandırma, özellik seçimi, aykırı değer tespiti, regresyon içeren neredeyse bütün teşhis etme uygulamalarında benzerlik ve uzaklık ölçülerine gereksinim duyulmaktadır. Literatürde, bu ölçülerle ilgili pek çok yöntem bulunmaktadır. Bundan dolayı, herhangi bir veri tabanı için en uygun benzerlik veya uzaklık ölçüsünün seçimi önemli bir konudur ve kullanılan veri tabanına ve araştırmacının amaçlarına göre farklılık göstermektedir. Bu bölümde, literatürde var olan benzerlik veya uzaklık ölçüleri hakkında detaylı bir bilgi verilmektedir [17].

Nesneler arasındaki benzerlikler veya uzaklıkların ölçümü birçok özelliğe göre değişmektedir. Bu yüzden, nesneler arası benzerlik ya da uzaklıkların hesaplanması yapılmadan önce özellik seçiminin yapılması gerekir. İki değişken arasındaki benzerlik ya da uzaklık, özelliğin değerine göre değişmektedir. Genel bir yaklaşım olarak, her bir özellik için benzerlik ya da uzaklık ölçüleri bulunmaktadır. Bundan dolayı, benzerlik ya da uzaklık ölçüleri aşağıda belirtildiği gibi pek çok kategoriye bölünebilir [17].

1. İkili Değişkenler için Benzerlik/Uzaklık Ölçüleri
2. Kategorik Değişkenler için Benzerlik/Uzaklık Ölçüleri
3. Nicel Değişkenler için Benzerlik/Uzaklık Ölçüleri
4. Sıralı Değişkenler için Benzerlik/Uzaklık Ölçüleri
5. İki Grup Değişken Arasındaki Mesafeler

2.1.1. Benzerlik ve Uzaklık Kavramları

Benzerlik kavramı, nesneleri gruplandırmaya yarayan belli başlı yardımcı araçlardır. Bazı benzerlik ya da uzaklık ölçülerini tanımlamak için, her şeyden önce kullanılacak olan veri tabanı hakkında bazı özelliklerin ortaya çıkarılması gerekmektedir. Örneğin, varsayalım ki bitki türleri kümelenmek istensin. Kümeleme yapılmadan önce bitki türleri arasında nasıl bir benzerliğin ya da benzemezliğin var olup olmadığının incelenmesi gerekir. Bu durum tespit edildikten sonra benzerlik ölçüleri ile nesnelerin birbirlerine olan benzerlik değerleri hesaplanır. Öte yandan, mesafe kavramını iki noktanın birbirlerine göre farklılığı olarak tanımlamak mümkündür [17]. Ayrıca, başka bir bakış açısıyla mesafe kavramına yaklaşıldığında, mesafeyi iki nesne arasındaki düzensizlik gibi görmek de mümkün olabilmektedir.

Kabul edelim ki, i ve j gibi iki nesne var olsun. Bunlar arasındaki benzerlik ölçüleri ve uzaklık ölçüleri sırasıyla s_{ij} ve d_{ij} ile temsil edilebilir ve aralarındaki ilişki,

$$s_{ij} = 1 - d_{ij}$$

şeklinde tanımlanır. Ayrıca, mesafe kavramı bir sayısal değerdir ve bu genellikle aşağıdaki özellikleri sağlar:

- $d_{ij} \geq 0$: iki nesne arasındaki mesafe daima pozitiftir.
- $d_{ii} = 0$: mesafe sıfırdır gerek ve yeter koşul kendisine göre bir ölçüdür.
- $d_{ij} = d_{ji}$: mesafe simetriktir.
- $d_{ij} \leq d_{ik} + d_{kj}$: mesafe üçgen eşitsizliğini sağlar.

Bu dört koşulu sağlayan mesafe ölçüsüne özel olarak *metrik* adı verilir. Bundan dolayı, mesafe ölçüsü ile metrik arasında bir ilişkilendirme yapılırsa, bütün mesafeler metrik değildir ancak her metrik bir mesafedir.

2.1.2. Benzerlikleri Ölçmek Neden Önemlidir?

İki nesne arasındaki benzerlik ölçümleri farklı nesneleri ayırt etmek için çok önemlidir. Benzerliği ölçmek için ana nedenler aşağıdaki gibidir:

- Nesneler birbirinden ayrılabilir.
- Herhangi bir kümeleme algoritması benzerlik ölçüsüne dayanan gruplar için uygulanabilir.
- Kümelemeden sonra nesneler pek çok grup oluşturur. Bu yüzden herhangi bir grubun karakteristiğini anlamak kolaydır.
- Kümelerin karakteristiği açıklanabilir.
- Ayrıca, veri nesnelerinin kümelenmesi daha düzenli şekilde bilgilerin alınmasına ve düzenlenmesine yardımcı olabilir.
- Benzerlik ölçüleri sınıflandırma yapmak için önemlidir. Bir nesne benzerlik ölçüsü yardımıyla grup olarak sınıflandırılabilir.
- Ayrıca benzerlik ölçümleri bu gruplandırma bilgilerine dayalı yeni bir nesnenin davranışını tahmin etmek için kulanılabilir.
- Veri kümesi içindeki yapı benzerlik ölçüsü kullanılarak açıklanabilir.
- Benzerlik ölçümleri uygulandıktan sonra veri, ortaya çıkarılmış ilişkiler kullanılarak basitleşitirilebilir. Bu basitleştirilmiş veri düzgün bir şekilde farklı veri madenleri tarafından kullanılabilir.
- Bu yüzden, verinin tahmini ve yapısını bildikten sonra karar alma ve planlama kolay olur.

Takip eden bölümlerde, farklı türden verilerin benzerliklerini veya uzaklıklarını ölçme yolları gösterilmiştir.

2.1.3. İkili Değişkenler için Benzerlik Ölçüleri

İkili veriler sadece 0/1, doğru/yanlış, evet/hayır, pozitif/negatif gibi ifadelerle temsil edilebilir. İkili veri olarak ifade edilen iki değişken arasındaki benzerliği veya uzaklığı ölçmek için öncelikle her bir değerin toplam oluşum sayısı toplanır. Aşağıda ikili iki değişken arasındaki mesafe hesaplama örneği verilmiştir [17].

K_1 ve K_2 gibi belli özelliklere sahip iki insan örneği göz önüne alınsın. Kabul edelim ki, uzunluğu 1.5 metreden uzun ve kilosu 60 kilogramdan fazla olan erkek ve kadın özellikleri olsun. Bu durumda, K_1 1.8 metre uzunluğunda, 58 kilogram ve erkek ise, bu durumda $K_1 = (1, 0, 1)$ olarak temsil edilir. K_2, 1.4 metre uzunluğunda, 50 kilogram ve bayan ise, bu durumda $K_2 = (0, 0, 0)$ olarak temsil edilir. Burada, hem K_1 hem de K_2 3-boyutlu nesnelerdir çünkü her bir nesne 3 değişken tarafından temsil edilmektedir.

Varsayalım ki,

m_{00} = her iki nesne için 0'a sahip olan özelliklerin toplam sayısı

m_{01} = i'nci nesne için 0'a ve j'nci nesne için 1'e sahip olan özelliklerin toplam sayısı

m_{10} = i'nci nesne için 1'e ve j'nci nesne için 0'a sahip olan özelliklerin toplam sayısı

m_{11} = her iki nesne için 1'e sahip olan özelliklerin toplam sayısı

Bu durumda, özelliklerin toplam sayısı ($\mathbb{F}$)=$m_{00} + m_{01} + m_{10} + m_{11}$ dir.

Tablo 2.1: Kontenjans Tablosu.

i / j	0	1
0	m_{00}	m_{01}
1	m_{10}	m_{11}

Tablo 2.1 m_{00}, m_{01}, m_{10}, m_{11} kavramlarını gösteriyor. K_1 ve K_2'nin yukarıdaki örneği için, $m_{00} = 1$, $m_{01} = 0$, $m_{10} = 2$, $m_{11} = 0$ dır. Literatürde şu anda ikili veriler için yirmiden fazla benzerlik veya uzaklık ölçüsü bulunmaktadır. Bunlar arasında en popüler olanları verelim. Not olarak, verilen bütün ölçüler Tablo 2.1'deki kontenjans tablosu baz alınarak tanımlamalar yapılmıştır.

2.1.3.1. Basit Eşleştirme Katsayısı

Bu benzerlik katsayısı $0 - 0$ ve $1 - 1$ ikili değişkenler arasındaki benzerliği hesaplamak için kullanılır ve i ile j nesneleri arasındaki basit eşleştirme katsayısı

$$s_{ij} = \frac{m_{00} + m_{11}}{\mathbb{F}} \tag{2.1}$$

şeklinde tanımlanır. Burada, $\mathbb{F}$ özelliklerin toplam sayısını göstermektedir. Aslında bakılırsa, basit eşleştirme katsayısı $0-0$ ve $1-1$ eşleşmelerinin tüm durumlar karşısında olabilirliğini ifade eden bir olasılık değerinden başka birşey değildir. Örneğin, K_1 ve K_2 değişkenleri arasındaki basit eşleştirme katsayısı

$$s_{K_1K_2} = \frac{1+0}{3} = \frac{1}{3} = 0,333$$

olarak bulunur.

2.1.3.2. Jaccard Katsayısı

Bu katsayının basit eşleştirme katsayısından farkı, 0-0 eşleştirmelerinin var olmadığı durumlarda kullanılmasıdır. Burada ikisi de sıfır olan eşleşmeler katsayıda pay ve paydaya yazılmazlar. Yani, kontenjans tablosundaki $0-0$ eşleşmesi haricindeki tüm durumlar karşısında $1-1$ eşleşmesinin olabilme durumunu belirten olasılık değerinden bahsedilmektedir. Örneğin, bir manav göz önüne alınsın. Manavda herkesin alabileceği kadar meyve bulunmaktadır. Kabul edelim ki, amacımız iki müşteri tarafından alınan benzer meyvelerin sayısını hesaplamak olsun. Basit eşleştirme katsayısını hesaplamada yapıldığı gibi iki müşteri tarafından alınan ortak olmayan meyvelerin sayısını hesaplamak zaman alıcı bir görevdir. Daha doğrusu, iki müşteri tarafından alınan sadece ortak meyvelerin toplam sayısını hesaplamak daha kolaydır. Jaccard katsayısı bu tip olaylarla ilgilenmektedir ve i ile j nesneleri arasındaki Jaccard katsayısı,

$$s_{ij} = \frac{m_{11}}{m_{01} + m_{10} + m_{11}} \tag{2.2}$$

şeklinde tanımlanır. K_1 ve K_2 özelliklerine ait örnek göz önüne alındığında, K_1 ve K_2 arasıdaki Jaccard katsayısı 0 dır. Ayrıca, Jaccard katsayısı ikili olmayan değişkenlere uygulamak için küme teorisi kullanılarak genelleştirilebilir [17]. Varsayalım ki, A ve B iki küme olmak üzere; aralarındaki Jaccard katsayısı

$$s_{AB} = \left|\frac{A \cap B}{A \cup B}\right| \tag{2.3}$$

olarak hesaplanır. Örneğin, $A = \{4, 5, 6, 7, 8\}$ ve $B = \{2, 4, 6, 8\}$ kümelerini düşünelim. Böylece, $A \cap B = \{4, 6, 8\}$ ve $A \cup B = \{2, 4, 5, 6, 7, 8\}$'dir. Bu durumda, $s_{AB} = \frac{3}{6} = 0.5$ olur.

2.1.3.3. Sorensen Katsayısı

Sorensen katsayısı Jaccard katsayısına çok benzerdir ve 1913 yılında Czekanowski tarafından ilk kez kullanıldı ve sonrasında Sorensen tarafından yeniden keşfedildi [18]. Bu katsayı aynı zamanda Czekanowski katsayısı ya da Dice katsayısı olarak da bilinir. Burada $0 - 0$ eşleştirmesi yok sayılarak $1 - 1$ eşleştirmesine 2 kat ağırlık verilir. Jaccard katsayısından farklı olarak bu katsayıda $1 - 1$ eşleştirmesinin olma olasılığı arttırılmıştır ve aşağıdaki gibi tanımlanır:

$$s_{ij} = \frac{2m_{11}}{2m_{11} + m_{01} + m_{10}} \tag{2.4}$$

Sorensen ve Jaccard katsayıları arasında çok yakın bir ilişki vardır ve bu ilişki

$$K_s(i,j) = \frac{s \cdot m_{11}}{s \cdot m_{11} + m_{01} + m_{10}} \quad ; s = 1, 2, ... \tag{2.5}$$

indeksi ile tanımlanır. Burada özel olarak $K_1(i,j)$ Jaccard katsayısını, $K_2(i,j)$ ise Sorensen katsayısını verir.

2.1.3.4. Hamming Mesafesi

İkili ifade olarak temsil edilen değişkenler arasındaki Hamming mesafesi farklı eşleşmelere sahip olan durumların yani, 0-1 ve 1-0 eşleşmelerinin toplamına eşittir [17]. Bu mesafe

$$d_{ij} = m_{01} + m_{10} \tag{2.6}$$

olarak tanımlanır. K_1 ve K_2 iki ikili değişkenleri tekrardan düşünüldüğünde, K_1 ve K_2 arasındaki Hamming mesafesi

$$d_{K_1K_2} = 0 + 2 = 2$$

olarak hesaplanmaktadır.

2.1.3.5. Russel-Rao Katsayısı

Russel-Rao katsayısı sadece $1-1$ eşleşmesinin bütün özelliklerin toplamına oranı olarak ifade edilebilir ve

$$s_{ij} = \frac{m_{11}}{\mathbb{F}} \tag{2.7}$$

şeklinde tanımlanmaktadır. Bu katsayı, bahsedilen diğer katsayılara benzer olarak kontenjans tablosundaki bütün durumlara göre $1-1$ eşleşmesinin olma olasılığını ifade eden bir olasılık değeridir. Bu doğrultuda, yukarıda verilen örnekteki K_1 ve K_2 değişkenleri için benzerlik katsayısı

$$s_{K_1K_2} = \frac{0}{1+0+2+0} = \frac{0}{3} = 0$$

olarak bulunur.

2.1.3.6. Rogers-Tanimiato Katsayısı

Bu katsayı $0-0$ ve $1-1$ eşleşmeleri pay ve paydada yer almakta ve aynı zamanda $0-1$ ve $1-0$ eşleşmelerinin ağırlıkları payda kısmında 2 kat ağırlık alarak olmaktadır ve

$$s_{ij} = \frac{m_{00} + m_{11}}{m_{00} + 2(m_{01} + m_{10}) + m_{11}} \tag{2.8}$$

şeklinde tanımlanır [19]. Aslında bakılırsa, basit eşleştirme katsayısına benzemektedir ancak burada kontenjans tablosundaki tüm durumlara karşı $0-0$ ve $1-1$ eşleştirmelerinin olma olasılıklarını hesaplarken, $0-1$ ve $1-0$ eşleştirmesine 2 kat ağırlık verilmektedir. Bu bağlamda, Bölüm 2.1.3'de tanımlanan K_1 ve K_2 değişkenleri için Rogers-Tanimiato katsayı

$$d_{K_1K_2} = \frac{1+0}{1+2(0+2)+0} = \frac{1}{5} = 0.2$$

olarak elde edilir.

2.1.3.7. Baroni-Urbani ve Buser Katsayısı

Bu katsayı 1976 yılında Baroni-Urbani ve Buser tarafından önerilmiştir ve var olarak eşleşenlerin yok olarak eşleşenlere çarpımı ile elde edildiği için diğer katsayılara nispeten daha zahmetli bir benzerlik katsayısı olup aşağıdaki formda tanımlanır.

$$s_{ij} = \frac{\sqrt{m_{11}m_{00}} + m_{11}}{\sqrt{m_{11}m_{00}} + m_{11} + m_{01} + m_{10}} \tag{2.9}$$

2.1.3.8. Diğer Katsayılar

Tablo 2.2: Literatürde bulunan diğer popüler benzerlik katsayıları.

Sokal ve Sneath-1 Katsayısı	$s_{ij} = \frac{2(m_{00}+m_{11})}{2m_{00}+m_{01}+m_{10}+2m_{11}}$
Sokal ve Sneath-2 Katsayısı	$s_{ij} = \frac{m_{11}}{m_{11}+2(m_{01}+m_{10})}$
Sokal ve Sneath-3 Katsayısı	$s_{ij} = \frac{m_{00}+m_{11}}{m_{01}+m_{10}}$
Sokal ve Sneath-4 Katsayısı	$s_{ij} = \frac{1}{4}(\frac{m_{11}}{m_{11}+m_{01}} + \frac{m_{11}}{m_{11}+m_{10}} + \frac{m_{00}}{m_{00}+m_{10}} + \frac{m_{00}}{m_{00}+m_{01}})$
Sokal ve Sneath-5 Katsayısı	$s_{ij} = \frac{m_{00}m_{11}}{\sqrt{(m_{11}+m_{10})(m_{11}+m_{01})(m_{10}+m_{00})+(m_{01}+m_{00})}}$
Kulczynski-1 Katsayısı	$s_{ij} = \frac{m_{11}}{m_{01}+m_{10}}$
Kulczynski-2 Katsayısı	$s_{ij} = \frac{m_{11}/(m_{11}+m_{10})+m_{11}/(m_{11}+m_{01})}{2}$
Goodman-Kruskal Lambda	$s_{ij} = \frac{t_1-t_2}{2\mathbb{F}-t_2}$
Anderberg D Katsayısı	$s_{ij} = \frac{t_1-t_2}{2\mathbb{F}}$
Yule Q Katsayısı	$s_{ij} = \frac{m_{00}m_{11}-m_{01}m_{10}}{m_{00}m_{11}+m_{01}m_{10}}$
Yule Y Katsayısı	$s_{ij} = \frac{\sqrt{m_{00}m_{11}}-\sqrt{m_{01}m_{10}}}{\sqrt{m_{00}m_{11}}+\sqrt{m_{01}m_{10}}}$
Ochiai Katsayısı	$s_{ij} = \sqrt{(\frac{m_{11}}{m_{11}+m_{10}})(\frac{m_{11}}{m_{11}+m_{01}})}$
Yayılım Benzerlik Ölçüsü	$s_{ij} = \frac{m_{00}m_{11}-m_{01}m_{10}}{\mathbb{F}^2}$
Hamann Katsayısı	$s_{ij} = \frac{(m_{00}+m_{11})-(m_{01}-m_{10})}{\mathbb{F}}$

Burada $\mathbb{F}$ bütün eşleşmelerin toplam sayısı ve $t_1 = \max(m_{11}, m_{10}) + \max(m_{01}, m_{00}) + \max(m_{11}, m_{01}) + \max(m_{10}, m_{00})$ ve $t_2 = \max(m_{11} + m_{01}, m_{10} + m_{00}) + \max(m_{11} + m_{10}, m_{01} + m_{00})$ dır.

2.1.4. Kategorik Değişkenler için Benzerlik Ölçüleri

Kategorik değişkene sahip veri tabanlarında sayılarla sadece farklı kategoriler temsil edilir. Örneğin, cinsiyet sınıfı 0 ve 1 ifadesi ile sırasıyla erkek ve kadını temsil eden kategorik değişkendir.

Kategorik veriler için benzerlik ve uzaklığı ölçmek sürekli verilerdeki ölçme işlemine göre daha zordur [20]. Kategorik özelliklere sahip olan iki değişken arasındaki mesafeyi hesaplamak için, öncelikle her bir özellik için olası kategorilerin sayısını hesaplamak zorunludur. İki kategori olduğunda, Basit Eşleştirme, Jaccard ve Hamming mesafeleri gibi ikili değişkenler için kullanılan mesafe ölçüleri kullanılabilir. Kategorilerin sayısı ikiden daha fazla ise, bu kategorileri ikili değişkenlerin bir kümesine dönüştürmeye ihtiyaç duyulur. Bir kategorik değişkeni ikili değişkene dönüştürmek için iki metot vardır.

1. Her bir kategori bir ikili değişken tarafından temsil edilir.
2. Her bir kategori birçok ikili değişken tarafından temsil edilir.

Yukarıdaki iki metotu kullanmak farklı mesafe ölçüleri oluşturur. Mesafe fonksiyonlarını hesaplamak orijinal değişkenlere dayanır [17]. Öncelikle, bir kategorik değişkenin değerini temsil etmek için gerekli ikili değişkenlerin toplam sayısı belirlenmelidir. Bu durumda, iki kategorik değişken arasındaki mesafe

p = i. nesne için 0'lara sahip olan ve j. nesne için 1'lere sahip olan değişkenlerin sayısı

q = i. nesne için 1'lere sahip olan ve j. nesne için 0'lara sahip olan değişkenlerin sayısı

n_k = Kategorik değişkenleri temsil eden ikili değişkenlerin sayısı

olmak üzere i ve j nesneleri arasındaki mesafe

$$d_{ij} = \frac{p+q}{n_k} \tag{2.10}$$

olarak tanımlanmaktadır.

2.1.4.1. Metot 1: Her Bir Kategori Bir Tek İkili Değişken Tarafından Temsil Edilir

Bu metotta her bir kategori bir ikili değişken tarafından temsil edilebilir. Bu durumda, iki değişken arasındaki mesafe, eşleşmeyenlerin sayısının bu kategorik değişkenleri temsil etmek için kullanılan ikili değişkenlerin toplam sayısına oranı olarak hesaplanır.

Varsayalım ki, renk ve cinsiyet gibi iki değişken var olsun. Cinsiyet; erkek $= 0$ ve kadın $= 1$ olan iki değere sahiptir. Renk ise; beyaz, mavi ve kırmızı gibi seçilmiş üç değer alsın. Kabul edelim ki, Ahmet bir kırmızı tişört, Leyla bir beyaz tişört ve Ayşe de bir mavi tişört giyen üç denek olsun. Renklerin her bir değeri bir ikili değişken ile gösterilir. Birinci koordinat olarak cinsiyeti ve ikinci koordinat olarak rengi düşünelim. Bu durumda, Ahmet, Leyla ve Ayşe nesnelerine karşılık gelen özellik vektörleri Ahmet $= (0, (1, 0, 0))$, Leyla $= (1, (0, 1, 0))$ ve Ayşe $= (1, (0, 0, 1))$ dir.

Nesneler arasındaki mesafeyi hesaplamak için öncelikle her bir koordinat için hesaplanmalıdır. Bu örnek için Hamming mesafesi kullanıldığında, her bir koordinat için $0 - 1$ ve $1 - 0$ eşleşmelerinin sayısı toplanır. Böylelikle,

- Ahmet ve Leyla arasınaki mesafe (1,2) dir ve iki değişken için toplam mesafe $1 + 2 = 3$ elde edilir.
- Ahmet ve Ayşe arasındaki mesafe (1,2) dir ve iki değişken için toplam mesafe $1 + 2 = 3$ elde edilir.
- Leyla ve Ayşe arasındaki mesafe (0,2) dir ve iki değişken için toplam mesefe $0+2 = 2$ elde edilir.

Daha sonra, iki kategorik nesne arasındaki mesafe eşleşmeyenlerin sayısının kategorik değişkenleri temsil etmek için kullanılan ikili değişkenlerin toplam sayısına oranı olarak hesaplanır. Böylece,

- Ahmet ve Leyla arasındaki mesafe (1/1,2/3) dir ve iki değişken için ortalama mesafe (1+2/3)/2=5/6=0.83
- Ahmet ve Ayşe arasındaki mesafe (1/1,2/3) dir ve iki değişken için ortalama mesafe (1+2/3)/2=5/6=0.83

- Leyla ve Ayşe arasındaki mesafe (0/1,2/3) dir ve iki değişken için ortalama mesafe (0+2/3)/2=1/3=0,33

şeklinde hesaplanır.

2.1.4.2. Metot 2: Her Bir Kategori Birçok İkili Değişken Tarafından Temsil Edilir

Bu metotta her bir kategori birçok ikili değişken tarafından temsil edilebilir. Örneğin, tişörtün rengini temsil etmek için bir önceki örnekte, iki ikili değişkenlere ihtiyaç duyulur. Daha sonra, üç renk; kırmızı renk 00, beyaz 01 ve mavi renk 10 olarak temsil edilebilir. Burada göstermelik değişkenler için etiketleme bazen keyfidir ama tutarlıdır.

Bölüm 2.1.4.1'deki örnek düşünüldüğünde, burada iki tane cinsiyet ve renk gibi kategorik değişkenler vardır. Cinsiyet ; erkek $= 0$ ve kadın $= 1$ değerlerine sahiptir. Renk ise; kırmızı, beyaz ve mavi gibi seçilmiş üç değere sahip olsun. Cinsiyeti temsil etmek için bir ikili göstermelik değişken kullanılır oysa ki renkleri temsil etmek için iki adet göstermelik ikili değişken kullanılır. Bu durumda, Ahmet $= (0, (0, 0))$, Leyla $= (1, (0, 1))$ ve Ayşe $= (1, (1, 0))$ şeklinde temsil edilir.

Nesneler arasındaki mesafeyi hesaplamak için her bir orijinal değişken için hesaplama olmalıdır.
Varsayalım ki Hamming mesafesi kullanılsın. Bu durumda,

- Ahmet ve Leyla arasındaki mesafe (1,1) dir ve toplam mesafe 1+1=2 elde edilir.
- Ahmet ve Ayşe arasındaki mesafe (1,1) dir ve toplam mesafe 1+1=2 elde edilir.
- Leyla ve Ayşe arasındaki mesafe (0,2) dir ve toplam mesafe 0+2=2 elde edilir.

İki kategorik nesne arasındaki mesafe eşleşmeyenlerin sayısını kategorileri temsil etmek için kullanılan ikili değişkenlerin toplam sayısına oranı olarak hesaplanır.

- Ahmet ve Leyla arasındaki mesafe (1/1,1/2) dir ve iki değişken için ortalama mesafe (1+1/2)/2=3/4 elde edilir.

- Ahmet ve Ayşe arasındaki mesafe (1/1,1/2) dir ve iki değişken için ortalama mesafe (1+1/2)/2=3/4 elde edilir.
- Leyla ve Ayşe arasındaki mesafe (0/1,2/2) dir ve iki değişken için ortalama mesafe (0+1)/2=1/2 elde edilir.

2.1.5. Nicel Değişkenler için Mesafe Ölçüleri

Nicel değişkenler nümerik bir ölçüt ile ölçülebilir. Bu yüzden, onlar bazı kategorileri temsil eden kategorik değişkenlerden ve de değişkenlerin sırasını temsil eden sıralı değişkenlerden farklıdırlar. Aynı zamanda, bu nicel değişkenlerin ölçü değerleri bir sayı olarak ortaya çıkar. Bu yüzden, herhangi bir matematiksel işlem nicel değişkenlere uygulanabilir. Nicel değişkenlere örnek olarak ağırlık, basınç miktarı,sıcaklık, yüzölçümü verilebilir [17].

Kabul edelim ki, yaş, günlük ortalama yapılan spor süresi ve aylık hastane masrafını ele alan bir durumu ölçmek gereksin. Bir nesneyi diğerinden ayırt etmek için üç nicel değişken kullanılsın. Üç özellik yaş(yıl), günlük ortalama yapılan spor süresi(dakika) ve aylık hastane masrafı(TL)'tır. Varsayalım ki, 1. Gözlem 18 yaşında, günlük ortalama 25 dakika spor yapıyor, aylık hastane masrafı olarak 100 TL harcıyor ve 2. Gözlem ise 35 yaşında, günlük ortalama 30 dakika spor yapıyor, hastaneye aylık 300 TL harcama yapıyor olsun.

Bu iki nesne 3-boyutlu noktalar olarak temsil edilir. Gözlem 1 $(18, 25, 100)$ koordinatlarına ve Gözlem 2 ise $(35, 30, 300)$ koordinatlarına sahiptir. Bu durumda bu iki nesne arasındaki benzerlik ya da mesafe bu koordinatlara bağlıdır.

2.1.5.1. Öklidyen Mesafe

Öklidyen mesafe nicel değişkenler için kullanılan mesafe ölçüleri arasında en çok kullanılan ölçü türüdür. n-boyutlu uzayda x_{ik} ve x_{jk} gibi iki nokta arasındaki Öklidyen mesafe

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{2.11}$$

şeklinde tanımlanır.

2.1.5.2. City Block Mesafesi

City Block mesafesi ayrıca Manhattan mesafesi olarak da bilinmektedir ve bütün özelliklerin mesafelerinin toplamı şeklinde hesaplanır. Yani, n-boyutlu uzayda iki nokta arasındaki City Block mesafesi

$$d_{ij} = \sum_{k=1}^{n}|x_{ik} - x_{jk}| \tag{2.12}$$

şeklinde tanımlanır.

2.1.5.3. Chebyshev Mesafesi

Sayısal değerlere sahip nicel değişkenler için kullanılmaktadır. Bu mesafe ölçüsünün diğer bir ismi ise Maksimum mesafedir. Bu mesafe özellikler arasındaki mesafenin maksimum değeridir ve

$$d_{ij} = \max_{k}|x_{ik} - x_{jk}| \tag{2.13}$$

olarak tanımlanır.

2.1.5.4. Minkowski Mesafesi

Minkowski mesafesi daha genel bir mesafe ölçüsüdür ve $m \geq 1$'nin değerine bağlıdır. Bu mesafe pek çok mesafe fonksiyonuna indirgenebilir. Mesela, $m = 1$ iken City Block mesafesine, $m = 2$ iken Öklidyen mesafesine ve $m = \infty$ iken Chebyshev mesafesine indirgenmektedir. Minkowski mesafesi hem sıralı hem de nicel değişkenler için kullanılabilir ve

$$d_{ij} = \sqrt[m]{\sum_{k=1}^{n}(x_{ik} - x_{jk})^m} \tag{2.14}$$

şeklinde tanımlanır.

2.1.5.5. Bray-Curtis Mesafesi

Bu mesafe ölçüsü genellikle botanik, ekoloji ve çevre bilimleri ile ilgili çalışmalarda kullanılmaktadır. Bütün koordinatlar pozitif olduğunda, bu mesafe fonksiyonunun değeri $[0, 1]$ arasında yer almaktadır. Ancak, her iki nesne sıfır koordinatlarına sahipse, bu durumda Bray-Curtis mesafesi tanımsız olacaktır [17]. Bu mesafe fonksiyonu

$$d_{ij} = \frac{\sum_{k=1}^{n} |x_{ik} - x_{jk}|}{\sum_{k=1}^{n} (x_{ik} + x_{jk})} \tag{2.15}$$

şeklinde tanımlanmaktadır.

2.1.5.6. Açısal Ayrışım

Buna bazen korelasyon katsayısı da denir. Öncelikle, iki vektör arasındaki açının kosinüsünü ölçer. Bu benzerlik ölçüsünü diğerlerinde ayıran en büyük farklılığı $[-1, +1]$ arasında değer almasıdır [17]. İki vektör benzer olduğunda, açısal ayrışım ölçüsü $+1$'e yakın bir değer olacaktır. Açısal ayrışımı tanımlamak için $\frac{0}{0} = 0$ olarak kabul edilir. İki vektör arasındaki Açısal ayrışım

$$s_{ij} = \frac{\sum_{k=1}^{n} (x_{ik} \times x_{jk})}{\sqrt{(\sum_{k=1}^{n} x_{ik}^2 \times \sum_{r=1}^{n} x_{jr}^2)}} \tag{2.16}$$

şeklinde tanımlanır.

Bu mesafenin isminin kökeni aşağıdaki gibidir. İki vektör arasındaki açının kosinüsü vektörlerin iç çarpımının normlarının çarpımına oranıdır ve

$$\cos\theta = \frac{< \mathbf{A} \cdot \mathbf{B} >}{||\mathbf{A}|| \cdot ||\mathbf{B}||} \tag{2.17}$$

biçiminde yazılır. Ayrıca, modül olarak adlandırılan **A**'nın normu koordinatlarının karesinin toplamının kareköküdür. Yani, $||A|| = \sqrt{a_1^2 + a_2^2 + ... + a_n^2}$'dır. Bu yüzden,

$$\cos\theta = \frac{a_1 b_1 + a_2 b_2 + ... + a_n b_n}{\sqrt{a_1^2 + a_2^2 + ... + a_n^2}\sqrt{b_1^2 + b_2^2 + ... + b_n^2}} \tag{2.18}$$

elde edilir.

2.1.5.7. Kolerasyon Katsayısı

Kolerasyon katsayısı lineer kolerasyon katsayısı ya da Pearson kolerasyon katsayısı olarak da bilinmektedir. $[-1,+1]$ arasında değerler alan Açısal ayrışımın özel bir durumudur [17]. Bu değerin elle hesaplanması zor olacağından bugün gelişen teknoloji sayesinde bu tür ölçümleri hesaplamak kolay bir hal almıştır. Bu mesafe

$$s_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \overline{x_i})(x_{jk} - \overline{x_j})}{\sqrt{(\sum_{k=1}^{n}(x_{ik} - \overline{x_i})^2 \sum_{k=1}^{n}(x_{jk} - \overline{x_j})^2)}} \tag{2.19}$$

şeklinde formülize edilir.

2.1.5.8. Mahalanobis Mesafesi

Mahalanobis mesafe ölçüsü kuadratik mesafe olarak da adlandırılmaktadır. Genel olarak, nesnelerin iki grubu arasındaki mesafeyi tanımlar. Kabul edelim ki, $\overline{x_i}$ ve $\overline{x_j}$ ortalamalarıyla iki grup var olsun. Bu durumda, Mahalanobis mesafesi

$$d_{ij} = \sqrt{(\overline{x_i} - \overline{x_j})^T \times S^{-1} \times (\overline{x_i} - \overline{x_j})} \tag{2.20}$$

şeklinde tanımlanır.

Örnek 2.1. Armstrong Colorado'nun Rocky dağlarında küçük memelilerin dokuz türünü yakalamış ve aşağıdaki tabloda söğüt ormancılığının yapıldığı ve ormancılığın hiç yapılmadığı iki habitat türü için yakaladıklarının tahmini yüzdelerini elde etmiştir.

Habitat türleri	Sc	Sy	Em	Pm	Cg	Pl	Ml	Mm	Zp
Söğüt Ormancılığı	70	58	5	0	4	0	31	5	35
Ormancılık Yok	10	11	20	20	9	8	11	46	44

Bu durumda, türlerin Öklidyen mesafesi Denklem (2.11)'den

$$\begin{aligned} d_{ij} &= \sqrt{\sum_{k=1}^{9}(x_{ik} - x_{jk})^2} \\ &= \sqrt{(70-10)^2 + (58-11)^2 + (5-20)^2 + ... + (35-44)^2} \\ &= \sqrt{8685} = 93.19 \end{aligned}$$

City Block mesafesi Denklem (2.12)'den

$$\begin{aligned} d_{ij} &= \sum_{k=1}^{9} |x_{ik} - x_{jk}| \\ &= |70-10| + |58-11| + |5-20| + ... + |35-44| \\ &= 225 \end{aligned}$$

Chebyshev mesafesi Denklem (2.13)'den

$$\begin{aligned} d_{ij} &= \max_{k} |x_{ik} - x_{jk}| \\ &= \max\{|70-10|, |58-11|, |5-20|, ..., |35-44|\} \\ &= |70-10| = 60 \end{aligned}$$

Bray-Curtis mesafesi Denklem (2.15)'den

$$\begin{aligned} d_{ij} &= \frac{\sum_{k=1}^{9} |x_{ik} - x_{jk}|}{\sum_{k=1}^{9} (x_{ik} + x_{jk})} \\ &= \frac{|70-10| + |58-11| + |5-20| + ... + |35-44|}{(70+10) + (58+11) + (5+20) + ... + (35+44)} \\ &= \frac{225}{387} = 0.58 \end{aligned}$$

olarak bulunur. Ayrıca, bu d_{ij} mesafelerini bir benzerlik ölçüsü olarak kullanmak için,

$$s_{ij} = 1 - d_{ij}$$

denkleminde gerekli basit cebirsel işlemler yapılarak benzerlik ölçüsü cinsinden de sonuçlar elde edilebilir.

2.1.6. Sıralı Değişkenler İçin Mesafe Ölçüleri

Keyfi bir sıralama ölçütüne dayanan ve değerler arasındaki mesafelerin anlamlı olmadığı ölçüm birimidir. Veri tabanlarında yapılan incelemeler doğrultusunda elde edilen gözlem değerlerine ait ölçüm sonuçları arasında sıralama yapılabilmektedir. Bu yüzden, bu değişkenler arasındaki nicel farklılıklar önemli değildir ancak

değişkenlerin sırası önemlidir [17]. Örneğin, öğrencilerin mezuniyet notları düşünüldüğünde, AA AB'den daha büyük ve BB de BC'den daha büyük olarak dizilsin. Burada, gözlemlerin dizilimi önemlidir ancak AA ile AB arasındaki mesafe önemli bir durum değildir. Ayrıca bazen, sayısal ifadeler sıralı değişkenleri temsil etmek amacıyla kullanılır. Sıralı ölçüye bir örnek verilirse,

- Bağıl Not: AA = Çok İyi, AB = İyi, BB = Orta, DD = Kötü
- Öncelik Sıraması: 1 = En İyi
- Memnuniyet Derecesi :1 = Hiç Memnun Değil, 100 = Çok Memnun

şeklinde olabilir. Sıralı değişkenler arasındaki mesafeyi hesaplamak için genellikle Standartlaştırılmış Rank Dönüşümü, Spearman Mesafesi, Footrule Mesafesi, Kendall Mesafesi, Cayley Mesafesi, Hamming Mesafesi, Chebysev/Maksimum Mesafesi ve Minkowski Mesafesi gibi metotlar kullanılır. Bu mesafeler aşağıdaki gibi tanımlanır.

2.1.6.1. Standartlaştırılmış Rank Dönüşümü

Sıralı değişkenler standartlaştırma yapılarak nicel değişkene dönüştürülür. Sıralamayı standartlaştırdıktan sonra, mesafe Bölüm 2.1.5'da tanımlanan mesafe yöntemleri ile nicel değişken olarak hesaplanır. Sıralı değişkenler tarafından temsil edilen iki nesne arasındaki mesafeyi tanımlamak için aşağıdaki adımlar gerçekleştirilerek sıralı ölçeği oran ölçeğine dönüştürülür [17].

- Sıralı değişken ranka dönüştürülür. (r=1)
- R en büyük rank olmak üzere, sıralama aşağıdaki formül kullanılarak [0,1] aralığında standartlaştılır.

$$x = \frac{r-1}{R-1} \tag{2.21}$$

- Bu sıralı değişkenler arasındaki mesafe nicel değişkenler olarak sıralı değişkenler şeklinde davranarak hesaplanır.

Bir sıralı değişken, bir nicel değişken olarak standartlaştırılsa, sadece bu mesafe kullanılır. Böyle dönüşüm tipleri kullanılmayacaksa, bu durumda, sıralı değişkenler için Spearman, Cayley ve Hamming gibi ya da Cheybsev/Maksimum, Ulam mesafeleri gibi diğer mesafe ölçüleri kullanılabilir.

Standartlaştırılmış rank dönüşümüne bir örnek verilse, bir şirketin yaptığı işe alım mülakatları için gözlemci niteliğinde insan kaynakları çalışanlarının görevlendirildiği düşünülsün. Her bir gözlemci iş başvurusunda bulunanları gözlemliyor ve her bir işe başvuranlar için uygunluğu, netliği, özgünlüğü, sağlamlığı ve anlamlı karşılaştırma gibi kriterler ortaya koyarak bu kişileri değerlendirmektedirler. Her beş kabul kriteri Çok Kötü $=-2$, Kötü $=-1$, Orta $=0$, İyi $=1$, Çok İyi $=2$ gibi özelliklere sahiptir. Kabul edelim ki, işe başvuran bir kişi için iki gözlemcinin değerlendirmesi aşağıdaki gibi olsun.

	Uygunluk	Netlik	Özgünlük	Sağlamlık	İş Becerisi ve Potansiyeli
Gözlemci 1	0	-1	0	1	0
Gözlemci 2	1	0	1	0	0

Amaç cevaplara göre bu iki gözlemci arasındaki benzersizliği/mesafeyi ölçmektir. Öncelikle, sıralı ölçek oran ölçeğine dönüştürülür. Orijinal endeks sıralanır ve bir ranka dönüştürülür. En büyük rank R=5 tir. Bu durumda, rank $[0,1]$ aralığı için standartlaştı. Örneğin, gözlemci 1'in 1 durumunda, $r=3$ olan sıralamaya dönüştüren $i=0$ vardır ve standartlaştırılmış rank $\frac{3-1}{5-1}=\frac{2}{4}=0.5$ tir. Aşağıdaki dönüşüm sağlanır:

Orjinal Endeks	-2	-1	0	1	2
Dönüştürülmüş Rank	1	2	3	4	5
Standartlaştırılmış Rank	0	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{3}{4}$	1

Gözlemci 1 ve Gözlemci 2'nin yeni koordinatları standartlaştırılmış rank kullanılarak elde edilir. Bu durumda, Gözlemci 1'in eski koordinatı olan $(0,-1,0,1,0)$'a karşılık $(\frac{2}{4},\frac{1}{4},\frac{2}{4},\frac{3}{4},\frac{2}{4})$ koordinatı ve Gözlemci 2'nin eski koordinatı olan $(1,0,1,0,0)$'a karşılık $(\frac{3}{4},\frac{2}{4},\frac{3}{4},\frac{2}{4},\frac{2}{4})$ koordinatı gelir. Bu durumda, Gözlemci 1 ile Gözlemci 2 arasındaki öklidyen mesafe

$$d_{12}=\sqrt{\left(\frac{2}{4}-\frac{3}{4}\right)^2+\left(\frac{1}{4}-\frac{2}{4}\right)^2+\left(\frac{2}{4}-\frac{3}{4}\right)^2+\left(\frac{3}{4}-\frac{2}{4}\right)^2+\left(\frac{2}{4}-\frac{2}{4}\right)^2}=0.5 \qquad (2.22)$$

şeklinde elde edilir. Yani, bu gözlemciler $\%50$ olasılıkla birbirine yakın değerlendirme yapmışlardır.

2.1.6.2. Spearman Mesafesi

Spearman mesafesi sıralı iki vektör arasındaki Öklidyen mesafenin karesidir ve aşağıdaki gibi hesaplanır.

$$d_{ij} = \sum_{p=1}^{n} (x_{ip} - x_{jp})^2 \tag{2.23}$$

Örneğin, **A** ve **B** gibi iki kişiye toplu taşıma hakkındaki tercihleri sorulsun. Burada $\mathbf{A} = [\text{Otobüs}, \text{Metro}, \text{Metrobüs}]$ ve $\mathbf{B} = [\text{Metro}, \text{Otobüs}, \text{Metrobüs}]$ şeklindeki sıralı vektörlerdir. Kabul edelim ki, model vektör ise $[\text{Otobüs}, \text{Metro}, \text{Metrobüs}]$ olsun. Bu durumda, **A**'nın koordinatları $(1, 2, 3)$ ve **B**'nin koordinatları ise $(2, 1, 3)$ olarak bulunmaktadır. Böylelikle, **A** ve **B** arasındaki Spearman mesafesi

$$d_{\mathbf{AB}} = (1-2)^2 + (2-1)^2 + (3-3)^2 = 2 \tag{2.24}$$

olarak hesaplanmaktadır.

2.1.6.3. Footrule Mesafesi

Footrule Mesafe iki sıralı vektörün özellik değerleri arasındaki mutlak farklarının toplamıdır. Bu mesafe nicel değişkenler için kullanılan City Block Mesafesi ve Manhattan Mesafesine çok benzemesi ile dikkat çekmektedir. Diğer bir adı ise Spearman Footrule mesafesidir. Bu mesafe

$$d_{ij} = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right| \tag{2.25}$$

şeklinde hesaplanır. Spearman mesafesindeki örneği göz önüne aldığımızda, Footrule mesafesi

$$d_{\mathbf{AB}} = |1-2| + |2-1| + |3-3| = 1 + 1 + 0 = 2 \tag{2.26}$$

olarak bulunur.

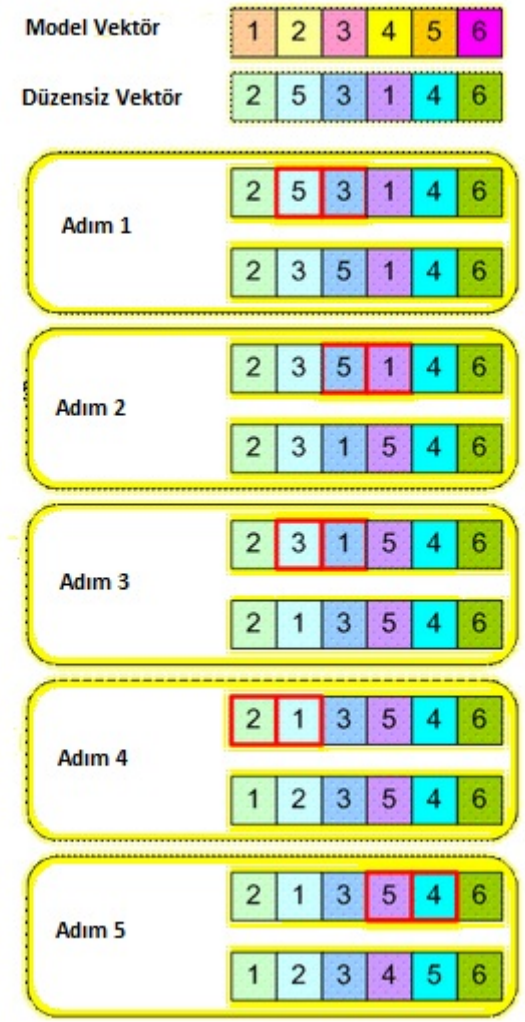
2.1.6.4. Kendall Mesafesi

Kendall mesafesi sıralı değişkenlerin düzensizliğini ölçmek amacıyla kullanılır ve model vektörle uyuşmayan uyumsuz çiftlerin minimum sayıda yer değiştirerek model vektöre dönüşmesini amaçlamaktadır. Uyumsuz çift kavramı, en az bir hücresi model vektörle uyuşmayan düzensiz vektördeki birbirine komşu olan hücrelerdir [21].

Kendall mesafesini hesaplamak için kullanılan algoritma, uyumsuz çiftlerin minimum sayıda yer değiştirmesidir:

- En az bir hücre model vektörle uyuşmayan düzensiz vektördeki birbirine komşu olan hücreler seçilir.
- Bu birbirine komşu olan hücrelerin sırası değiştirilir.

Altı ürün hakkında önem sırasını belirleyecek A ve B gibi iki müşteri olsun. Sıra vektörleri $A = [1, 2, 3, 4, 5, 6]$ ve $B = [2, 5, 3, 1, 4, 6]$ olarak verilsin. A'yı model vektör ve B'yi ise düzensiz vektör olarak düşünelim. Bu durumda, verilen koşullar altında minimum düzeyde işlem yapılarak düzensiz vektörü model vektöre dönüştürelim. Bu adımlar aşağıda diyagramlarla gösterilsin.



1. adımda düzensiz vektördeki bitişik olan 3 ve 5 hücreleri yer değiştirmiştir. 2. adımda ise, birinci adımdaki yer değiştirme ile oluşan düzensiz vektördeki bitişik olan 1 ve 5 hücreleri yer değiştirmiştir. Düzensiz şekilde bitişik olan hücrelerin minimum sayıda yer değiştirerek model vektöre ulaşması beklenir. Adımlar bu şekilde devam ettirildiğinde beşinci adımda, bitişik olan 5 ve 4 hücreleri yer değiştirerek model vektöre ulaşılmıştır. Böylelikle, A ve B vektörleri arasındaki Kendall mesafesi 5 olur.

2.1.6.5. Cayley Mesafesi

Cayley mesafesi model vektörle en az bir hücresi uyuşmayan düzensiz vektörlerin herhangi bir çift hücresinin minimum yer değiştirmesini toplayarak sıralı değişkenin düzensizliğini ölçer. Birbirine komşu olan hücreleri gerektiren Kendall mesafesinin aksine, Cayley mesafesinin hesaplanmasında düzensiz vektörlerde herhangi bir çift hücre

seçilebilir [22].

Cayley mesafesinin hesaplanması için algoritma, seçilen herhangi bir çift hücrenin yer değiştirmesinin minimum sayıda olması ile gerçekleşmektedir.

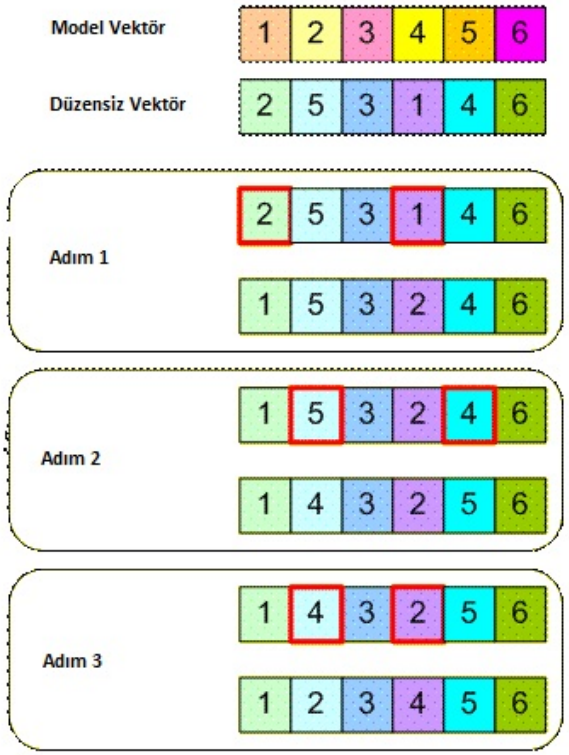
- Model vektörle en az bir hücresi uyuşmayan düzensiz vektörde herhangi bir çift seçilir.
- Bu çiftlerin yerleri değiştirilir.

Kendall mesafesini hesaplama probleminde olduğu gibi, Cayley mesafesini hesaplama probleminde de yer değiştirme yapmaktan ziyade amaç minimum yer değişikliği yapmaktır. Öncelik seçilen model vektör ve düzensiz vektör arasında en çok sayıda uyuşan hücreleri verecek şekilde yer değiştiren bir çift vermesidir. Uyuşan hücreler yer değiştirme sayısını azaltacaktır. Bu yüzden,

$$d_{\text{Kendall}} \geq d_{\text{Cayley}} \tag{2.27}$$

yazılır. Sıralama numaralandıktan sonra, mesafesi değişmeyecek olan sıralama invaryanttır. Örneğin, 1'den 10'a kadar olan bir sıralama 10'dan 1'e kadar olarak yeniden numaralandırılabilir. Diğer sıralı mesafelerde yeniden numaralandırma yapıldığında mesafe değişirken, Cayley ve Hamming mesafesinde yeniden numaralandırma yapılarak mesafe değişmemektedir [22].

Şimdi, Kendall mesafesinde göz önüne alınan örnek Cayley mesafesi için de ele alınsın. Buradaki amaç, düzensiz vektörü model vektöre dönüştürmek için hücrelerin minimum sayıda yer değiştirmesini sağlamaktır. Bu durum aşağıdaki diyagramla gösterildiğinde,



1. adımda düzensiz vektördeki 1 ve 2 hücreleri yer değiştirilmiştir. Ardından ikinci adımda, birinci adımda oluşan düzensiz vektördeki 4 ve 5 hücreleri yer değiştirmiştir. Hücreler bitişik veya bitişik olmayacak şekilde keyfi yer değiştirmesi devam ettirildiğinde son adımda 4 ve 2 hücreleri yer değiştirir. Böylelikle, her bir adımda hücrelerin keyfi bir şekilde yer değiştirilmesiyle model vektör oluşturulmuş ve A ile B vektörleri arasındaki Cayley mesafesi 3 olarak bulunmuştur.

2.1.6.6. Chebyshev/Maksimum Mesafe

Chebyshev mesafesi ayrıca maksimum mesafe olarakta bilinmektedir. Bu mesafe hem sıralı değikenler için hem de nicel değişkenler için kullanılmaktadır [17]. Bu mesafe iki nesnenin koordinatları arasındaki farklarının mutlak değerinin maksimumu olarak hesaplanır ve

$$d_{ij} = \max_k |x_{ik} - x_{jk}| \tag{2.28}$$

şeklinde gösterilir.

Varsayalım bir araç A noktasından B noktasına gitsin ve bunun için iki yol olsun. Birinci yolu kullanarak, 3 km. yol ya da ikinci yol kullanılarak 4 km. yol gitmesi mecburi olsun. Başlangıç noktası olan $A = (0, 0)$ ile ve $B = (3, 4)$ ile gösterilsin. Bu durumda, Chebyshev mesafesi herhangi bir yol boyunca A noktasından B noktasına gidilen maksimum yolun uzunluğu olarak hesaplanır. Yani;

$$d_{AB} = \max\{|0-3|, |0-4|\} = \max\{3, 4\} = 4$$

olarak hesaplanmaktadır.

2.1.6.7. Minkowski Mesafesi

Bu metrik mesafesinin genelleştirilmiş halidir. $m = 1$ olduğunda City Block mesafesi ve $m = 2$ olduğunda Öklidyen mesafesi elde edilmektedir. Cheybsev mesafesi $m = \infty$ olduğunda Minkowski mesafesinin özel durumu olur. Bu mesafe hem sıralı değişkenler hem de nicel değişkenler için kullanılabilir ve

$$d_{ij} = \sqrt[m]{\sum_{k=1}^{n} |x_{ik} - x_{jk}|^m} \tag{2.29}$$

şeklinde hesaplanır.

2.1.7. İki Grup için Mesafe Ölçüleri

Çoğu kümeleme algoritmaları hiyerarşik bir yapıya sahiptir. Hiyerarşik kümeleme tekniklerinde iki küme ya da bir küme ve bir nesne arasındaki mesafeyi hesaplamak için genellikle iki grup arasındaki mesafe ölçüleri kullanılır [23].

Aşağıda $C_1 = \{y_1, y_2, ..., y_r\}$ ve $C_2 = \{z_1, z_2, ..., z_s\}$ kümeleri sırasıyla bir bölümlemeden elde edilen iki kümeyi göstersin.

2.1.7.1. Ortalamaya Bağlı Mesafe

Nicel veriler için iki küme arasındaki benzersizliği ölçmenin popüler yolu iki kümenin ortalamaları arasındaki mesafeyi ölçmektir. C_1 ve C_2 gibi iki nicel veri kümesi var olsun.

Bu durumda, kümelerin ortalaması

$$\mu(C_j) = \frac{1}{|C_j|} \sum_{x \in C_j} x \ \ ; j = 1, 2 \tag{2.30}$$

olmak üzere, $d(.,.)$'ye göre C_1 ve C_2 gibi iki küme arasındaki ortalamaya bağlı mesafe

$$\begin{aligned} D_{\text{ortalama}}(C_1, C_2) &= d(\mu(C_1), \mu(C_2)) \\ &= d\Big(\frac{1}{|C_1|} \sum_{x \in C_1} x, \frac{1}{|C_2|} \sum_{y \in C_2} y\Big) \\ &= \frac{1}{|C_1||C_2|} d\Big(\sum_{x \in C_1} x, \sum_{y \in C_2} y\Big) \\ &= \frac{1}{|C_1||C_2|} \sum_{x \in C_1,\ y \in C_2} d(x, y) \\ &= \frac{1}{|C_1||C_2|} \sum_{x \in C_1,\ y \in C_2} (C_1, C_2) \end{aligned} \tag{2.31}$$

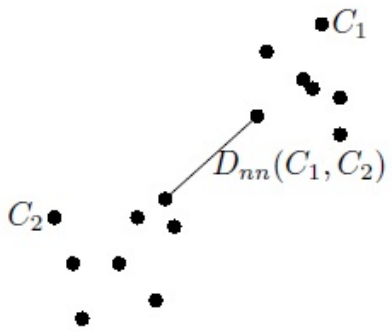
olarak tanımlanır.

2.1.7.2. En Yakın Komşu Mesafesi

$d(.,.)$'ye göre C_1 ve C_2 gibi iki küme arasındaki en yakın komşu mesafesi

$$D_{nn}(C_1, C_2) = \min_{1 \le i \le r,\ 1 \le j \le s} d(y_i, z_j) \tag{2.32}$$

olarak tanımlanır. Şekil 2.1'de, iki boyutluda en yakın komşu mesafesinin bir örneği verilmiştir.



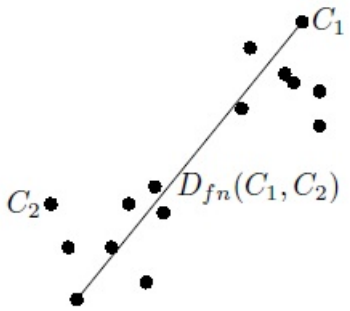
Şekil 2.1: İki küme arasındaki en yakın komşu mesafesi.

2.1.7.3. En Uzak Komşu Mesafesi

C_1 ve C_2 gibi iki küme arasındaki en uzak komşu mesafesi $d(.,.)$'ye göre

$$D_{fn}(C_1, C_2) = \max_{1 \leq i \leq r,\ 1 \leq j \leq s} d(y_i, z_j) \tag{2.33}$$

olarak tanımlanır. Şekil 2.2, iki boyutluda en uzak komşu mesafesinin bir örneğidir.



Şekil 2.2: İki küme arasındaki en uzak komşu mesafesi.

2.1.7.4. Ortalama Komşu Mesafesi

$d(.,.)$'ye göre C_1 ve C_2 gibi iki küme arasındaki ortalama komşu mesafesi

$$D_{\text{ort}} = \frac{1}{rs} \sum_{i=1}^{r} \sum_{j=1}^{s} d(y_i, z_j) \tag{2.34}$$

olarak tanımlanır.

2.1.8. Normalleştirme Metotları

Pek çok istatistiksel analiz yönteminde normalleştirme işlemi kullanılmaktadır. Buradaki tartışma, özelliklerin değerlerini nasıl $[0, 1]$ aralığına dönüşeceğidir.

Varsayalım $[f_{\min}, f_{\max}]$ aralığında değer alan fakat $[0, 1]$ aralığına dönüştürülmesi gereken özellikler var olsun. Orijinal özellik f, normalleştirilmiş özellik ise δ tarafından gösterilsin. Belli bir özellik değerini normalleştirmek için pek çok yol vardır. Öncelikle, bütün negatif değerler pozitife dönüştürülür ve daha sonra her bir sayı pay kısmından daha büyük olan bazı değerler tarafından bölünür [17]. Aşağıda orijinal özellik değerlerini

normalleştirilmiş değerlere dönüştürmek için bazı teknikler hakkında kısaca bilgi verilmiştir.

- Verilen özellik vektörünün aralığı biliniyorsa, bu durumda, özellik aşağıdaki denklem kullanılarak normalleştirilebilir.

$$\delta = \frac{f - f_{\min}}{f_{\max} - f_{\min}} \tag{2.35}$$

  Burada $f_{\min}$ özellik vektörünün minimum değerini ve $f_{\max}$ ise maksimum değerini gösterir. Eğer $f_{\min} = f$ ise, bu durumda, $\delta = 0$ olur. Ayrıca $f_{\max} = f$ olursa, $\delta = 1$ olur. Verilen bir veri kümesi için $f_{\min} = 0$ ise, Denklem (2.35) $\delta = \frac{f}{f_{\max}}$'a indirgenir.

- Kabul edelim ki, belli bir özelliğin maksimum değeri, bilinmesin ama özellik daima 0 değerini yada bazı pozitif değerler alabilsin. Belli özellik için n tane olası değerin bir toplamı varsa, bu durumda, i'nci özelliğin normalleştirilmesi aşağıdaki gibi yapılabilir.

$$\delta_i = \frac{f}{\sum_{i=1}^{n} f_i} \tag{2.36}$$

  Diğer yandan, Denklem (2.36) kullanılarak yapılan normalleştirme, Denklem (2.35) kullanılarak yapılan normalleştirmeden çok daha düşük değerde olacaktır çünkü $f_{\max} \leq (\sum_{i=1}^{n} f_i)$'dir.

- Belli bir özelliğin maksimum değeri bilinmiyor ve ayrıca negatif değer alıyorsa, bu durumda, normalleştirme aşağıdaki denklem kullanılarak yapılabilir.

$$\delta_i = \frac{|f|}{\sum_{i=1}^{n} |f_i|} \tag{2.37}$$

- Negatif değerlerin normalleştirilmesi: Pozitif ya da 0 değerli veri kümeleri için yukarıda tartışılan normalleştirme teknikleri kullanılır. Ancak, veri kümesinin bazı elemanları veya bileşenleri negatif değerli olduğu durumlarda, bu değerlerin minimumunun mutlak değeri bütün bileşenlere eklenir. Bu durumda, negatif değerlerden biri sıfır, diğerleri pozitif olacaktır. Böylelikle, yukarıda bahsedilen

normalleştirme tekniklerinden biri en son durumda elde edilen veriye uygulanabilir [17].

Varsayalım ki, veri kümesi $[-6, -9, 0, 6, 7]$ olsun. Bu sayıların minimumu -9'dur. Şimdi, $|-9| = 9$ sayısı veri kümesindeki beş değerlere eklenir. Bu durumda, değiştirilmiş sayılar $[-6+9, -9+9, 0+9, 6+9, 7+9] = [3, 0, 9, 15, 16]$'dır. Böylelikle, yukarıda bahsedilen tekniklerden herhangi biri bu veri kümesini normalleştimek için uygulanabilir.

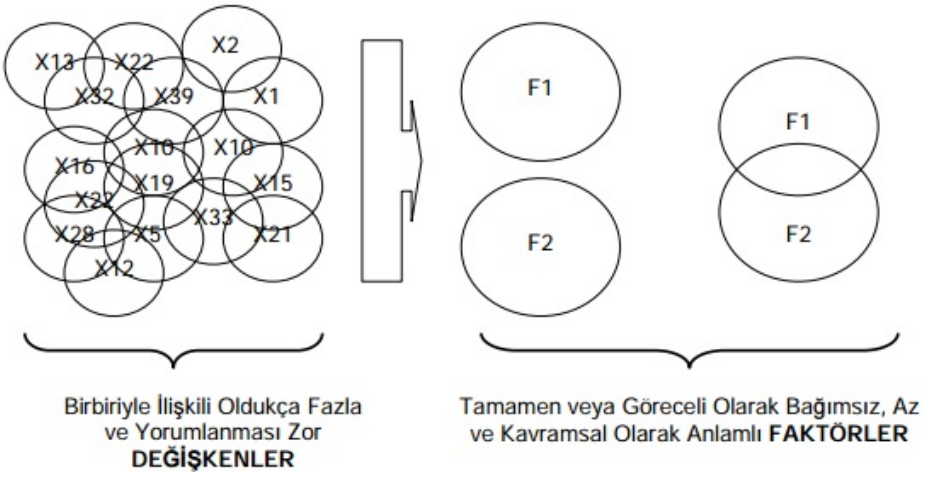
- z-Skor Normalleştirilmesi: Bu bir istatistiksel normalleştirme tekniğidir. Burada varsayım verilerin Normal dağılıma sahip olmasıdır. Normal dağılıma sahip herhangi bir veri kümesi ortalaması 0, varyansı 1 olan standart normal dağılıma

$$Z = \frac{X - \mu}{s} \tag{2.38}$$

dönüşümü kullanılarak dönüştürülür. Burada X orijinal veri kümesini, Z standart dönüşümlü veri kümesini, s standart sapmayı ve μ veri kümesinin ortalamasını temsil eder.

2.2. FAKTÖR ANALİZİ

Faktör analizi, birbirlerine bağımlı olan çok sayıdaki değişkeni, birbirlerinden bağımsız olacak şekildeki daha az sayıdaki değişkenlerle ifade etmeye yarayan istatistiksel analiz yöntemlerinden biridir [24]. Çok değişkenli istatistiksel analiz teknikleri arasında geniş bir uygulama alanına sahip olan faktör analizi, özellikle son kırk yıl içerisinde teknolojinin gelişmesiyle birlikte değişken sayılarının artması ve karmaşık hale gelmesinden kaynaklı çeşitli bilim dallarında kullanılmaktadır. Faktör analizi, ilk olarak 20. yüzyılın başlarında Spearman tarafından geliştirildi. Ancak, teknolojinin günden güne gelişmesiyle 1970'li yıllardan sonra faktör analizinin kullanımı sıklaşmıştır. Faktör analizinin iki temel amacı bulunmaktadır. Bunlardan birincisi, değişken sayısını azaltmaktır. İkinci amacı ise, değişkenler arasındaki ilişkilerin yapısını araştırmaktır. Böylelikle, değişkenler faktör analizi kullanılarak sınıflandırılabilmektedir. Buna ek olarak, faktör analizinin diğer bir amacı ise, veri matrisini diğer çok değişkenli yöntemler için hazır hale getirmektir [25]. Faktör analizi süreci gerçekleştirilirken bilgi kaybının minimum seviyede gerçekleşmesi hedeflenmektedir. Ayrıca, faktör analizinde ele alınan değişkenler arasında bağımlı ya da bağımsızlık gibi bir durum yoktur. Burada, bütün değişkenler birbirleriyle ilişkili değişkenlerdir. Bundan dolayı da, faktör analizi çok değişkenli analiz teknikleri olan varyans analizi, çoklu regresyon yöntemi, ayırma analizi, kanonik korelasyon gibi bir ya da daha fazla bağımlı ve bağımsız değişken arasındaki bağımlılık yapısını inceleyen yöntemlerden ayrılır [26].



Şekil 2.3: Faktör analizinin amacının görselleştirilmesi.

Faktör analizinin yukarıda bahsedilen temel bilgileri ve amaçları Tablo 2.3'de verilen bilgiler kullanılarak örnekleme yapılarak açıklanacaktır.

Tablo 2.3: Çocuğun Çevre ve Sıçrama Ölçümleri (cm).

Çocuk	KALÇAÇ	KARINÇ	GÖGÜSÇ	AKTİFS
Emin	63,5	57,75	62	25
Cemal	74,25	71	70,25	18,1
Dilek	70,25	60	64,75	20,1
Serhat	76,75	73,5	72,75	18,3

Tablo 2.3'de 4 adet çocuğa ilişkin kalça çevresi, karın çevresi, göğüs çevresi ve aktif sıçrama yüksekliğine ait bilgiler bulunmaktadır. Bu bahsedilen dört değişkenden ilk üç tane değişken çocukların vücutlarına ait bazı kısımların çevre ölçülerini gösterirken ve diğer değişken ise sıçrama yüksekliklerini belirtir. Bundan dolayı, bu örnekteki değişkenlerin iki faktörlü bir yapıya sahip olduğu ön bilgisi söz konusudur. Bu faktörler kısaca çevre ölçümleri faktörü ve sıçrama uzunluğu faktörü olarak adlandırılabilir. Ancak bu dört değişken arasında istatistiksel bakımdan faktörleşme olabilmesi için aşağıdaki özelliklerinin sağlanması gerekmektedir:

1. Birinci ve ikinci faktörü oluşturan değişkenlerin kendi içlerindeki ilişkilerinin yüksek olması gerekir.

2. Birinci faktördeki değişkenlerle ikinci faktördeki değişkenler arasındaki ilişki düşük olmalıdır [26].

Tablo 2.3'de verilen veri tabanına Pearson Korelasyon katsayısı uygulandığında yukarıda bahsedilen özelliklerin Tablo 2.4'de sağlandığı görülmektedir.

Tablo 2.4: Beş değişken arasındaki Pearson Korelasyon ilişki katsayıları.

Değişken	KALÇAÇ	KARINÇ	GÖGÜSÇ	AKTİFS
KALKAÇ	1	0,927	0,964	-0,966
KARINÇ	0,927	1	0,99	-0,84
GÖĞÜSÇ	0,964	0,99	1	-0,881
AKTİFS	-0,966	-0,84	-0,881	1

Yukarıdaki örnekte olduğu gibi verinin az ve düzenli olarak verilmesine pek rastlanmamaktadır. Karmaşık haldeki çok sayıda değişkene sahip olan veri tabanları verildiğinde, faktörleri belirlemek neredeyse imkansız hale gelecektir. Bundan dolayı da,

bu tür durumlarla karşılaşıldığında faktör analizini kullanmak birtakım işleri kolaylaştıracaktır [26].

2.2.1. Faktör Analizinin Aşamaları

Faktör Analizi genelde beş aşamada incelenir:

1. Verinin faktörlenebilir bir yapıda olup olmadığının ve gerekli varsayımların ve kısıtlayıcıların sağlanıp sağlanmadığının incelenmesi
2. Faktörleşmeyi gösterecek olan faktör yükleri matrisinin faktör türetme yöntemlerinden biri ile elde edilmesi
3. Özdeğerlerin incelenmesi, yamaç grafiğinin çizimi, vb. yaklaşımlarla kaç faktörün dikkate alıncağına ya da değişkenlerin kaç faktör altında toplanabileceğine karar verilmesi
4. İkinci aşamanın bir alt bölümü olarak da düşünülebilecek olan ve faktörleri daha kolay yorumlayabilecek bir yapıya getirme amacını güden faktör döndürme aşaması
5. Elde edilen bulguların tümel olarak yorumlanması [26].

2.2.2. Faktör Analizinin Uygulanabilirliği

Faktör analizinin uygulanabilirliğine ilişkin bazı sınırlamalar aşağıda verilmiştir. Bu sınırlamalar faktör analizi yapılmadan önce mutlaka incelenmelidir.

2.2.2.1. Korelasyon Katsayılarının İncelenmesi

Bir veri tabanının korelasyon katsayılarının incelenmesi, bu veri tabanına faktör analizinin uygulanıp uygulanmaması konusunda önemli bir kriterdir. Veri tabanına ait korelasyon katsayılarının büyük çoğunluğunun $0,30$ veya daha fazla bir ilişki göstermesi gerekir. Aksi durumda, bu veri tabanına faktör analizinin uygulanmasına gerek yoktur [26, 27]. Tablo 2.4'deki korelasyon matrisinde bütün korelasyon ilişki katsayılarının $0,30$'dan daha büyük olduğu görülmektedir. Bu sonuç doğrultusunda, bu veriye faktör analizi uygulanabilir.

2.2.2.2. Kısmi Korelasyon Katsayılarının İncelenmesi

Korelasyon matrisinde değişkenler arası ilişkilerin yüksek ve anlamlı olması, korelasyon matrisinin faktörler içerdiğinin kuvvetli bir doğrulayıcısı olmamaktadır. Bu bağlamda, kısmi korelasyon katsayılarının incelenmesi faktör analizinin uygulanabilirliğini araştırmak için iyi bir yaklaşımdır. Korelasyon katsayısı iki değişken arasındaki ilişkiyi gösterirken diğer değişkenlerin etkilerini dikkate almaz. Ancak, bazen geriye kalan değişkenlerin etkisi ortadan kaldırıldıktan sonra, iki değişken arasındaki ilişkinin miktarı incelenmek istenebilir. Diğer bir deyişle, ikincil ilişkilerin etkisi ortadan kaldırıldıktan sonra, iki değişken arasındaki gerçek ilişki incelenmek istenebilir. Bu inceleme kısmi korelasyon katsayıları yardımıyla yapılır.

Kabul edelim ki, X_1, X_2, X_3 birer değişken olsunlar. Bu durumda, X_3 değişkeni sabit tutulup, X_1 ve X_2 değişkenleri arasındaki ilişki incelenirse, kısmi korelasyon katsayısı

$$r(X_1, X_2 : X_3) = \frac{r(X_1, X_2) - r(X_1, X_3)r(X_2, X_3)}{\sqrt{[1 - r^2(X_1, X_3)][1 - r^2(X_2, X_3)]}} \tag{2.39}$$

olarak tanımlanır. Benzer şekilde, X_1, X_2, X_3 ve X_4 birer değişken olsun. Bu durumda, X_3 ve X_4 sabit tutulup, X_1 ve X_2 değişkenleri arasındaki ilişki incelenirse, kısmi korelasyon katsayısı

$$r(X_1, X_2 : X_3, X_4) = \frac{r(X_1, X_2 : X_3) - r(X_1, X_4 : X_3)r(X_2, X_4 : X_3)}{\sqrt{[1 - r^2(X_1, X_4 : X_3)][1 - r^2(X_2, X_4 : X_3)]}} \tag{2.40}$$

şeklinde formülize edilir [28, 27]. Bu bilgiler doğrultusunda, Tablo 2.3'ün kısmi korelasyon matrisi aşağıdaki gibi elde edilir.

Tablo 2.5: Tablo 2.3'deki verilerin kısmi korelasyon matrisi.

Değişken	KALÇAÇ	KARINÇ	GÖGÜSÇ	AKTİFS
KALKAÇ	-	-	-	-
KARINÇ	-0,862	-	-	-
GÖĞÜSÇ	0,939	0,979	-	-
AKTİFS	-0,961	-0,797	-0,837	-

Tablo 2.5'de görüldüğü üzere, değişkenler arasında pozitif ve negatif yönlü yüksek bir ilişki bulunmaktadır. Bu doğrultuda, verilen veri tabanının faktör içerdiği söylenilebilir.

2.2.2.3. Korelasyon Matrisinin Determinantının Alınması

Bir korelasyon matrisinin determinantı daima 0 ile 1 arasında değer almaktadır. Eğer korelasyon matrisinin determinantı 1 veya 1'e yakın değer alıyorsa, bu veriler arası ilişkinin düşük olduğu anlamına gelir. Böylelikle, değişkenlerin birbirlerine olan bağımlılığı azalmaktadır. Bundan dolayı, bu tür verilere faktör analizi uygulanması doğru değildir. Öte yandan, korelasyon matrisinin determinant değeri 0 veya 0'a yakın değerler alıyorsa, bu veriler arası ilişkinin yüksek olduğu anlamına gelir. Böylelikle, değişkenlerin birbirlerine olan bağımlılığı fazladır [26]. Bundan dolayı, bu tür verilere faktör analizi uygulanabilir. Tablo 2.3 için korelasyon matrisinin determinant değeri $0,0000085 \cong 0,00001$ olarak elde edilir. Böylelikle, bu veritabanının faktörlenebilir olduğu görülmektedir.

2.2.2.4. Korelasyon Matrisinin Tersinin Alınması

Yukarıdaki bölümlerde, bir veri tabanına faktör analizinin uygulanabilirliği incelenirken, değişkenler arasındaki korelasyonların yeterli olup olmadığına bakılmıştır. Faktör analizinin uygulanabilirliğini farklı bir bakış açısı ile gerçekleştiren yöntem korelasyon matrisinin tersinin alınmasıyla gerçekleştirilen durumdur. Bu incelemede, korelasyon matrisinin tersinin köşegen elemanları incelenir. Korelasyon matrisinin tersinin köşegen elemanlarına *varyans şişirme değeri* denilir ve VIF_j ile gösterilir. Veri tabanına faktör analizi uygulanabilmesi için VIF_j değerinin 5 veya 10'un üzerinde olması beklenilir [26]. Bu durumda, Tablo 2.4 matrisinin tersi aşağıdaki gibidir.

Tablo 2.6: Pearson Korelasyon Matrisinin Tersi.

Değişken	KALÇAÇ	KARINÇ	GÖGÜSÇ	AKTİFS
KALKAÇ	**396,59**	274,44	-506,822	167,297
KARINÇ	274,644	**255,679**	-424,107	160,438
GÖĞÜSÇ	-506,822	-422,107	**733,819**	-199,345
AKTİFS	167,297	106,438	-199,345	**76,3943**

Tablo 2.6'de görüldüğü üzere, VIF_j değerlerinin 5 veya 10'dan fazlasıyla büyük olduğu görülmektedir. Böylelikle, Tablo 2.3 veri tabanına faktör analizinin uygulanabilceği söylenebilir.

2.2.2.5. Kaiser-Meyer-Olkin Örneklem Yeterliği Ölçüsünün Elde Edilmesi

Bu ölçüt faktör analizinin uygunluğunun saptanmasında kullanılan diğer bir ölçüt türüdür. Bu ölçüt türünde veri tabanının kolerasyon katsayı değerleri ile kısmi korelasyon katsayı değerleri kullanılmaktadır. Ayrıca, Kaiser-Meyer-Olkin örneklem yeterliliği ölçüsü kısaca KMO olarak ifade edilmektedir. Bu ölçüt aşağıdaki gibi tanımlanmaktadır.

$$\text{KMO} = \frac{\sum_{i \neq j} \sum r_{ij}^2}{\sum_{i \neq j} \sum r_{ij}^2 + \sum_{i \neq j} \sum w_{ij}^2} \tag{2.41}$$

Burada, $\sum_{i \neq j} \sum r_{ij}^2$ korelasyon katsayılarının toplamı ve $\sum_{i \neq j} \sum w_{ij}^2$ kısmi korelasyon katsayılarının toplamıdır. KMO değeri 0 ile 1 arasında yer almaktadır. KMO oranın $0,5$'in üzerinde olması gerekir. Oran ne kadar yüksek olursa veri seti faktör analizi yapmak için o kadar iyidir denilebilir [26]. KMO verileri ve yorumları aşağıdaki gibidir:

Tablo 2.7: KMO için Nitelendirmeler.

KMO	Örneklem Yeterliliği
0,9-1	Çok İyi
0,8-0,89	İyi
0,7-0,79	Orta
0,6-0,69	Kötü
0,5-0,59	Çok Kötü
0,5'in altı	Kabul Edilemez

Tablo 2.3'deki veri tabanına göre elde edilen Pearson korelasyon katsayıları ve kısmi korelasyon katsayıları doğrultusunda KMO değeri

$$\begin{aligned} \text{KMO} &= \frac{(0,927)^2 + (0,964)^2 + ... + (-0,881)^2}{\Big[(0,927)^2 + (0,964)^2 + ... + (-0,881)^2\Big] + \Big[(-0,862)^2 + (0,939)^2 + ... + (-0,837)^2\Big]} \\ &= 0,516 \end{aligned}$$

dır. KMO değeri her bir değişken için de hesaplanabilir. Bu durumda,

$$\text{KMO}_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} w_{ij}^2} \tag{2.42}$$

olarak tanımlanır. Bu duruma bir örnek verilse, Tablo 2.3'de verilen veritabanındaki birinci değişken olan kalça çevresi değişkeni için KMO değeri

$$\begin{aligned} \text{KMO}_{\text{KALÇAÇ}} &= \frac{\left[(0,927)^2+(0,964)^2+(-0,966)^2\right]}{\left[(0,927)^2+(0,964)^2+(-0,966)^2\right]+\left[(-0,862^2)+(0,939)^2+(-0,961)^2\right]} \\ &= 0,516 \end{aligned}$$

olarak hesaplanır. Bu doğrultuda, KMO nitelendirmesine göre, örneklemin faktörlenebilirlik yeterliği çok kötü düzeydedir.

2.2.3. Faktör Analizinin Modeli

Faktör analizi bir denklem sistemi olarak ifade edilebilir ve genel anlamıyla $j = 1, 2, ..., p$ olmak üzere,

$$z_j = \alpha_{j1}F_1 + \alpha_{j2}F_2 + ... + \alpha_{jm}F_m + \varepsilon_j \tag{2.43}$$

eşitliği yazılabilir. Burada, α_{jm}, j'nci değişkenin m'nci faktör üzerindeki yükü, F_1, F_2, ..., F_m değişkenleri m adet temel faktör, ε_j artık faktör yükü ve z_j ise teoriksel olarak var olduğu düşünülen ancak bir takım ölçümler sonucu ölçülen j. gözlem değerinin gizil(gözlenemeyen) faktörlerdir. Böylece, bu model her bir gözlenen değişkenin artık değişkenlerle birlikte bu faktörlerin bir lineer kombinasyonu olan m adet temel faktörü bulundurduğunu varsayar [29]. Bu denklem sisteminin kanonik hali ise aşağıdaki gibidir:

$$\begin{aligned} z_1 &= \alpha_{11}F_1 + \alpha_{12}F_2 + ... + \alpha_{1m}F_m + \varepsilon_1 \\ z_2 &= \alpha_{21}F_1 + \alpha_{22}F_2 + ... + \alpha_{2m}F_m + \varepsilon_2 \\ . &= . \quad . \quad . \quad . \\ . &= . \quad . \quad . \quad . \\ . &= . \quad . \quad . \quad . \\ z_p &= \alpha_{p1}F_1 + \alpha_{p2}F_2 + ... + \alpha_{pm}F_m + \varepsilon_p \end{aligned} \tag{2.44}$$

Bu denklem kümesine *faktör örüntüsü* ya da sadece *örüntü* denir. Faktör analizi

sadece örüntüyü vermez, ayrıca değişkenler ve faktörler arasındaki ilişkiyi de verir. Bu ilişkiyi gösteren tabloya *faktör yapısı* veya sadece *yapı* denir [30].

2.2.4. Faktör Türetme Yöntemleri

Her şeyden önce faktör türetmenin birtakım amaçları bulunmaktadır. Bu amaçlardan en belirgin olanları veri tabanında bulunan değişkenlerin bağımlılığını yok edecek şekilde verideki karmaşıklığı gidermek ve karmaşık halde bulunan veri yapısını en iyi şekilde açıklamaktır. Literatürde pek çok sayıda faktör türetme yöntemi bulunmasına rağmen, faktör türetmenin temelinde temel bileşenler analizi (PCA), temel koordinatlar analizi (PCoA), en çok olabilirlik yöntemi, alfa faktörleştirme yöntemi, görüntü faktörleştirme yöntemi bulunmaktadır. Bu yöntemlerin seçimi ve kullanımı araştırmacının amacına göre farklılık göstermektedir. Bu bölümde, faktör türetme yöntemi olarak temel bileşenler analizi ile ilgili temel bilgiler verilmiştir.

2.2.4.1. Temel Bileşenler Analizi

18. yüzyılın başlarında Karl Pearson tarafından temel bileşenler analizi üzerine çalışmalar yapılmıştır. Daha sonra 1933 yılında Hotelling tarafından bu yöntem geliştirilmiştir. Bu yöntem aralarında korelasyon bulunan karmaşık yapıdaki p adet değişkenin $k < p$ olacak şekilde birbirinden bağımsız basit yapı oluşturan k adet değişkeni, lineer bileşenleri olan değişkenlerle ifade etme yöntemidir. Bu analiz yöntemiyle değişkenleri birbirinden bağımsız hale getirmek ve veri tabanının boyutunun indirgenmesi amaçlanmaktadır [31]. Bunların yanı sıra, kullanılan diğer istatistiksel analiz yöntemlerine kolaylık sağlamak gibi görevi de bulunmaktadır.

Temel bileşenler analizinde, m tane gözlem ve p tane değişkene sahip olan bir X veri matrisi düşünülsün. X veri matrisinin p adet değişkenin lineer bileşenlerini bulabilmek için varyans-kovaryans ya da korelasyon matrisinin özdeğer-özvektörleri kullanılır. Eğer, orijinal veri matrisi kullanılacaksa, varyans-kovaryans matrisi kullanılır. Aksine, standartlaştırılmış veri matrisi kullanılacaksa, korelasyon matrisi kullanılır. Bu iki yöntemle hesaplanan ifadelerinin sonuçları birbirinden farklı çıkmaktadır. Bu durumda, hangi yöntemin kullanılmasının uygun olup olmayacağı ile ilgili ortaya bir sorun

çıkmaktadır. Uzman araştırmacılar, değişkenler eğer aynı veya karşılaştırılabilir birime sahipseler varyans-kovaryans matrisinin kullanılmasını önermektedirler. Öte yandan, bu durumların sağlanmaması durumunda ise, korelasyon matrisinin kullanılmasını önermektedirler.

$X_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \\ . \\ . \\ . \\ x_{m1} \end{pmatrix}$, $X_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ x_{32} \\ . \\ . \\ . \\ x_{m2} \end{pmatrix}$, ... , $X_p = \begin{pmatrix} x_{1p} \\ x_{2p} \\ x_{3p} \\ . \\ . \\ . \\ x_{mp} \end{pmatrix}$ olmak üzere, X veri matrisi aşağıdaki gibi gösterilir.

$$X = \begin{pmatrix} x_{11} & x_{12} & . & . & . & x_{1p} \\ x_{21} & x_{21} & . & . & . & x_{2p} \\ x_{31} & x_{32} & . & . & . & x_{3p} \\ . & . & & . & & . \\ . & . & & . & & . \\ . & . & & . & & . \\ x_{m1} & x_{m2} & . & . & . & x_{mp} \end{pmatrix}$$

Matematiksel olarak temel bileşenler (faktörler) $X_1, X_2, ..., X_p$ değişkenlerinin lineer kombinasyonlarıdır ve X veri matrisinin temel bileşenleri

$$\begin{aligned} Y_1 &= \alpha_{11}X_1 + \alpha_{21}X_2 + ... + \alpha_{p1}X_p \\ Y_2 &= \alpha_{12}X_1 + \alpha_{22}X_2 + ... + \alpha_{p2}X_p \\ . &= . \quad . \quad . \quad . \\ . &= . \quad . \quad . \quad . \\ . &= . \quad . \quad . \quad . \\ Y_p &= \alpha_{1p}X_1 + \alpha_{2p}X_2 + ... + \alpha_{pp}X_p \end{aligned} \tag{2.45}$$

olarak ifade edilir. Ancak, orijinal veri matrisi yerine standartlaştırılmış veri matrisi kullanılırsa, temel bileşenler

$$\begin{aligned} Y_1 &= \alpha_{11}Z_1 + \alpha_{21}Z_2 + ... + \alpha_{p1}Z_p \\ Y_2 &= \alpha_{12}Z_1 + \alpha_{22}Z_2 + ... + \alpha_{p2}Z_p \\ . &= . \quad . \quad . \quad . \\ . &= . \quad . \quad . \quad . \\ . &= . \quad . \quad . \quad . \\ Y_p &= \alpha_{1p}Z_1 + \alpha_{2p}Z_2 + ... + \alpha_{pp}Z_p \end{aligned} \tag{2.46}$$

şeklinde olur. Burada, $Y_1, Y_2, ..., Y_p$ değişkenleri temel bileşenler (faktörler), Z_1, Z_2,...,Z_p orjinal veri matrisinin standartlaştırılmış değişkenleri ve α_{ij} i. değişkenin j.temel bileşendeki faktör yüküdür. Ayrıca, dikkate alınacak olan faktör yüklerinin bütün değişkenlerdeki değişimin $\%70$'inden fazlasını açıklaması gerekir. Bu toplam değişimi açıklama değeri veri matrisinden oluşan özdeğerler (λ) sayesinde hesaplanmaktadır ve $k = 1, 2, ..., p$ olmak üzere

$$k\text{'ıncı temel bileşenin açıkladığı değişkenlik oranı} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p} \tag{2.47}$$

olarak ifade edilir. Burada λ_k, k. bileşene ait özdeğerdir. Özetlenirse, temel bileşenler yöntemi için aşağıdaki adımlar izlenir.

- m ölçümündeki p değişkene ait veri matrisinin korelasyon matrisi oluşturulur,
- Korelasyon matrisinin öz değerleri ve öz vektörleri hesaplanır,
- Öz değerlerden temel bileşenlerin toplam varyansı açıklama oranları bulunur,
- Elde edilen özdeğer ve özvektör matrisi tarafından faktör yükleri bulunur,
- Faktör yükleri matrisinin transpozesi ile standartlaştırılmış veri matrisi çarpılarak temel bileşen değerleri bulunur [31].

2.2.5. Önemli Faktörlerin Belirlenmesi

Birbirine bağlı olan çok sayıdaki değişkeni birbirinden bağımsız olacak şekildeki daha az sayıda faktör üreterek veri tabanı hakkında açıklama yapmak önemli bir işlemdir. Bunun yanı sıra, bu üretilen faktörlerden hangisinin önemli bir konuma sahip olduğunun tespiti de önemli bir yer teşkil eder. Bu değerlendirmeyi yaparken bazı kriterler bulunmaktadır. Üretilen faktörler arasında önemli faktörleri belirlemeyi sağlayan birtakım uygulaması kolay yaklaşımlar başlıklar halinde aşağıda belirtilmiştir.

2.2.5.1. Özdeğer Ölçütü

Bu ölçüt bir faktörün açıklayıcılığının en azından bir değişkenin açıklayıcılığı kadar olması düşüncesiyle Kaiser tarafından önerilmiştir ve Kaiser ölçütü olarak da bilinmektedir. Bu ölçüt temel bileşenler analizinde sıklıkla tercih edilen bir yaklaşımdır. Bu yaklaşım yukarıda da belirtildiği gibi bir faktörün açıklayıcılığının en azından bir değişkenin açıklayıcılığı kadar olması mantığına dayanır [26]. Özdeğerler ölçütüne göre, özdeğerleri 1'den büyük olan faktörler *anlamlı faktörler* olarak nitelendirilirken, özdeğeri 1'den küçük olan faktörler ise *anlamsız faktörler* olarak nitelendirilmektedir. Tablo 2.3'te veri tabanına özdeğer ölçütü uygulandığında, $3,785$ değerine sahip olan birinci özdeğer 1'den büyük olması nedeniyle anlamlı faktör olarak alınabileceği söylenebilir.

2.2.5.2. Varyans Yüzdesi Ölçütü

Önemli faktörlerin belirlenmesinde kullanılan diğer önemli bir yaklaşım türü ise, varyans yüzdesi ölçütüdür. p ile veri tabanındaki değişken sayısı ve m ile önemli özdeğerlerin sayısı ifade edilirse, bu durumda

$$\Big(\sum_{j=1}^{m}\frac{\lambda_i}{p}\Big)\geq\frac{2}{3}\quad\text{veya}\quad\Big(\sum_{j=1}^{m}\frac{\lambda_i}{p}\Big)\geq 0,666 \tag{2.48}$$

koşulunu sağlayan en küçük m değeri temel bileşen sayısını belirler [26]. Buna göre, ilk özdeğer olan $3,785$ değeri için bu koşul $3,785/4 = 0,94625$ şeklinde sağlanmaktadır. Dolayısıyla, türetilecek önemli faktör sayısının 1 tane olması yeterlidir. Ancak, bu koşul sağlanmasaydı bu durumda, ikinci özdeğerin eklenmesiyle koşulun sağlanıp sağlanmadığı

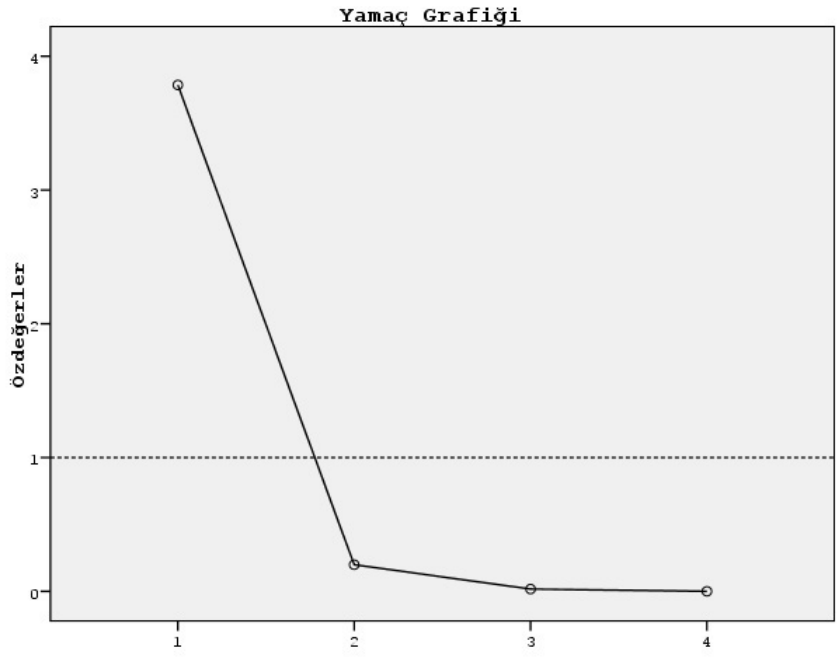
incelenecekti. Eğer, ilk iki özdeğerle bu koşul sağlansaydı, önemli faktörlerin sayısının 2 olduğu söylenebilirdi.

Ayrıca, sosyal bilimlerde ilgilenilen durumun hassasiyetine göre $0,666$ değeri değişebilmektedir.

2.2.5.3. Yamaç Grafiği Yaklaşımı

Önemli faktörleri belirlemedeki diğer önemli bir yaklaşım yamaç grafiğidir. Bu yaklaşımda dikey eksen özdeğerleri, yatay eksen faktör sayısını gösterir. Grafikte dik eğim veren noktalar alınır. Yüzeysel, düz eğim veren noktalar alınmaz. Grafiğin yatay eğime geçtiği noktadan itibaren yatay bir çizgi çizilir. Bu çizginin üzerinde kalan noktaların sayısı boyut olarak kabul edilmektedir [32].

Tablo 2.9'deki verilere ilişkin yamaç grafiği Şekil 2.4'de verilmiştir. Özdeğer ölçütü yaklaşımı olan anlamlı faktörleri belirleyen 1 kriteri Şekil 2.4'e uygulandığında 1 özdeğer ölçütünü sadece Faktör1 aştığından faktör sayısının 1 alınması gerektiği görülmektedir. Bu sayının, özdeğer ölçütü ve varyans yüzdesi ölçütünde ortaya çıkan faktör sayısıyla aynı olduğuna dikkat edelim.



Şekil 2.4: Tablo 2.3 Verisi için Yamaç Grafiği.

2.2.6. Özdeğerler-Özvektörler Matrisi

A $= [a_{ij}]$ $n \times n$ tipindeki bir karesel matris olmak üzere, A$x = \lambda x$ olacak şekilde λ skaler sayı ve sıfırdan farklı x vektörlerini bulma problemine *özdeğer-özvektör problemi* denir. Burada, λ skaler sayısı A matrisinin *özdeğerleri* ya da *karakteristik değerleri* olarak adlandırılmaktadır. A$x = \lambda x$ denklemi üzerinde basit cebirsel işlemler yapılırsa, A$x = \lambda x$ denklemi,

$$(\mathrm{A} - \lambda I)x = 0 \tag{2.49}$$

şeklinde yazılabilir. Burada, I $n \times n$ tipindeki birim matristir. Denklem (2.49)'den elde edilen λ özdeğerlerine karşılık gelen x vektörlerine ise *özvektör* ve bu özvektörlerin oluşturduğu matrise ise *özvektörler matrisi* denilir. Faktör analizinde, veri matrisinin korelasyon matrisi elde edildikten sonra, korelasyon matrisinin özdeğerleri ile özvektörlerini içeren matris oluşturulur. Bununla birlikte, gerekli birtakım işlemler yapılarak bir sonraki bölümde bahsedilecek olan faktör yükleri matrisi elde edilir. Tablo 2.4'de verilen veri matrisine ilişkin Pearson korelasyon matrisine ait özvektörler matrisi Tablo 2.8'de verilmiştir.

Tablo 2.8: Tablo 2.4'de Verilen Korelasyon Matrisinin Özdeğerleri ve Özvektörler Matrisi.

Değişken	1	2	3	4
KALKAÇ	-0.509	-0.228	0.651	0.513
KARINÇ	-0.496	0.56	-0.517	0.413
GÖĞÜSÇ	-0.506	0.361	0.299	-0.723
AKTİFS	0.486	0.709	0.466	0.206
Özdeğerler	3,785	0,198	0,016	0,00071

2.2.7. Faktör Yükleri Matrisi, Adlandırılması ve Özellikleri

Veri tabanındaki birbirleriyle ilişkili olan değişkenlerin faktörlerle ilişkisini ortaya koyan katsayıya *faktör yükü* adı verilir. Faktör yükü değişkenlerle faktör arasındaki ilişkiyi ortaya koyan bir değer olduğundan dolayı, değerlerin yüksek olması beklenir. Bir faktörle yüksek düzeyde ilişki veren değişkenlerin oluşturduğu bir küme var ise, bu bulgu o maddelerin birlikte bir kavramı-yapıyı-faktörü ölçtüğü anlamına gelmektedir [33].

Faktör yükleri, verilerin korelasyon matrisinin her bir özdeğerin karekökü ile o özdeğere karşılık gelen özvektör elemanları çarpılarak elde edilir.

$$\alpha_{ji} = e_{ji}\sqrt{\lambda_i} \quad ; j = 1, 2, ..., p, \quad i = 1, 2, ..., m < p \tag{2.50}$$

Burada α_{ji} j. değişkenin i. faktör üzerindeki yüküdür. Aslında, faktör yükü i.faktör ile j.değişken arasındaki korelasyon katsayısını vermektedir. Yukarıdaki ilişki, matris formunda aşağıdaki gibi ifade edilir.

$$\mathrm{A}_{p\times m} = \mathrm{C}_{p\times m}\mathrm{D}_{\sqrt{\lambda}} \tag{2.51}$$

Burada, $m \times m$ boyutlu $\mathrm{D}_{\sqrt{\lambda}}$ matrisi, köşegen elemanları $\sqrt{\lambda_i}$ ($i = 1, 2, \ldots, m$) olan köşegen bir matristir. A matrisi $p \times m$ boyutlu ise, *Faktör Yükleri Matrisi* ya da *Faktör Yapı Matrisi* olarak adlandırılır. Bu matris, faktör analizi açısından çok önemli olup, hangi değişkenlerin hangi faktör etrafında yoğunlaştığını göstermektedir. Ayrıca, faktörlerin yorumlanmasına büyük katkı sağlar [34, 35, 36]. C matrisi ise $p \times m$ boyutlu olup özvektörler matrisidir.

Tablo 2.3'de verilen örneğin faktör yükleri matrisi aşağıdaki gibi elde edilir:

Tablo 2.9: Faktör Yükleri Matrisi ve Özdeğerler.

Değişken	Faktör 1	Faktör 2	Faktör 3	Faktör 4
KALÇAÇ	-0.991	$-0,101$	$0,083$	0
KARINÇ	$-0,965$	$0,249$	$-0,066$	0
GÖGÜSÇ	$-0,986$	$0,16$	$0,038$	0
AKTİFS	$0,946$	$0,316$	$0,059$	0
Özdeğerler	$3,785$	$0,198$	$0,016$	$0,00071$
$\%$	$94,623$	$4,968$	$0,408$	$4,5 \cdot 10^{-16}$
Birikimli$\%$	$94,623$	$99,592$	100	100

Bu bilgiler çerçevesinde Tablo 2.9'deki özdeğerler satırından Faktör 1'in varyansı olan $\lambda_1 = 3,785$ toplam varyansın büyük bir kısmını açıklarken Faktör 2'nin varyansı olan $\lambda_2 = 0,198$ de toplam varyansın büyük bir kısmını açıklamaktadır. Bu bilgi $\%$ satırından daha iyi yorumlanabilir. Buna göre, Faktör 1'in toplam varyansın $\%94,623$'ünü $(3,785/4)$, Faktör 2'nin toplam varyansın $\%4,968$'ini, Faktör 3'ün toplam varyansın $\%0,408$'ini ve Faktör 4'ün toplam varyansın $\%4,5 \cdot 10^{-16}$'sını açıkladığı bilgisine

ulaşılmaktadır. Birikimli $\%$ satırından da toplam varyansın $\%99,592$'sinin [(3,785+0,198)/7] ilk iki faktör tarafından açıklandığı görülmektedir. Dolayısıyla, verideki toplam değişimin aslında $\%94,623$'sini Faktör 1 ile dört değişkenli veri tabanının açıklanması yeterli olmasına rağmen, burada ilk iki faktörün açıkladığı yüzdelik orana bakarak ileriki bölümlerde saçılım grafiğinin iki boyutlu gösterilişinin sağlanması gerçekleştirilecektir. Ayrıca, ilk üç değişkenin ikinci faktör üzerindeki ağırlığı birinci faktöre göre daha fazla iken, AKTİFS değişkeninin birinci faktör üzerindeki ağırlığının daha fazla olduğu görülmektedir.

Faktör yükleri Pearson korelasyon katsayıları olup -1 ile $+1$ arasında değişmektedir ve onun gibi yorumlanır. Değişkenlerle faktörler arasındaki ilişkiyi çok iyi bir şekilde açıklayabilmek için faktör yükü değerinin $0,7$ ve üzerinde bir değere sahip olması gerekir [26]. Faktör yükü değerleri ve buna ilişkin yorumlar aşağıdaki Tablo 2.10'da verilmiştir.

Tablo 2.10: Faktör Yükü Değerlerinin Yorumu.

Faktör Yükü Değerleri	Yorumu
0,7'in üstü	En İyi Açıklayan Değer
0,4-0,69	Kabul Edilebilir Değer
0,3-0,39	En Düşük Düzeydeki Anlamlı Değer
0,3'ün altı	Anlamsız Değer

Her değişkenin faktör yüklerinin kareleri toplamı 1'e eşittir. Örneğin, kalça çevresi değişkeni için $(-0,99)^2 + (-0,101)^2 + (0,082)^2 + (0,0136)^2 = 1$ olarak elde edilmektedir. Kalça çevresi değişkeninin birinci faktör tarafından açıklanan varyansı $(-0,991)^2 = 0,982$ olup, birinci faktör kalça çevresi değişkenindeki toplam varyansın $0,982$'sini açıklamaktadır denir. Bu varyans "*açıklanan varyans*", "*ortak varyans*" olarak adlandırılır. Bu kavram kısaca ilgili değişkendeki değişimin $\%$ kaçının üretilen faktörler tarafından açıklandığını ifade etmektedir.

Ortak varyanslar genellikle h^2 ile gösterilmektedir. Böylelikle, ilk iki faktörün yükleri ve ortak varyanslar tablosu Tablo 2.11'de verilir.

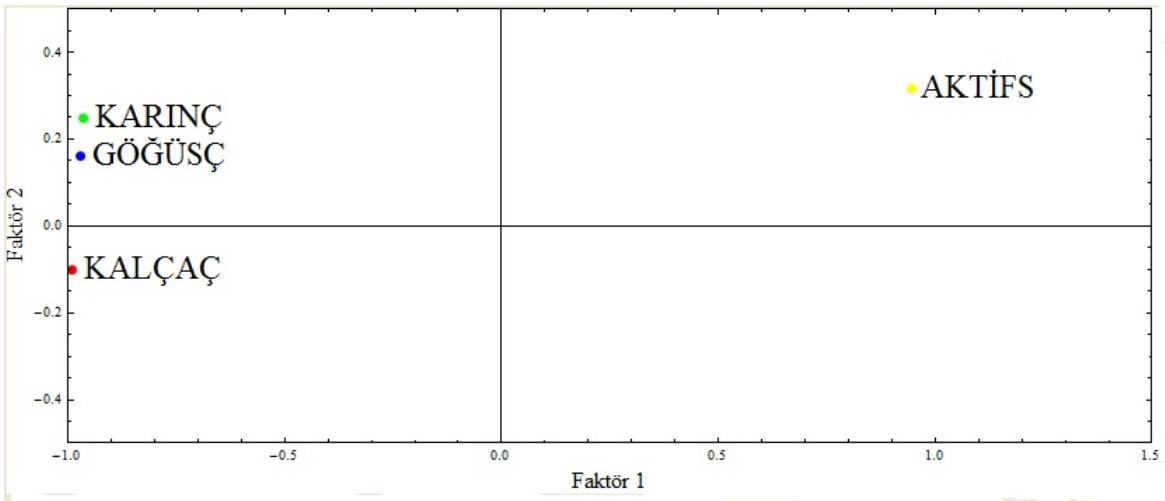
Tablo 2.11: İlk İki Faktör Yükleri ve Ortak Varyanslar.

Değişken	Faktör 1	Faktör 2	h_1^2	h_2^2	h_{1+2}^2
KALÇAÇ	$-0,991$	$-0,101$	$0,982$	$0,0102$	$0,9922$
KARINÇ	$-0,965$	$0,249$	$0,931$	$0,062$	$0,993$
GÖGÜSÇ	$-0,986$	$0,16$	$0,972$	$0,0256$	$0,9976$
AKTİFS	$0,946$	$0,316$	$0,894$	$0,0998$	$0,9938$
Özdeğerler	$3,785$	$0,198$	$3,785$	$0,198$	$3,983$
$\%$	$94,623$	$4,968$	$94,623$	$4,968$	$99,591$
Birikimli$\%$	$94,623$	$99,592$	$94,623$	$99,592$	$99,592$

Tablo 2.11'e göre, 1. ve 2. faktör en az $\%99,22$ değeri ile "KALÇAÇ" değişkenindeki toplam varyansı, en fazla $\%99,76$ ile de "GÖGÜSÇ" değişkenin varyansını açıklar.

2.2.8. Faktör Yüklerinin Saçılım Grafiği

Saçılım grafiği, iki farklı değişken arasındaki ilişkinin sebebini göstermeksizin ilişkinin ne kadar güçlü olup olmadığını noktasal olarak iki boyutlu düzlem üzerinde göstermeye yarayan görsel bir araçtır [37]. Bir önceki bölümde, beş değişkeninin Faktör 1 ve Faktör 2 yüklerine ait bilgiler Tablo 2.11'de verilmiştir. Bu faktör yüklerine ait veriler doğrultusunda değişkenler arasındaki ilişkinin daha da anlaşılır bir şekilde ifade edilmesi saçılım grafiği tarafından verilmektedir. Şekil 2.5'de verilen ilk iki faktör yüklerine ait saçılım grafiğinden de görüldüğü üzere değişkenler çevre ölçülerine ve sıçrama ölçülerine göre farklı gruplar oluşturmuştur.



Şekil 2.5: Faktör Yüklerinin Saçılım Grafiği.

2.2.9. Faktör Katsayıları ve Skorları

Bütün değişkenler faktör yapısı içerisinde farklı ağırlıkta bulunmaktadır. Bu değişkenlerden bazıları ana rol oynarken bazıları yardımcı rol oynarlar. Belirlenen faktör yüklerinden faydalanılarak bütün değişkenlerin faktör skorları elde edilebilir [9]. Eğer orijinal değişken değerleri kullanılırsa, orijinal değişken değerlerinin ortalama değerlerden farkı alınarak elde edilen değer ile faktör yüklerinin çarpımlarının toplaması faktör skorunu verir ve matematiksel olarak

$$\text{FS}_i = \alpha_{1i}(x_i - \overline{x}) + \alpha_{2i}(x_i - \overline{x}) + ... + \alpha_{pi}(x_i - \overline{x}); \quad j = 1, 2, ..., p \tag{2.52}$$

şeklinde ifade edilir. Burada, α_{ji} notasyonu j. değişkenin i. faktörünü ve x_i notasyonu ise x veri matrisinin i. değişkenini ifade etmektedir. Faktör skorları, her bir veri değeri $Z = \frac{X - \overline{X}}{\sigma}$ ile standartlaştırarak da elde edilebilir.

Daha önceki bölümlerde elde edilen bilgiler ve bu bölümde bahsedilen bilgiler doğrultusunda, Tablo 2.3'de verilen KALÇAÇ, KARINÇ, GÖĞÜSÇ ve AKTİFS değişenlerinin faktör skorlarını Denklem (2.52)'yi kullanarak elde edebilmek için öncelikle bu değişkenlere ait temel istatistiksel değerleri ve standartlaştırılmış veri matrisi verilsin.

Tablo 2.12: Değişkenlere ait Temel İstatistiksel Bilgiler.

Değişken	Ortalama	Standart Sapma
KALÇAÇ	$71,1875$	$5,782066$
KARINÇ	$65,5625$	$7,84319$
GÖĞÜSÇ	$67,4375$	$4,930243$
AKTİFS	$20,375$	$3,211827$

Böylelikle, değişkenlere ait standartlaştırılmış veri matrisi Tablo 2.13'de verilir.

Tablo 2.13: Tablo 2.3'de verilen Veri Matrisinin Standartlaştırılmış Z Veri Matrisi.

Değişken	ZKALÇAÇ	ZKARINÇ	ZGÖĞÜSÇ	ZAKTİFS
Emin	$-1,32954$	$-0,99608705$	$-1,102886815$	$1,4399903$
Cemal	$0,464799$	$0,69327658$	$0,570458697$	$-0,708319$
Dilek	$-0,162139$	$-0,70921398$	$-0,54510497$	$-0,085621$
Serhat	$0,962026$	$1,01202444$	$1,07753309$	$-0,64604974$

Faktör skorunun elde edilmesinin daha da anlaşılır hale gelmesi için Emin'in birinci

faktöre ilişkin skoru ile Serhat'ın ikinci faktöre ilişkin skoru yukarıda tanımlanan hususlar doğrultusunda aşağıdaki gibi hesaplanır.

$$\begin{aligned}
\text{FS}_{1,\text{Emin}} &= (-0,991) \times \text{ZKALÇAÇ} + (-0,965) \times \text{ZKARINÇ} + (-0,986) \times \text{ZGÖĞÜSÇ} \\
&+ (0,946) \times \text{ZAKTİFS} \\
&= (-0,991) \times (-1,32954) + (-0,965) \times (-0,99608705) \\
&+ (-0,986) \times (-1,102886815) + (0,946) \times (1,4399903) \\
&= 1,31757414 + 0,961224003 + 1,0874464 + 1,362230824 \\
&\cong 4,728
\end{aligned}$$

ve

$$\begin{aligned}
\text{FS}_{2,\text{Serhat}} &= (-0,101) \times \text{ZKALÇAÇ} + (0,249) \times \text{ZKARINÇ} + (0,16) \times \text{ZGÖĞÜSÇ} \\
&+ (0,316) \times \text{ZAKTİFS} \\
&= (-0,101) \times (0,962026) + (0,249) \times (1,01202444) + (0,16) \times (1,07753309) \\
&= (0,316) \times (-0,64604974) \\
&= -0,09716463 + 0,25199409 + 0,17240529 - 0,20415172 \\
&\cong 0,12
\end{aligned}$$

Diğer bileşenler için benzer şekilde işlemler gerçekleştirilse faktör skorları Tablo 2.14'de verilir.

Tablo 2.14: 4 Faktör için Değişkenlerin Z Skorları ile Özdeğerlerinin Çarpılması Sonucunda Elde Edilen Yeni ve Dik Faktör Skorları.

Çocuk	FS_1	FS_2	FS_3	FS_4
Emin	$4,72$	$0,16$	$-3,7 \cdot 10^{-3}$	$-6,4 \cdot 10^{-5}$
Cemal	$-2,36$	$-6,16 \cdot 10^{-3}$	$-0,026$	$-8,4 \cdot 10^{-3}$
Dilek	$1,35$	$-0,27$	$6,4 \cdot 10^{-3}$	$-4,9 \cdot 10^{-5}$
Serhat	$-3,6$	$0,12$	$0,01$	$1,1 \cdot 10^{-4}$
Ortalama	0	0	0	0
Standart Sapma	$3,74$	$0,19$	$0,016$	$0,004$

Tablo 2.14'deki faktör skorları yaklaşık değerler olduğundan dolayı, bu tablodan elde edilen ortalamaların yaklaşık sıfıra ve varyans değerlerinin ise yaklaşık olarak özdeğerlere eşit olduğu görülmektedir. Bundan dolayı, temel bileşenler yöntemi son

derece ilginç yaklaşımlar bütünüdür. Ayrıca, üretilen temel bileşenler birbirlerinden bağımsızdır. Yani, $i \neq j$ olmak üzere

$$\text{cov}(\text{FS}_i, \text{FS}_j) = 0 \tag{2.53}$$

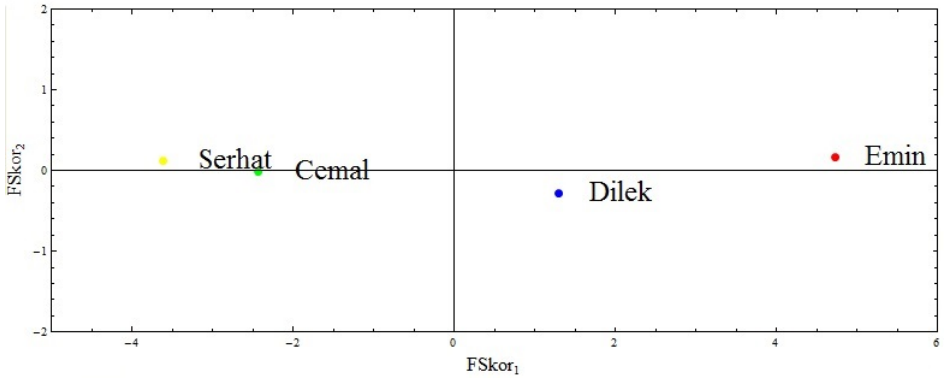
dır. Ancak yukarıdaki örnekte tablolarda yaklaşık değerler alınarak hesap yapıldığından dolayı, üretilen temel bileşenlerin kovaryans değerleri sıfıra yakın değerler çıkacaktır. Temel bileşenlerin kovaryans matrisi aşağıdaki gibi olacaktır.

Tablo 2.15: Temel Bileşenlerin Kovaryans Matrisi.

	FS_1	FS_2	FS_3	FS_4
FS_1	$10,633425$	$-0,01042021$	$-0,000013625$	$0,004817621$
FS_2	$-0,01042021$	$0,028219591$	$0,000603312$	$0,00000102695$
FS_3	$-0,000013625$	$0,000603312$	$0,000165087$	$0,0000041305$
FS_4	$0,004817621$	$0,00000102695$	$0,0000041305$	$0,0000013215$

2.2.10. Faktör Skorlarının Saçılım Grafiği

Faktör skorları yardımıyla çizilen saçılım grafikleri, gözlemlerdeki ilişkinin görülmesi sürecinde sıklıkla kullanılır. Bu örnek için toplam değişkenliğin $\%99,592$'sini açıklayan ilk iki faktör için faktör skorlarının saçılım grafiğinin yapılması uygun görülür. Böylelikle, çocukların ilk iki faktöre göre faktör skorlarının saçılım grafiği Şekil 2.6'de gösterilmiştir.



Şekil 2.6: Çocukların Faktör Skorlarına Göre Saçılım Grafiği.

2.2.11. Faktör Analizi Sonuçlarının Kullanım Yerleri

Faktör analizinin yukarıdaki bölümlerde bahsedilen amaçlarının dışında da kullanım alanları bulunmaktadır. Bunlardan bazıları şunlardır:

- Faktör yükleri, bir anlamda değişkenlerin ilgili faktör üzerindeki objektif ağırlıklarıdır. Bu nedenle, bu yükler kullanılarak elde edilen faktör skorları gözlemlere ilişkin birer objektif değerlendirme skoru olarak düşünülebilmekte ve bu skorlar yardımıyla gözlemleri sıraya dizdirme işlemleri yapılabilmektedir. Bu süreçte değişkenler arasındaki ilişkiler de dikkate alınmaktadır. Sıraya dizdirme işlemleri çeşitli dallarda sıklıkla kullanılmaktadır. Örneğin, ülkelerin, bölgelerin, illerin sağlık/ekonomik/gelişmişlik gibi pek çok özelliği gösteren değişken vardır. Bu göstergelerin her biri açısından ülkeler, bölgeler, illler gibi değişkenler dikkate alınarak sıralamada faktör yükleri yardımıyla elde edilen faktör skorları dikkate alınabilmektedir [26].

- Faktör analizi, kümeleme analizindeki bazı kararsızlıkları gidermede de kullanılmaktadır. Bu amaçla da faktör skorlarından yararlanılır. Bu tür bir yaklaşım, değişkenler arasındaki korelasyon/kovaryans matrisinden yola çıkarak faktör analizinin gözlemler arasındaki ilişkileri saptamadaki kullanımına örnek oluşturur. Bu yaklaşım, faktör analizinin önemli uygulama alanlarından biridir. Yine, çok fazla değişken olduğu durumlarda, değişkenler arasında faktörleşme varsa, ilgili faktör skorları dikkate alınarak kümeleme analizi yapmak olasıdır [26].

- Faktör türetme amacıyla sıklıkla kullanılan ve birçok kaynakta ayrı bir bölüm altında ele alınan temel bileşenler yöntemi yardımıyla farklı regresyon modelleri oluşturmak da söz konusudur. Bu tür yaklaşımlar, bağımsız değişkenler arasında çoklu bağlantı olduğu durumlarda tercih edilmektedir [26].

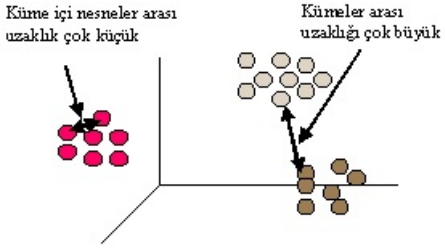
2.3. KÜMELEME ANALİZİ ve DENDROGRAM

Benzerlik ya da mesafe ölçülerinin sayısının çok olmasından dolayı kümeleme analizi yapmak uğraştırıcı bir süreçtir. Bundan dolayı da kümeleme analizinin tam anlamıyla net bir tanımı bulunmamaktadır [38]. Kümeleme problemi karmaşık ve büyük veri tabanlarında uygun bir parametre seçimi ve optimal çözümü elde etmek için etkili bir araştırma gerektirir. Ayrıca, en iyi olacak şekilde tek bir ölçü tanımlamak genellikle zor bir süreçtir. Bununla birlikte, problemi aynı anda optimize etmek için birden fazla amaç belirlemek de zordur. Bu sadece sayısal olarak zahmet gerektiren bir süreç değil aynı zamanda kesin bir çözüm elde etmemeye yol açabilecek sorunlarla da karşılaşmaya neden olabilir. Bu yüzden, sağlam, hızlı ve yakın çözümler sağlayan hem tek hem de çok amaçlı optimizasyon tekniklerinin kullanılması uygun görülmektedir [17].

Bu bölümde, sıklıkla kullanılan bazı kümeleme algoritmaları hakkında bilgiler verilmiştir. Ancak, her şeyden önce "Kümeleme nedir?", "Kümelemenin amaçları nelerdir?", "Kümeleme sonuçları ne tür görsellerle ifade edilir?" gibi sorulara cevap aranır.

2.3.1. Kümeleme Tanımı

Çeşitli nesnelerin belli kriterler çerçevesinde grup oluşturmasına *kümeleme* adı verilir. Kümelemede, benzer veya aynı özelliklere sahip olan nesneler aynı küme içerisinde, farklı özelliklere sahip nesneler ise farklı kümeler içerisinde yer alır.



Şekil 2.7: Kümeleme Yapısı.

Şekil 2.7'de görüldüğü gibi, nesneler arasında her ne kadar yakın mesafe varsa, bu nesneler birbirlerine benzemektedir [39]. Öte yandan, her ne kadar nesneler arası mesafe artarsa bu nesnelerin birbirlerinden farklı özellikler taşıdığı anlamına gelmektedir. Matematiksel olarak kümeleme, bazı benzerlik yada mesafe metriklerine göre uzayları

K bölgeye bölümleme sürecidir [17]. Burada K'nın değeri, bilinen bir önsel bilgi olabilir yada olmayabilir. Herhangi bir kümeleme tekniğinin amacı, verilen **X** veri kümesinin

$$\begin{aligned} \sum_{j=1}^{n} u_{kj} &\geq 1;\ k = 1, ..., K \\ \sum_{k=1}^{K} u_{kj} &= 1;\ j = 1, ..., n \ \text{ ve} \\ \sum_{k=1}^{K} \sum_{j=1}^{n} u_{kj} &= n \end{aligned}$$

koşullarını sağlayan $U(\mathbf{X})$ bölümleme matrisini geliştirmektir. $K \times n$ boyutundaki $U(\mathrm{X})$ bölümleme matrisi $U = [u_{kj}]$ olarak temsil edilebilir. Burada, u_{kj} C_k kümesi için $\overline{x_j}$ modelinin elemanıdır.

2.3.2. Kümeleme Analizinin Amaçları

Küçük boyutlu veri tabanlarındaki değişkenleri gözlemlemek, sınıflandırmak, aralarındaki ilişkiyi araştırmak, düzenlemek gibi bazı birtakım işlemlerin yapılması nispeten kolay bir süreçtir. Fakat, bu durum büyük boyutlu veri tabanları için aynı kolaylıkta değildir. Çünkü, büyük veri tabanlarındaki değişkenlerin karmaşık halde dizilmiş olması, benzer özellikteki değişkenleri ayırt edememeye sebep olmaktadır. Bu nedenle, araştırmacılar bu gibi durumlar karşısında istatistiksel analiz yöntemlerinden biri olan kümeleme analizini kullanmaktadır [26]. Bu sorunlar doğrultusunda, kümeleme analizinin amaçları verilen bir veri tabanınındaki verileri düzenlemek, belli kriterler çerçevesinde gruplara ayırarak büyük ve karmaşık yapıları anlaşılır bir hale getirmek ve uygun bir modeli bulmaktır. Bu sürecin gerçekleşmesi araştırmacıya çok büyük kolaylıklar sağlamaktadır. Diğer yandan, bu süreç gerçekleştirilirken bilgi kaybı ortaya çıkmaktadır. Kümeleme analizinin asıl amacı bu tür sınıflandırmalar yapılırken bilgi kaybını minimize etmektir.

2.3.3. Kümeleme Analizinin Kullanım Alanları

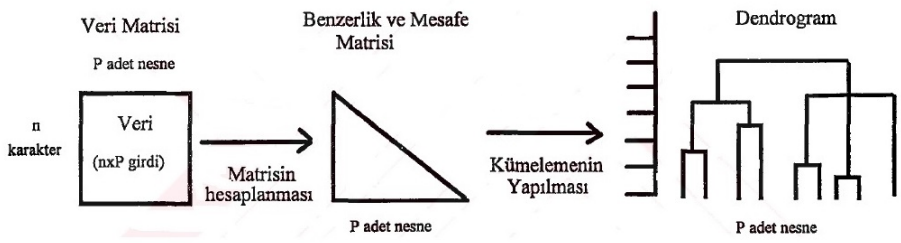
Kümeleme teknikleri, araştırma problemlerinde geniş bir uygulama alanına sahiptir. Özellikle, kamu kuruluşlarının ve büyük şirketlerin geleceğe yönelik planlama ve stratejik çalışmalarında önemli bir rol oynamaktadır. Ayrıca, tıpta hastalıkların tedavilerinin sınıflandırılmasında, psikiyatri dalında önemli hastalıkların belirtilerini teşhis etmede kullanılmaktadır [40]. Bunların yanı sıra, günümüzde teknolojinin devamlı olarak gelişmesiyle büyük veri kitlelerinden kurtulup özet veriler elde etmeyi amaçlayan "*veri madenciliği*" önemli görevler üstlenmiştir ve veri madenciliği çalışmalarında kümeleme analizi gibi pek çok istatistiksel analiz teknikleri kullanılmaktadır. Veri madenciliği kısmındaki bazı uygulama alanlarına örnek olarak örüntü tanıma, görüntü işleme, ekonomi bilimi (özellikle pazar araştırmasında), dünya çapında ağ (world wide web) üzerinde doküman sınıflandırılması, benzer ortak arkadaş grupları keşfetme, istatistik, biyoloji ve makine öğrenme verilebilir [39].

Yukarıda bahsedilen kullanım alanlarından da anlaşılacağı üzere kümeleme analizi tıptan psikiyatriye, ekonomiden veri madenciliğine, ziraatten arkelojiye kadar pek çok alanda kullanılan çok değişkenli istatistiksel analiz tekniğidir.

2.3.4. Kümeleme Analizinin Uygulama Aşamaları

Kümeleme analizinin aşamaları Şekil 2.8'de görüldüğü gibi aşağıdaki biçimde sıralanabilir.

i) Değişkenler ve veri matrisi belirlenir.
ii) Değişkenler arasındaki benzerlik ya da uzaklıklara göre benzerlik ya da uzaklık matrisi oluşturulur.
iii) Uygun kümeleme algoritması kullanılarak bir önceki adımda elde edilen benzerlik ya da uzaklık matrisine göre kümeleme yapılır.
iv) Oluşturulan küme şekillerle özet bilgi haline getirilip veri tabanı hakkında yorumlar yapılır [2].



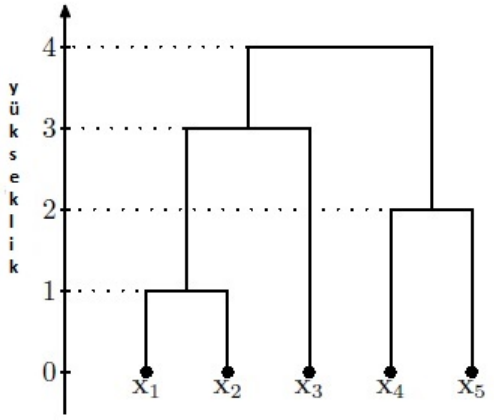
Şekil 2.8: Kümeleme Analizinin Aşamaları.

2.3.5. Dendrogram

Dendrogram, kümeleme sürecinde veritabanlarını özet bilgi haline getiren ve birtakım değerler verilmiş bir ağaç diyagramı olarak nitelendirilen görsel bir ifadedir. Bununla birlikte, bir dendrogram $A \cap B \neq \varnothing$ olmak üzere A ve B veri noktalarının bütün alt kümeleri için

$$h(A) \leq h(B) \Leftrightarrow A \subseteq B$$

özelliğini sağlayan bir yükseklikle ilişkili olan her bir iç düğümde n-ağaca sahip bir göstergedir [23]. Burada, $h(A)$ ve $h(B)$ sırasıyla A ve B'nin yüksekliğini gösterir. Şekil 2.9'de beş veri noktalı bir dendrogram örneği gösterilmektedir [23]. Burada, noktalı çizgiler iç düğümlerin yüksekliğini ifade ediyor. (x_i, x_j) veri noktalarının her bir çifti için, h_{ij} hem x_i hem de x_j'ye sahip olan en küçük kümeyi belirleyen iç düğümün yüksekliği olsun. Bu durumda, h_{ij}'nin küçük bir değeri x_i ve x_j arasında yüksek bir benzerlik gösterir. Örneğin, Şekil 2.9'de verilen dendrogramda $h_{12} = 1$, $h_{23} = h_{13} = 3$ ve $h_{14} = 4$'tür [23].



Şekil 2.9: Beş veri noktasının bir dendrogramı.

Dendrogramdaki yükseklikler aşağıdaki ultrametrik şartı sağlar [23].

$$h_{ij} \leq \max\{h_{ik}, h_{jk}\} \tag{2.54}$$

Aslında, Denklem (2.54)'de verilen ultrametrik şart ayrıca bir dendrogram için gerek ve yeter koşuldur. Matematiksel olarak bir dendrogram, $c : [0, \infty) \to E(D)$ fonksiyonu

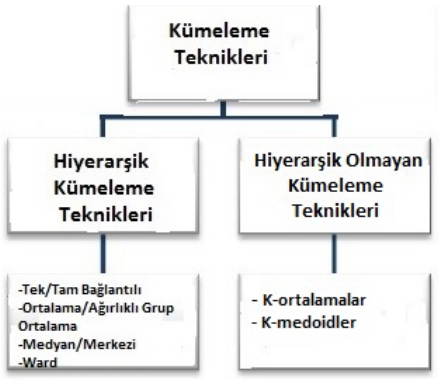
1. $h \leq h'$ise,$c(h) \subseteq c(h')$
2. $c(h)$ daima $D \times D$'nin elemandır.
3. Bazı küçük $\delta > 0$ için $c(h + \delta) = c(h)$

şartlarını sağlamasıyla gerçekleşir [23]. Burada, D verilen veri kümesi ve $E(D)$, D ile denklik ilişkilerinin kümesidir. Bir örnek olarak, aşağıda verilen c fonksiyonu Şekil 2.9'de verilen dendrogramın bilgilerini içermektedir.

$$c(h) = \begin{cases} \{(i,i) \mid i = 1,2,3,4,5\} & ;0 \leq h < 1 \\ \{(i,i) \mid i = 3,4,5\} \cup \{(i,j) \mid i,j = 1,2\} & ;1 \leq h < 2 \\ \{(i,j) \mid i,j = 1,2\} \cup \{(i,j) \mid i,j = 4,5\} \cup \{(3,3)\} & ;2 \leq h < 3 \\ \{(i,j) \mid i,j = 4,5\} \cup \{(i,j) \mid i,j = 1,2,3\} & ;3 \leq h < 4 \\ \{(i,j) \mid i,j = 1,2,3,4,5\} & ;4 \leq h \end{cases}$$

2.3.6. Bazı Kümeleme Teknikleri

Kümeleme sürecinde hangi benzerlik ya da mesafe ölçüsünün seçileceğine karar verdikten sonra, uygulamada hangi kümeleme tekniğinin seçileceğine karar verme gereksinimi duyulur [2]. Kümeleme tekniklerinin seçimi kullanılacak olan veri tabanının türüne ve amaçlarına göre farklılık göstermektedir [41]. Bundan dolayı da, literatürde pek çok sayıda kümeleme algoritması bulunmaktadır. Bu algoritmalar hiyerarşik ve hiyerarşik olmayan kümeleme teknikleri olmak üzere iki ana sınıfa ayrılabilir. Bu tekniklerin her birinin ortak amacı kümeler arasındaki farklılıkları ve kümeler içi benzerlikleri maksimum düzeye çıkarmaktır. Yani, ortak amaç küme içi homojenlik arttırılırken kümeler arası homojenlik azaltılır [42].



Şekil 2.10: Bazı kümeleme teknikleri.

2.3.7. Hiyerarşik Olmayan Kümeleme Teknikleri

Hiyerarşik olmayan kümeleme teknikleri en temel kümeleme analizi tekniklerinden biridir. Bu teknikte, n ile veri tabanındaki nesne sayısı ve k ile oluşturulacak küme sayısı gösterilsin. $k \leq n$ olmak üzere hiyerarşik olmayan kümeleme teknikleri n adet nesneyi k adet kümeye bölümler. Böylece, oluşan her bölüm bir kümeyi gösterir. Kümeler tarafsız bölme ölçütü olarak nitelendirilen bir ölçüte uygun oluşturulduğu için aynı kümedeki nesneler birbirlerine benzerken, farklı kümedeki nesneler birbirinden farklıdırlar [39]. Bununla ilgili bazı teknikler bulunmaktadır ve bu teknikler genellikle bir kümenin özeti olarak görülen küme merkezi ya da ağırlık merkezi fikrine dayanır. Örneğin, bir küme reel değerli noktalara sahip olduğunda, küme merkezi genellikle küme içinde bulunan noktaların aritmetik ortalamasıyla tanımlanır. Eğer bir küme sözel yazılardan veya dosyalardan oluşuyorsa, bu kümenin ağırlık merkezi küme içinde bulunan cümlelerin anahtar kelimelerinin listesi olabilir [17]. Bazı popüler hiyerarşik olmayan kümeleme teknikleri K-ortalamalar algoritması, K-medoidler kümeleme algoritmasıdır. Aşağıdaki alt başlıklarda, bu algoritmalardan bahsedilmiştir.

2.3.7.1. K-Ortalamalar Kümeleme Algoritması

K-ortalamalar algoritması 1967 yılında Mac Quenn tarafından geliştirilen en eski kümeleme yöntemlerinden biridir. Bu algoritma, n adet veriden oluşan bir veri tabanında K adet küme oluşturur [43]. K-ortalamalar algoritmasına göre kümeleme yapılırken, kümeler içi benzerliklerin maksimum olması beklenilir. Öte yandan ise kümeler arası benzerliklerin minimum düzeyde olması amaçlanır. Bundan dolayı, kümeleme sürecinde

türüne özgü olarak hata kareleri toplamı kullanılır:

$$J = \sum_{j=1}^{n} \sum_{k=1}^{K} u_{kj} \times ||\overline{x_j} - \overline{z_k}||^2 \tag{2.55}$$

Burada, k kümesinin merkezi $\overline{z_k}$ ile ve verinin j'nci noktası $\overline{x_j}$ ile gösterilmektedir.

K-ortalamalar algoritmasının adımları aşağıda gösterilmiştir:

Adım 1: $x_1, x_2, ..., x_n$ noktalarından rastgele olarak K tane $z_1, z_2, ..., z_k$ olacak şekilde küme merkezleri seçilir.

Adım 2: $i = 1, 2, ..., n$ ve $j \in 1, 2, ..., k$ olmak üzere C_j kümesi için x_i noktası atanır gerek ve yeter koşul $||x_i - z_j|| < ||x_i - z_p||; \;\; p = 1, 2, ..., K$ ve $j \neq p$'dır.

Adım 3: Aşağıdaki gibi yeni küme merkezleri $z_1^*, z_2^*, ..., z_k^*$ hesaplanır:

$$z_i^* = \frac{\sum_{x_j \in C_i} x_j}{n_i}; \;\; i = 1, 2, ..., K \tag{2.56}$$

Burada, n_i C_i kümesine ait elemanların sayısıdır.

Adım 4: $i = 1, 2, ..., K$ olmak üzere, $z_i^* = z_i$ ise, bu durumda algoritma sonlanır. Aksi durumda, 2'nci adıma geri dönülerek devam edilir.

Küme merkezlerinin güncellenmesi J'nin merkezlere göre diferansiyeli alınıp sıfıra eşitlenmesiyle elde edilir [17]. Bu analizin temel amacı aşağıdaki gibi J'nin minimize olmasıdır.

$$\frac{\mathrm{d}J}{\mathrm{d}\overline{z_k}} = 2 \sum_{j=1}^{n} u_{kj}(\overline{x_j} - \overline{z_k})(-1) = 0; \;\; k = 1, 2, ..., K \tag{2.57}$$

$$\sum_{j=1}^{n} u_{kj}\overline{x_j} - \overline{z_k} \sum_{j=1}^{n} u_{kj} = 0 \tag{2.58}$$

$$\overline{z_k} = \frac{\sum_{j=1}^{n} u_{kj}\overline{x_j}}{\sum_{j=1}^{n} u_{kj}} \tag{2.59}$$

olarak bulunur. Kümeleme sürecinde $\sum_{j=1}^{n} u_{kj}$ ifadesi hiçbir şey ifade etmemektedir ama bu k'ncı kümeye ait elemanların sayısıdır yani, n_k dır. Bu yüzden, güncellenen denklem aşağıdaki gibi özetlenir:

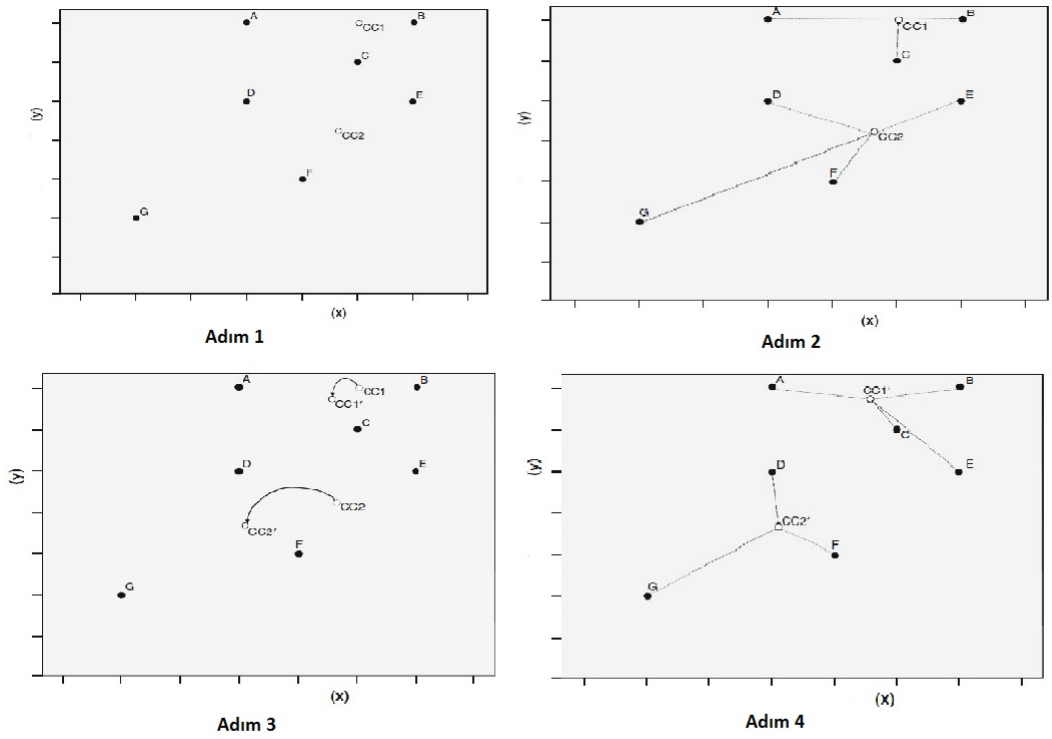
$$\overline{z_j^*} = \frac{\sum_{\overline{x_j} \in C_i} \overline{x_j}}{n_i} \tag{2.60}$$

J'nin minimumunu gerçekten sağlamak için ikinci türev,

$$\frac{d^2 J}{d\overline{z_k}^2} = 2 \sum_{j=1}^{n} u_{kj} \tag{2.61}$$

pozitif olmalıdır. $\sum_{j=1}^{n} u_{kj} \geq 1$ olmasından dolayı yukarıdaki denklemin sağ tarafının pozitif olduğu aşikar bir şekilde görülmektedir.

Bu algoritma için keyfi alınan A,B,C,D,E,F,G nesneleri ve CC1 ve CC2 gibi iki kümenin merkezleri düşünülsün. Bu durumda, K-ortalamalar algoritmasını daha iyi anlayabilmek amacıyla aşağıdaki şekil verilmiştir.



Şekil 2.11: K-ortalamalar algoritmasının adımları.

K-ortalamalar algoritması pek çok sayıda dezavantaja sahiptir. K-ortalamalar optimal olmayan değerlere yakınsayabilir ama başlangıç kümesinin merkezlerine bağlıdır. Ayrıca, bu algoritma kümelerin bir hiperküre olduğunu kabul eder. Son olarak, aykırı

değerler için K-ortalamalar sağlam olmayabilir. Bu nedenlerden dolayı pek çok kümeleme algoritmaları geliştirilmiştir [17].

2.3.7.2. K-Medoidler Kümeleme Algoritması

Hiyerarşik olmayan kümeleme algoritmaları içinde K-ortalamalar algoritması gibi önemli bir yere sahip olan diğer bir yöntem ise K-medoidler algoritmasıdır. Bu algoritma 20. Yüzyılın sonlarına doğru Kaufmann ve Rousseeuw tarafından K-ortalamalar algoritmasının gürültü ve aykırı değerler sonucu oluşan dezavantajları ortadan kaldırmak amacıyla geliştirilmiştir. K-ortalamalar algoritması küme merkezini küme içinde bulunan noktaların aritmetik ortalaması olarak bulduğundan dolayı, küme merkezi bulunurken istisnai noktalar küme merkezini değiştirir. Ancak, bu durumun K-medoidler algoritması için aynı olduğu söylenemez. Çünkü, K-medoidler algoritmasında küme merkezi için küme içerisinde bulunan noktaların orta noktasına bakılır [39]. Diğer yandan, K-medoidler algoritmasının temeli, verinin çeşitli yapısal özelliklerini temsil eden k tane nesneyi bulma esasına dayanır.

Bu algoritmanın adımları aşağıdaki gibi izlenir:

1. K-medoidler algoritmasına göre kümeleme yapılırken, ilk olarak karışık halde verilmiş olan veri tabanı sıralanır.

2. Sıralama işlemi yapıldıktan sonra, her verinin başlangıçta rastgele belirlenmiş olan merkez noktalarına göre uzaklığı alınır. Veriler en yakın olduğu merkez noktasının kümesine dahil olur.

3. Bu adımdan sonra, her küme için elemanlarının ortalaması alınır. K-medoidler algoritmasında küme elemanı olmayan bir değer merkez noktası olamaz. Bu nedenle, küme ortalamasına en yakın olan nokta yeni merkez noktası olur.

4. Sonraki adımda, tekrar her verinin merkez noktalarına olan uzaklığı hesaplanır ve veriler en yakın olduğu merkez noktasının kümesine dahil edilir. Küme

elemanlarının ortalaması alınıp, ortalamaya en yakın noktalar yeni merkez noktaları olarak belirlenir.

5. Kümeleme işlemi sonucu, bir sonraki adımda aynı çıkana kadar bu işlem tekrarlanır.

Görüldüğü gibi K-medoidler kümeleme algoritmasını K-ortalamalar algoritmasından ayıran özellik, merkez noktaların belirlenme şeklidir. Küme elemanı olmayan bir değerin merkez noktası kabul edilmemesi ise gürültülü verilerin kümelere dahil edilmesine rağmen, küme üzerindeki etkilerini ortadan kaldırır.

2.3.8. Hiyerarşik Kümeleme Teknikleri

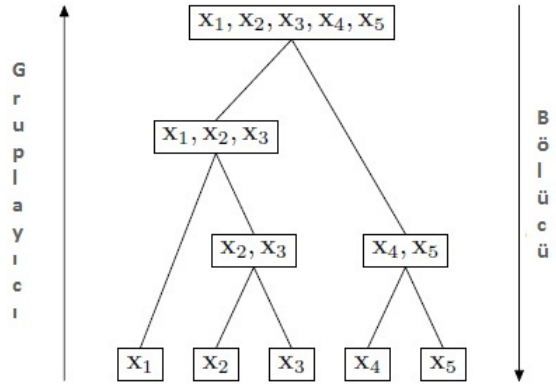
Hiyerarşik kümeleme yöntemi, kümelerden bir eleman silinmesi ya da eklenmesiyle oluşan ağaca benzeyen bir yapıyı ifade eden aşamalar grubudur [45]. Hiyerarşik kümeleme algoritmaları veri tabanındaki değişkenler arasındaki uzaklık ya da benzerliklerini göz önünde bulundurarak veriler için bir küme oluşturmaya ve bu kümelere girecek olan birimlerin hangi uzaklık veya benzerlik ölçüsünün kullanılacağını belirlemeye çalışan yöntemlerdir [46]. Hiyerarşik kümelemelerin sonuçları birimler arasındaki ilişkiyi özetleyen ve açıklayan dendrogramlar tarafından görsel olarak ifade edilir.

Hiyerarşik kümeleme tekniklerinin uygulanabilmesi için gözlem sayısının çok büyük olmaması gerekir. Bu veritabanındaki gözlemlerin büyüklüğü 1000'e kadar olabilmektedir. Eğer büyük veri tabanlarında hiyerarşik kümeleme tekniği kullanılacaksa, bu durumda gözlemlerden örneklem alınması gerekir [26].

Bu algoritmanın temelinde gruplayıcı ve bölücü olmak üzere iki tür hiyerarşik kümeleme tekniği vardır:

Gruplayıcı hiyerarşik kümeleme tekniklerinde her birim veya her gözlem ilk başta bir küme olarak düşünülür. Ardından birbirlerine en yakın iki küme veya gözlemler yeni bir küme oluşturacak şekilde birleştirilir. Adımlar bu şekilde devam ettirildiğinde, bütün birimler tek bir küme oluşturana kadar bu süreç devam eder. Bu süreç dendrogram veya ağaç diyagramı olarak ifade edilen şekillerle gösterilir. Bölücü hiyerarşik kümeleme

tekniklerinde ise süreç gruplayıcı hiyerarşik yöntemin tam tersi olmakla birlikte benzerlikleri az olan birimler ayrıştırılarak küçük kümeler oluşturulur. Ayrıca, burada verilen verilerdeki her bir birim tek kalana kadar bu süreç devam eder.



Şekil 2.12: Gruplayıcı Hiyerarşik Kümeleme ve Bölücü Hiyearşik Kümeleme.

Gruplayıcı hiyerarşik kümeleme algoritması çok yaygındır ve bu bölümde, sadece gruplayıcı hiyerarşik kümeleme algoritmasının metotları işlenecektir. Bunun için herşeyden önce gruplayıcı kümeleme algoritmasında gerekli olacak olan Lance-Williams formülü aşağıda tanımlanmıştır.

Tanım 2.1. (Lance-Williams Formülü): Lance-Williams formülü gruplayıcı hiyerarşik kümeleme algoritmalarında iki küme arasındaki mesafeyi hesaplamak için kullanılır. Lance-Williams C_i ve C_j gibi iki kümenin birleşiminden oluşmuş $C = C_i \cup C_j$ ve C_k kümeleri arasındaki mesafeyi veren bir formülü önermektedir [23]. Bu formül

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \alpha_i D(C_k, C_i) + \alpha_j D(C_k, C_j) \\ &+ \beta D(C_i, C_j) + \gamma |D(C_k, C_i) - D(C_k, C_j)| \end{aligned} \tag{2.62}$$

tarafından tanımlanır. Burada, $D(.,.)$ iki küme arasındaki mesafedir. Hiyerarşik kümeleme algoritmalarında α_i, α_j, β ve γ parametrelerinin uygun bir değer seçimi kullanılır. Bu değerler Tablo 2.16'de verilmektedir [23].

Tablo 2.16: Lance-Williams formülündeki parametreler için bazı yaygın olarak kullanılan değerler, burada, $n_i = |C_i|$ (C_i'deki noktaların sayısı) ve $\sum_{ijk} = n_i + n_j + n_k$'dır.

Algoritma	α_i	α_j	β	γ
Tek Bağlantı	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{-1}{2}$
Tam Bağlantı	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Ward Metotu	$\frac{n_i+n_j}{\sum_{ijk}}$	$\frac{n_j+n_k}{\sum_{ijk}}$	$\frac{-n_k}{\sum_{ijk}}$	0
Grup Ortalaması	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
Ağırlıklı Grup Ortalaması	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Merkez	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$\frac{-n_i n_j}{(n_i+n_j)^2}$	0
Medyan	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{-1}{4}$	0

2.3.8.1. Tek Bağlantı Kümeleme Algoritması

Tek bağlantı metodu en basit hiyerarşik kümeleme metotlarından biridir. İlk olarak 1951 yılında Florek tarafından ve daha sonra 1957 yılında McQuitty ve Sneath tarafından tanıtılmıştır. Ayrıca, tek bağlantı metodu en yakın komşu metodu, minimum metodu ve bağlantılılık metodu gibi isimlerleriyle de bilinmektedir [23]. Bu yöntem, farklı veri yapılarındaki kümelenmeleri tanımlayabilmesi bakımından sıklıkla tercih edilmektedir.

İki grup arasındaki mesafeyi ölçmek için en yakın komşu mesafesi kullanılır [23]. C_i, C_j ve C_k veri noktalarının üç kümesi olsun. Bu durumda, C_k ve $C_i \cup C_j$ arasındaki mesafe Lance-Williams formülünden elde edilebilir. $D(.,.)$ iki küme arasındaki mesafe olmak üzere,

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \frac{1}{2}D(C_k, C_i) + \frac{1}{2}D(C_k, C_j) - \frac{1}{2}|D(C_k, C_i) - D(C_k, C_j)| \\ &= \frac{1}{2}\Big[D(C_k, C_i) + D(C_k, C_j) - |D(C_k, C_i) - D(C_k, C_j)|\Big] \end{aligned} \quad (2.63)$$

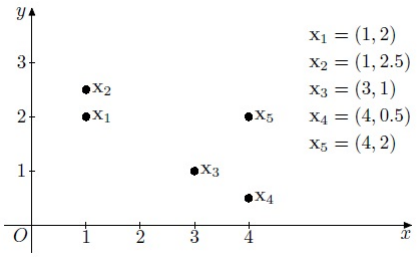
Şimdi kabul edelim ki, $D(C_k, C_i) > D(C_k, C_j)$ olsun. Bu durumda, Denklem (2.63)

$$
\begin{aligned}
D(C_k, C_i \cup C_j) &= \frac{1}{2}D(C_k, C_i) + \frac{1}{2}D(C_k, C_j) - \frac{1}{2}|D(C_k, C_i) - D(C_k, C_j)| \\
&= \frac{1}{2}\Big[D(C_k, C_i) + D(C_k, C_j) - |D(C_k, C_i) - D(C_k, C_j)|\Big] \\
&= \frac{1}{2}[2D(C_k, C_j)] \\
&= D(C_k, C_j) \\
&= \min\{D(C_k, C_i), D(C_k, C_j)\}
\end{aligned}
$$

olarak elde edilir. Benzer şekilde, $D(C_k, C_j) > D(C_k, C_i)$ kabul edilirse, $D(C_k, C_i \cup C_j) = \min\{D(C_k, C_i), D(C_k, C_j)\}$ sonucu elde edilir. Bu algoritmanın aşamaları aşağıdaki gibidir:

1. Verilere göre benzersizlik matrisi oluşturulur.
2. Benzersizlik matrisindeki en küçük uzaklığa sahip olan iki birim ya da gözlem birleştirilir.
3. Bir sonraki uzaklıklar hesaplanıp benzersizlik matrisi oluşturulur.
4. Bu süreçte tüm birim ya da gözlemler tek bir küme oluşturana kadar bu süreç tekrarlanarak devam edilir.

Örneğin, Şekil 2.13'de verilen veri noktaları için, öklidyen mesafe kullanılarak hesaplanan benzersizlik matrisi Tablo 2.17 ile tanımlanmıştır. Öklidyen mesafe kullanılarak tek bağlantılı hiyerarşik kümeleme algoritması bu verilere uygulanırsa, bu durumda x_1 ve x_2 arasındaki mesafe diğer mesafelere göre en küçük olduğundan dolayı, x_1 ve x_2 bu algoritmanın ilk aşamasında birleştirilerek bir küme haline getirilir.



Şekil 2.13: Beş nokta ile iki boyutlu veri kümesi.

Tablo 2.17: Şekil 2.13 da verilen veri kümelerinin benzersizlik mesafesi.

	x_1	x_2	x_3	x_4	x_5
x_1	0	0,5	2,24	3,35	3
x_2	0,5	0	2,5	3,61	3,04
x_3	2,24	2,5	0	1,12	1,41
x_4	3,35	3,61	1,12	0	1,5
x_5	3	3,04	1,41	1,5	0

Şimdi, $\{x_1, x_2\}$, ve x_3, x_4, x_5 arasındaki mesafe Denklem (2.63)'den aşağıdaki gibi elde edilir.

$$D(\{x_1, x_2)\}, x_3) = \min\{d(x_1, x_3), d(x_2, x_3)\} = 2,24$$

$$D(\{x_1, x_2)\}, x_4) = \min\{d(x_1, x_4), d(x_2, x_4)\} = 3,35$$

$$D(\{x_1, x_2)\}, x_5) = \min\{d(x_1, x_5), d(x_2, x_5)\} = 3$$

x_1 ve x_2 birleştirildikten sonra, benzersizlik matrisi aşağıdaki gibi olur.

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0	2,24	3,35	3
x_3	2,24	0	1,12	1,41
x_4	3,35	1,12	0	1,5
x_5	3	1,41	1,5	0

Algoritmanın ikinci aşamasında, x_3 ve x_4 arasındaki mesafe en küçük olduğu için $\{x_1, x_2\}$ kümesinden ayrı olarak birleştirilip bir $\{x_3, x_4\}$ kümesi oluşturulur. Daha sonra, $\{x_3, x_4\}$ ve diğer kalan gruplar arasındaki mesafe aşağıdaki gibi hesaplanır.

$$\begin{aligned} D(\{x_3, x_4\}, \{x_1, x_2\}) &= \min\{d(x_1, x_3), d(x_1, x_4), d(x_2, x_3), d(x_2, x_4)\} \\ &= \min\{D(\{x_1, x_2\}, x_3), D(\{x_1, x_2\}, x_4)\} \\ &= 2,24 \end{aligned}$$

ve

$$\begin{aligned} D(\{x_3, x_4\}, x_5) &= \min\{d(x_3, x_5), d(x_4, x_5)\} \\ &= 1,41 \end{aligned}$$

x_3 ve x_4 birleştirildikten sonra, benzersizlik matrisi aşağıdaki gibi elde edilir.

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0	2,24	3
$\{x_3, x_4\}$	2,24	0	1,41
x_5	3	1,41	0

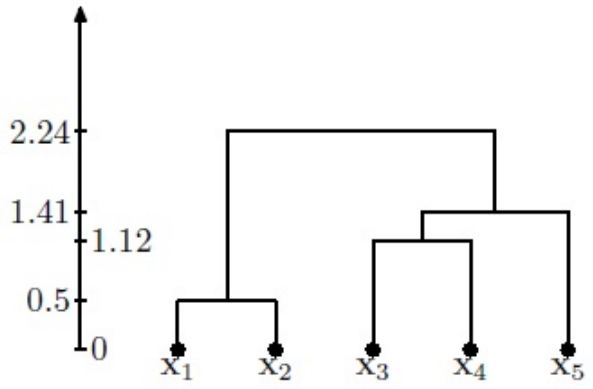
Algoritmanın üçüncü aşamasında, $\{x_3, x_4\}$ ve x_5 birleştirilir ve daha sonrasında $\{x_1, x_2\}$ ile $\{x_3, x_4, x_5\}$ arasındaki mesafe hesaplanır.

$$\begin{aligned} D(\{x_1, x_2\}, \{x_3, x_4, x_5\}) &= \min\{d(x_1, x_3), d(x_1, x_4), d(x_1, x_5), d(x_2, x_3), d(x_2, x_4), d(x_2, x_5)\} \\ &= \min\{D(\{x_1, x_2\}, \{x_3, x_4\}), D(\{x_1, x_2\}, x_5))\} \\ &= 2,24 \end{aligned}$$

Benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	$\{x_3, x_4, x_5\}$
$\{x_1, x_2\}$	0	2,24
$\{x_3, x_4, x_5\}$	2,24	0

Algoritmanın dördüncü aşamasında, bütün noktalar tek bir kümede birleştirilir. Bu kümelemenin dendrogramı Şekil 2.14'de gösteriliyor.



Şekil 2.14: İki boyutlu beş elemanlı veri kümesine tek bağlantı metodu uygulanarak üretilen dendrogram.

2.3.8.2. Tam Bağlantı Kümeleme Algoritması

Tek bağlantı metodunun aksine, tam bağlantı metodu iki küme arasındaki benzersizliği ölçmek için en uzak komşu mesafesini kullanır [23]. C_i, C_j ve C_k veri noktalarının üç grubu olsun. Bu durumda, C_k ve $C_i \cup C_j$ kümeleri arasındaki mesafe Lance-Williams formülünden aşağıdaki gibi elde edilir:

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \frac{1}{2}D(C_k, C_i) + \frac{1}{2}D(C_k, C_j) + \frac{1}{2}|D(C_k, C_i) - D(C_k, C_j)| \\ &= \frac{1}{2}\Big[D(C_k, C_i) + D(C_k, C_j) + |D(C_k, C_i) - D(C_k, C_j)|\Big] \end{aligned} \quad (2.64)$$

Şimdi kabul edelim ki, $D(C_k, C_i) > D(C_k, C_j)$ olsun. Bu durumda, Denklem (2.64)

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \frac{1}{2}D(C_k, C_i) + \frac{1}{2}D(C_k, C_j) + \frac{1}{2}|D(C_k, C_i) - D(C_k, C_j)| \\ &= \frac{1}{2}\Big[D(C_k, C_i) + D(C_k, C_j) + |D(C_k, C_i) - D(C_k, C_j)|\Big] \\ &= \frac{1}{2}[2D(C_k, C_i)] \\ &= D(C_k, C_i) \\ &= \max\{D(C_k, C_i), D(C_k, C_j)\} \end{aligned}$$

olarak elde edilir. Benzer şekilde, $D(C_k, C_j) > D(C_k, C_i)$ kabul edilirse, $D(C_k, C_i \cup C_j) = \max\{D(C_k, C_i), D(C_k, C_j)\}$ sonucu elde edilir. Tablo 2.17'de verilen benzersizlik matrisine tam bağlantı metodu uygulanarak birinci aşamada x_1 ve x_2 birleştirilerek bir küme oluşturulur. Böylece, $\{x_1, x_2\}$ ve kalan üç nokta arasındaki mesafe

$$D(\{x_1, x_2\}, x_3) = \max\{d(x_1, x_3), d(x_2, x_3)\} = 2,5$$

$$D(\{x_1, x_2\}, x_4) = \max\{d(x_1, x_4), d(x_2, x_4)\} = 3,61$$

$$D(\{x_1, x_2\}, x_5) = \max\{d(x_1, x_5), d(x_2, x_5)\} = 3,04$$

olarak hesaplanır. Algoritmanın birinci aşamasında x_1 ve x_2 birleştirildikten sonra, benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0	2,5	3,61	3,04
x_3	2,5	0	1,12	1,41
x_4	3,61	1,12	0	1,5
x_5	3,04	1,41	1,5	0

Daha sonra, x_3 ve x_4 arasındaki mesafe en az olduğundan dolayı algoritmanın ikinci aşamasında x_3 ve x_4 birleştirilir. x_3 ve x_4 birleştirildikten sonra, $\{x_3, x_4\}$ ve kalan gruplar arasındaki mesafeler

$$\begin{aligned} D(\{x_3, x_4\}, \{x_1, x_2\}) &= \max\{d(x_1, x_3), d(x_1, x_4), d(x_2, x_3), d(x_2, x_4)\} \\ &= \max\{D(\{x_1, x_2\}, x_3), D(\{x_1, x_2\}, x_4)\} \\ &= 3,61 \end{aligned}$$

ve

$$\begin{aligned} D(\{x_3, x_4\}, x_5\}) &= \max\{d(x_3, x_5), d(x_4, x_5)\} \\ &= 1,5 \end{aligned}$$

olarak hesaplanır. x_3 ve x_4 birleştirildikten sonra, benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0	3,61	3,04
$\{x_3, x_4\}$	3,61	0	1,5
x_5	3,04	1,5	0

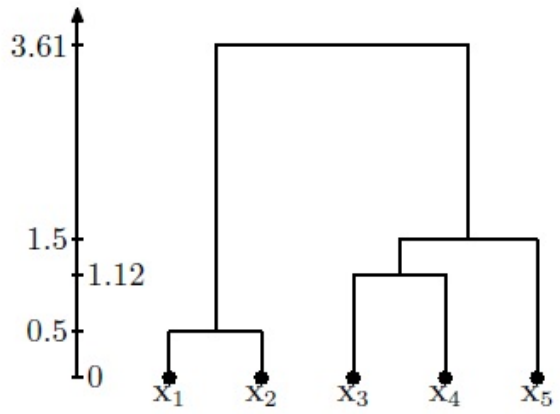
Algoritmanın üçüncü aşamasında, $\{x_3, x_4\}$ ve x_5 arasındaki mesafe en az olduğundan dolayı $\{x_3, x_4\}$ ve x_5 grupları birleştirilir. $\{x_3, x_4, x_5\}$ elde edildikten sonra, $\{x_1, x_2\}$ grubu ile arasındaki mesafe

$$\begin{aligned} D(\{x_1, x_2\}, \{x_3, x_4, x_5\}) &= \max\{d_{13}, d_{14}, d_{15}, d_{23}, d_{24}, d_{25}\} \\ &= \max\{D(\{x_1, x_2\}, \{x_3, x_4\}), D(\{x_1, x_2\}, x_5)\} \\ &= 3,61 \end{aligned}$$

olarak hesaplanır. Burada, $i = 1, 2$ ve $j = 3, 4, 5$ olmak üzere $d_{ij} = d(x_i, x_j)$'dır ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	$\{x_3, x_4, x_5\}$
$\{x_1, x_2\}$	0	3,61
$\{x_3, x_4, x_5\}$	3,61	0

Şekil 2.13'deki verilere öklidyen mesafe uygulanarak elde edilen benzersizlik matrisine tam bağlantı metodu uygulanıp ve daha sonra elde edilen bu süreçleri özetleyen ve açıklayan dendrogram Şekil 2.15'de ifade edilmektedir.



Şekil 2.15: İki boyutlu beş elemanlı veri kümesine tam bağlantı metodu uygulanarak üretilen dendrogram.

2.3.8.3. Grup Ortalamalı Kümeleme Algoritması

Grup ortalamalı kümeleme algoritması ayrıca "UPGMA" olarak temsil edilen *aritmetik ortalama kullanılarak ağırlıksız grup ortalaması metodu* olarak da bilinmektedir [23]. Bu algoritma Sokal ve Michener tarafından önerilmiştir. Grup ortalamalı metotta, iki grup arasındaki mesafe her gruptan alınan veri noktalarının bütün olası çiftleri arasındaki mesafenin ortalaması olarak tanımlanır [23]. Bu algoritma süresince elde edilen sonuçların bir özeti niteliği taşıyan dendrogram genellikle tam bağlantı algoritmasında elde edilen dendrogramla benzerlik göstermektedir. Ancak her bir yöntemde farklı benzerlik ya da uzaklık ölçüsü kullanılmasından kaynaklı, birleşme sonucundaki değerler farklı olabilmektedir. Şimdi, C_i, C_j ve C_k veri noktalarının üç grubu olsun. Bu durumda, C_k ve $C_i \cup C_j$ arasındaki mesafe Lance-Williams formülünden aşağıdaki gibi elde edilir.

$$D(C_k, C_i \cup C_k) = \frac{|C_i|}{|C_i| + |C_j|} D(C_k, C_i) + \frac{|C_j|}{|C_i| + |C_j|} D(C_k, C_j) \tag{2.65}$$

Burada, $D(.,.)$ iki küme arasındaki mesafedir. C_1, C_2 ve C_3 boştan farklı, karşılıklı örtüşmeyen kümeler olsun. $n_i = |C_i|$, $n_j = |C_j|$ ve $\sum(C_i, C_j)$ C_i ve C_j kümeleri arasındaki mesafelerin toplamı olmak üzere,

$$D(C_i, C_j) = \frac{1}{n_i n_j} \sum(C_i, C_j) \;\; ; 1 \leq i < j \leq 3 \tag{2.66}$$

denklemi ortalama komşu mesafesinden sağlanır [23]. Ayrıca, $\sum(C_1, C_2) + \sum(C_1, C_3) = \sum(C_1, C_2 \cup C_3)$ olduğundan dolayı,

$$\begin{aligned} D(C_1, C_2 \cup C_3) &= \frac{n_2}{n_2 + n_3} D(C_1, C_2) + \frac{n_3}{n_2 + n_3} D(C_1, C_3) \\ &= \frac{n_2}{n_2 + n_3} \cdot \frac{1}{n_1 \cdot n_2} \sum(C_1, C_2) + \frac{n_3}{n_2 + n_3} \cdot \frac{1}{n_2 \cdot n_3} \sum(C_1, C_3) \\ &= \frac{1}{n_1(n_2 + n_3)} \sum(C_1, C_2 \cup C_3) \end{aligned} \tag{2.67}$$

elde edilir.

Şekil 2.13'da verilen veri noktalarına öklidyen mesafe uygulandığında, elde edilen benzersizlik matrisine ortalama bağlantılı kümeleme algoritması uygulansın. Bu durumda, algoritmanın birinci aşamasında x_1 ve x_2 arasındaki mesafe diğer mesafelere göre en az olduğundan dolayı, x_1 ve x_2 birleştirilerek bir küme oluşturulur. x_1 ve x_2 birleştikten sonra, $\{x_1, x_2\}$ ve kalan üç adet veri noktaları arasındaki mesafeler,

$$D(\{x_1, x_2\}, x_3) = \frac{1}{2} d(x_1, x_3) + \frac{1}{2} d(x_2, x_3) = 2,37$$

$$D(\{x_1, x_2\}, x_4) = \frac{1}{2} d(x_1, x_4) + \frac{1}{2} d(x_2, x_4) = 3,48$$

$$D(\{x_1, x_2\}, x_5) = \frac{1}{2} d(x_1, x_5) + \frac{1}{2} d(x_2, x_5) = 3,02$$

olarak hesaplanır ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0	2,37	3,48	3,02
x_3	2,37	0	1,12	1,41
x_4	3,48	1,12	0	1,5
x_5	3,02	1,41	1,5	0

Algoritmanın ikinci aşamasında yukarıdaki benzersizlik matrisi göz önünde bulundurularak x_3 ve x_4 arasındaki mesafe en küçük olduğundan dolayı bu noktalar birleştirilir. $\{x_3, x_4\}$ ve diğer kümeler arasındaki mesafe aşağıdaki gibi olur:

$$D(\{x_1, x_2\}, \{x_3, x_4\}) = \frac{1}{2}D(\{x_1, x_2\}, x_4) + \frac{1}{2}D(\{x_1, x_2\}, x_3) = 2,93$$

$$D(\{x_3, x_4\}, x_5) = \frac{1}{2}d(x_3, x_5) + \frac{1}{2}d(x_4, x_5) = 1,46$$

x_3 ve x_4 birleştirildikten sonra, benzersizlik matrisi

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0	2,93	3,02
$\{x_3, x_4\}$	2,93	0	1,46
x_5	3,02	1,46	0

şeklinde olur. Algoritmanın üçüncü aşamasında, $\{x_3, x_4\}$ ve x_5 arasındaki mesafe en az olduğundan, $\{x_3, x_4\}$ ve x_5 birleştirilir. Bu durumda, $\{x_1, x_2\}$ ve $\{x_3, x_4, x_5\}$ arasındaki mesafe

$$\begin{aligned} D(\{x_1, x_2\}, \{x_3, x_4, x_5\}) &= \frac{2}{3}D(\{x_1, x_2\}, \{x_3, x_4\}) + \frac{1}{3}D(\{x_1, x_2\}, x_5) \\ &= 2,96 \end{aligned}$$

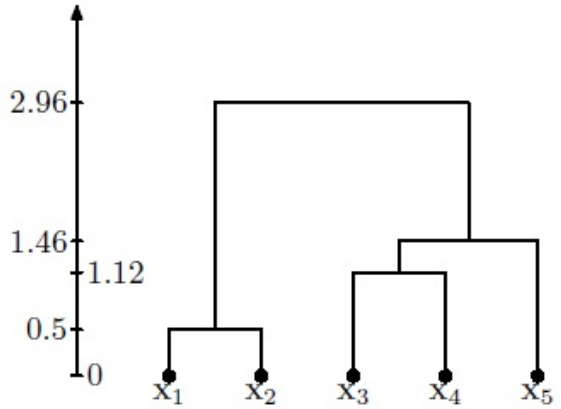
olarak bulunur. Böylece, benzersizlik matrisi

	$\{x_1, x_2\}$	$\{x_3, x_4, x_5\}$
$\{x_1, x_2\}$	0	2,96
$\{x_3, x_4, x_5\}$	2,96	0

şeklinde elde edilir. Ayrıca, mesafeler Denklem (2.67) tarafından hesaplanmaktadır. Örneğin, son aşamada $\{x_1, x_2\}$ ve $\{x_3, x_4, x_5\}$ arasındaki mesafe

$$D(\{x_1, x_2\}, \{x_3, x_4, x_5\}) = \frac{1}{6}(d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}) = 2,96$$

olarak hesaplanır. Bu kümelemenin dendrogramı Şekil 2.16'de verilmektedir.



Şekil 2.16: Beş adet ikili veri kümesine grup ortalaması metodu uygulanarak üretilen dendrogram.

2.3.8.4. Ağırlıklı Grup Ortalaması Kümeleme Algoritması

Lance-Williams formülü kulllanılarak, iki küme arasındaki mesafe

$$D(C_k, C_i \cup C_j) = \frac{1}{2}\Big(D(C_k, C_i) + D(C_k, C_j)\Big) \tag{2.68}$$

olarak hesaplanır [23]. Burada, C_i, C_j ve C_k veri noktalarının kümesidir. Şekil 2.13'de verilen beş veri noktasına ağırlıklı grup ortalaması metodu uygulanırsa, ilk aşamada diğer metotlarda olduğu gibi x_1 ve x_2 birleştirilir. x_1 ve x_2 birleştirildikten sonra, kümeler arasındaki mesafe

$$D(\{x_1, x_2\}, x_3) = \frac{1}{2}(d(x_1, x_3) + d(x_2, x_3)) = 2.37$$

$$D(\{x_1, x_2\}, x_4) = \frac{1}{2}(d(x_1, x_4) + d(x_2, x_4)) = 3.48$$

$$D(\{x_1, x_2\}, x_5) = \frac{1}{2}(d(x_1, x_5) + d(x_2, x_5)) = 3.02$$

olarak bulunur ve böylelikle uzaklık matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0	2,37	3,48	3,02
x_3	2,37	0	1,12	1,41
x_4	3,48	1,12	0	1,5
x_5	3,02	1,41	1,5	0

Bu metodun ikinci aşamasında, x_3 ve x_4 birleştirilir. x_3 ve x_4 birleştirildikten sonra, kümeler arası mesafe

$$D(\{x_3, x_4\}, \{x_1, x_2\}) = \frac{1}{2}(2.37 + 3.48) = 2.93$$

$$D(\{x_3, x_4\}, x_5) = \frac{1}{2}(1.41 + 1.5) = 1.46$$

olarak bulunur ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0	2,93	3,02
$\{x_3, x_4\}$	2,93	0	1,46
x_5	3,02	1,46	0

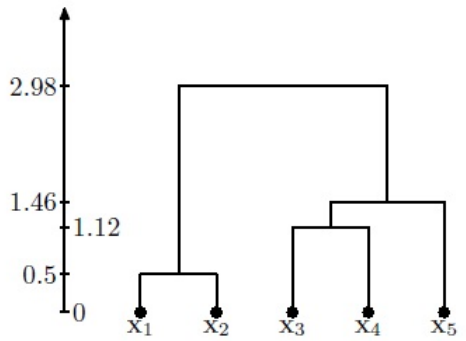
$\{x_3, x_4\}$ ve x_5 kümeleri bu metodun üçüncü aşamasında birleştirilir. Bu durumda, mesafe aşağıdaki gibi olur:

$$D(\{x_1, x_2\}, \{x_3, x_4, x_5\}) = \frac{1}{2}(2.93 + 3.02) = 2.98$$

Böylelikle, bu durumda benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	$\{x_3, x_4, x_5\}$
$\{x_1, x_2\}$	0	2,98
$\{x_3, x_4, x_5\}$	2,98	0

Bu kümelemenin bütün süreçleri Şekil 2.17'de gösterilen dendrogram tarafından temsil edilmektedir.



Şekil 2.17: Beş adet veri kümelerine ağırlıklı grup ortalaması metodu uygulanarak üretilen dendrogram.

2.3.8.5. Merkezi Kümeleme Algoritması

Gürültü ve aykırı değerlerden en az etkilenen hiyerarşik kümeleme algoritmasıdır. Bu algoritmada genellikle kare öklidyen mesafe kullanılmaktadır. Merkezi metotta kümeler arasındaki mesafe Denklem (2.62)'deki Lance-Williams formülününden ve Tablo 2.16'den faydalanarak aşağıdaki formda hesaplanabilir [23].

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \frac{|C_i|}{|C_i| + |C_j|} D(C_k, C_i) + \frac{|C_j|}{|C_i| + |C_j|} D(C_k, C_j) \\ &- \frac{|C_i| \cdot |C_j|}{(|C_i| + |C_j|)^2} D(C_i, C_j) \end{aligned} \tag{2.69}$$

Burada, C_i, C_j ve C_k verilerin birer grubudur. C ve C' herhangi iki örtüşmeyen kümeler olmak üzere, bu kümeler arasındaki mesafe Denklem (2.69)'den

$$\begin{aligned} D(C, C') &= \frac{1}{|C| \cdot |C'|} \sum_{x \in C,\, y \in C'} d(x, y) - \frac{1}{2|C|^2} \sum_{x, y \in C} d(x, y) \\ &- \frac{1}{2|C'|^2} \sum_{x, y \in C'} d(x, y) \end{aligned} \tag{2.70}$$

Burada, $d(.,.)$ benzersizlik matrisi tarafından hesaplanan mesafe fonksiyonudur. Aslında, C_1, C_2 ve C_3 boştan farklı ve karşılıklı örtüşmeyen kümeler olsun. Bu durumda, $i.$ ve $j.$ kümeler arasındaki uzaklık

$$\begin{aligned} D(C_i, C_j) &= \frac{1}{n_i \cdot n_j} \sum (C_i, C_j) - \frac{1}{2n_i^2} \sum (C_i) \\ &- \frac{1}{2n_j^2} \sum (C_j) \;\; ; 1 \leq i < j \leq 3 \end{aligned} \tag{2.71}$$

olarak tanımlanır [23]. Burada, $n_i = |C_i|$, $n_j = |C_j|$ ve $\sum(C_i, C_j)$ ise C_i ve C_j kümeleri arasındaki mesafelerin toplamını ifade eder ve

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

olarak hesaplanır. C_i arasındaki mesafelerin toplamı

$$\sum(C_i) = \sum_{x,y \in C_i} d(x,y)$$

olmak üzere,

$$\begin{aligned} D(C_1, C_2 \cup C_3) &= \frac{1}{n_1(n_2+n_3)} \sum(C_1, C_2 \cup C_3) - \frac{1}{2n_1^2} \sum(C_1) \\ &- \frac{1}{2(n_2+n_3)^2} \sum(C_2 \cup C_3) \end{aligned} \tag{2.72}$$

yazılır. Bu eşitliği gösterelim. Bu durumda, öncelikle Denklem (2.69) kullanılarak

$$\begin{aligned} D(C_1, C_2 \cup C_3) &= \frac{n_2}{n_2+n_3} D(C_1, C_2) + \frac{n_3}{n_2+n_3} D(C_1, C_3) \\ &- \frac{n_2 n_3}{(n_2+n_3)^2} D(C_2, C_3) \end{aligned}$$

yazılır. Yukarıdaki denkleme Denklem (2.71) uygulanır ve bazı cebirsel işlemler yapılırsa,

$$\begin{aligned} D(C_1, C_2 \cup C_3) &= \frac{n_2}{n_2+n_3} \left(\frac{1}{n_1 \cdot n_2} \sum(C_1, C_2) - \frac{1}{2n_1^2} \sum(C_1) - \frac{1}{2n_2^2} \sum(C_2) \right) \\ &+ \frac{n_3}{n_2+n_3} \left(\frac{1}{n_1 \cdot n_3} \sum(C_1, C_3) - \frac{1}{2n_1^2} \sum(C_1) - \frac{1}{2n_3^2} \sum(C_3) \right) \\ &- \frac{n_2 \cdot n_3}{(n_2+n_3)^2} \left(\frac{1}{n_2 \cdot n_3} \sum(C_2, C_3) - \frac{1}{2n_2^2} \sum(C_2) - \frac{1}{2n_3^2} \sum(C_3) \right) \\ &= \sum(C_1, C_2) \Big[\frac{1}{n_1(n_2+n_3)} \Big] + \sum(C_1, C_3) \Big[\frac{1}{n_1(n_2+n_3)} \Big] \\ &+ \sum(C_2, C_3) \Big[-\frac{1}{(n_2+n_3)^2} \Big] + \sum(C_1) \Big[-\frac{n_2}{2n_1^2(n_2+n_3)} - \frac{n_3}{2n_1^2(n_2+n_3)} \Big] \\ &+ \sum(C_2) \Big[-\frac{1}{2n_2(n_2+n_3)} + \frac{n_3}{2n_2(n_2+n_3)^2} \Big] \\ &+ \sum(C_3) \Big[-\frac{1}{2n_3(n_2+n_3)} + \frac{n_2}{2n_3(n_2+n_3)^2} \Big] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n_1(n_2+n_3)}\sum(C_1, C_2 \cup C_3) - \frac{1}{2n_1^2}\sum(C_1) \\
&- \frac{1}{2(n_2+n_3)^2}\left(\sum(C_2) + \sum(C_3) + 2\sum(C_2, C_3)\right) \qquad (2.73)
\end{aligned}
$$

elde edilir. Burada

$$\sum(C_1, C_2) + \sum(C_1, C_3) = \sum(C_1, C_2 \cup C_3)$$

ve

$$\sum(C_2) + \sum(C_3) + 2\sum(C_2, C_3) = \sum(C_2 \cup C_3)$$

denklemleri kullanılırsa,

$$
\begin{aligned}
D(C_1, C_2 \cup C_3) &= \frac{1}{n_1(n_2+n_3)}\sum(C_1, C_2 \cup C_3) - \frac{1}{2n_1^2}\sum(C1) \\
&- \frac{1}{2(n_2+n_3)^2}\sum(C_2 \cup C_3)
\end{aligned}
$$

bulunarak ispat tamamlanır. Şekil 2.13'de verilen veri kümesine merkezi kümeleme algoritması uygulanırsa, ilk aşamada diğer metotlarda olduğu gibi x_1 ve x_2 birleştirilir. x_1 ve x_2 birleştirildikten sonra, mesafeler

$$D(\{x_1, x_2\}, x_3) = \frac{1}{2}\left(d(x_1, x_3) + d(x_2, x_3)\right) - \frac{1}{4}d(x_1, x_2) = 2.245$$

$$D(\{x_1, x_2\}, x_4) = \frac{1}{2}\left(d(x_1, x_4) + d(x_2, x_4)\right) - \frac{1}{4}d(x_1, x_2) = 3.355$$

$$D(\{x_1, x_2\}, x_5) = \frac{1}{2}\left(d(x_1, x_5) + d(x_2, x_5)\right) - \frac{1}{4}d(x_1, x_2) = 2.895$$

olarak bulunur ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0	2.245	3.355	2.895
x_3	2.245	0	1,12	1,41
x_4	3.355	1,12	0	1,5
x_5	2.895	1,41	1,5	0

İkinci aşamada, x_3 ve x_4 arasındaki mesafe en az olduğundan dolayı x_3 ve x_4 birleştirilir ve mesafeler

$$D(\{x_3, x_4\}, \{x_1, x_2\}) = \frac{1}{2}(2.245 + 3.355) - \frac{1}{4}(1.12) = 2.52$$

$$D(\{x_3, x_4\}, x_5) = \frac{1}{2}(1.41 + 1.5) - \frac{1}{4}(1.12) = 1.175$$

olarak bulunur ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0	2.52	2.895
$\{x_3, x_4\}$	2.52	0	1.175
x_5	2.895	1.175	0

Üçüncü aşamada, $\{x_3, x_4\}$ ve x_5 birleştirilir ve mesafe

$$D(\{x_1, x_2\}, \{x_3, x_4, x_5\}) = \frac{2}{3}(2.52) + \frac{1}{3}(2.895) - \frac{2}{9}(1.175)$$

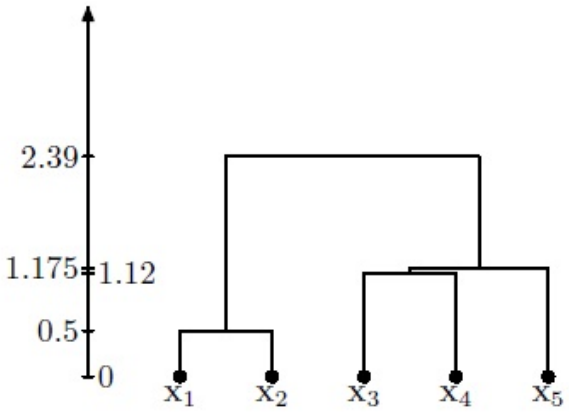
olarak hesaplanır ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	$\{x_3, x_4, x_5\}$
$\{x_1, x_2\}$	0	2.39
$\{x_3, x_4, x_5\}$	2.39	0

Mesafeler Denklem (2.70) tarafından hesaplanır. Örneğin, son aşamada $\{x_1, x_2\}$ ve $\{x_3, x_4, x_5\}$ arasındaki mesafe

$$\begin{aligned} D(\{x_1, x_2\}, \{x_3, x_4, x_5\}) &= \frac{1}{6}(d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}) - \frac{1}{8}(2d_{12}) \\ &- \frac{1}{18}(2d_{34} + 2d_{35} + 2d_{45}) \\ &= \frac{1}{6}(2.24 + 3.35 + 3 + 2.5 + 3.61 + 3.04) - \frac{1}{4}(0.5) \\ &- \frac{1}{9}(1.12 + 1.41 + 1.5) \\ &= 2.957 - 0.125 - 0.448 \cong 2.39 \end{aligned}$$

olarak hesaplanır. Bu kümelemenin tüm süreçleri Şekil 2.18'de gösterilen dendrogram tarafından temsil edilmektedir.



Şekil 2.18: Şekil 2.13'de verilen veri kümelerine merkezi metot uygulanarak üretilen dendrogram.

2.3.8.6. Medyan Kümeleme Algoritması

Merkezi kümeleme algoritması metodunda birleştirilecek olan iki grubun boyutları arasında oldukça büyük fark varsa, bu iki grubun birleşmesi sonucunda oluşacak olan yeni kümenin merkezi, boyutu büyük olan kümeye daha yakın olacaktır [47]. Hatta bu merkez büyük boyutlu küme içerisinde bile yer alabilir. Merkezi metotta bu tür dezavantajları gidermek amacıyla 1967 yılında Gower tarafından medyan kümeleme algoritması geliştirilmiştir. Medyan kümeleme algoritmasıyla oluşturulan yeni grubun merkezi grupların boyutundan bağımsızdır. Ancak, her algoritmada olduğu gibi bu algoritmanın da dezavantajı bulunmaktadır. Bu dezavantaj medyan metodun geometriksel olarak yorumlanamamasıdır. Bundan dolayı da, kolerasyon katsayıları gibi ölçüler bu yöntem için uygun değildir [23]. Medyan kümeleme algoritmasında yeni oluşturulan gruplar ve diğer gruplar arasındaki mesafe

$$D(C_k, C_i \cup C_j) = \frac{1}{2}D(C_k, C_i) + \frac{1}{2}D(C_k, C_j) - \frac{1}{4}D(C_i, C_j) \tag{2.74}$$

olarak hesaplanmaktadır. Burada, C_i, C_j ve C_k veriler için üç kümedir. Şekil 2.13'de verilen veri kümesi ele alındığında, medyan kümeleme algoritmasının ilk aşamasında diğer metotlarda olduğu gibi x_1 ve x_2 birleştirilir. x_1 ve x_2 birleştirildikten sonra, mesafeler

$$D(\{x_1, x_2\}, x_3) = \frac{1}{2}\left(d(x_1, x_3) + d(x_2, x_3)\right) - \frac{1}{4}d(x_1, x_2) = 2.245$$

$$D(\{x_1, x_2\}, x_4) = \frac{1}{2}\left(d(x_1, x_4) + d(x_2, x_4)\right) - \frac{1}{4}d(x_1, x_2) = 3.355$$

$$D(\{x_1, x_2\}, x_5) = \frac{1}{2}\left(d(x_1, x_5) + d(x_2, x_5)\right) - \frac{1}{4}d(x_1, x_2) = 2.895$$

olarak bulunur ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0	2.245	3.355	2.895
x_3	2.245	0	1,12	1,41
x_4	3.355	1,12	0	1,5
x_5	2.895	1,41	1,5	0

Bu metodun ikinci aşamasında x_3 ve x_4 değişkenleri arasındaki mesafe en az olduğundan dolayı, bu değişkenler birleştirilir. Daha sonra, kümeler arasındaki mesafeler

$$D(\{x_3, x_4\}, \{x_1, x_2\}) = \frac{1}{2}(2.245 + 3.355) - \frac{1}{4}(1.12) = 2.52$$

$$D(\{x_3, x_4\}, x_5) = \frac{1}{2}(1.41 + 1.5) - \frac{1}{4}(1.12) = 1.175$$

olarak hesaplanır ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0	2.52	2.895
$\{x_3, x_4\}$	2.52	0	1.175
x_5	2.895	1.175	0

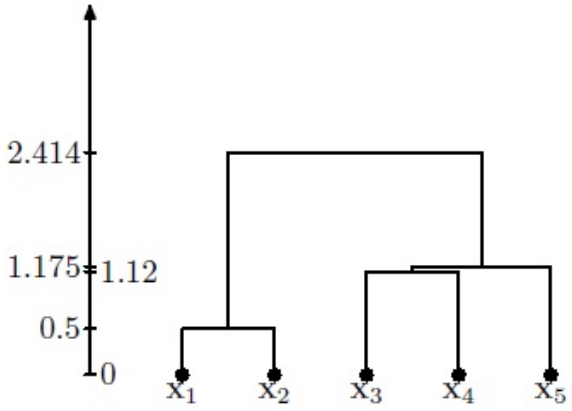
Bu metodun üçüncü aşamasında $\{x_3, x_4\}$ ve x_5 arasındaki mesafe en az olduğundan dolayı, bunlar birleştirilir. Bu durumda mesafe

$$D(\{x_1, x_2\}, \{x_3, x_4, x_5\}) = \frac{1}{2}(2.52 + 2.895) - \frac{1}{4}(1.175) = 2.414$$

olarak bulunur ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	$\{x_3, x_4, x_5\}$
$\{x_1, x_2\}$	0	2.414
$\{x_3, x_4, x_5\}$	2.414	0

Bu kümelemenin bütün süreçleri Şekil 2.19'de gösterilen dendrogram tarafından temsil edilmektedir.



Şekil 2.19: Beş adet ikili veri kümesine medyan metot uygulanarak üretilen dendrogram.

2.3.8.7. Ward'ın Kümeleme Algoritması

Ward Jr. ve Hook 1963 yılında kümelerin birleşimi sonucunda oluşan bilgi kayıplarını minimum seviyeye getirecek olan $P_n, P_{n-1}, ..., P_1$ bölümlemelerini araştıran bir hiyerarşik kümeleme metodu önermişlerdir. Bilgi kaybı genellikle hata kareler toplamı (ESS) tarafından hesaplanmaktadır. Böylece, Ward kümeleme metodu genellikle *minimum varyans metodu* olarakta bilinmektedir [23].

C veri noktaların bir grubu olsun. Bu veri grubunun ortalaması

$$\mu(C) = \frac{1}{|C|}\sum_{x\in C} x$$

olmak üzere C ile ilişkili ESS

$$\text{ESS}(C) = \sum_{x\in C}(x-\mu(C))(x-\mu(C))^T$$

ya da

$$\begin{aligned}\text{ESS}(C) &= \sum_{x\in C} xx^T - \frac{1}{|C|}\left(\sum_{x\in C} x\right)\left(\sum_{x\in C} x\right)^T \\ &= \sum_{x\in C} xx^T - |C|\mu(C)\mu(C)^T \end{aligned} \tag{2.75}$$

olur. Varsayalım kümeleme de $C_1, C_2, ..., C_k$ gibi k-tane grup olsun. Bu durumda, bilgi

kaybı

$$\text{ESS} = \sum_{i=1}^{k} \text{ESS}(C_i) \tag{2.76}$$

ESS'lerin toplamı tarafından temsil edilir. Kümeleme sürecinin ilk adımında her bir gözlem bir küme olduğundan dolayı ESS sıfıra eşit olur ve Ward algoritması bilgi kaybındaki minimum artıştaki birleşme sonuçları olan iki grup birleştirilerek süreç devam eder. Veri tabanına kare öklidyen mesafesi uygulanarak elde edilen benzersizlik matrisi, kümeleme süreci boyunca Lance-Williams formülünden faydalanarak aşağıdaki gibi hesaplanmaktadır.

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \frac{|C_k| + |C_i|}{\sum_{ijk}} D(C_k, C_i) + \frac{|C_k| + |C_j|}{\sum_{ijk}} D(C_k, C_j) \\ &- \frac{|C_k|}{\sum_{ijk}} D(C_i, C_j) \end{aligned} \tag{2.77}$$

Burada, $\sum_{ijk} = |C_i| + |C_j| + |C_k|$'dır. Bunu doğrulamak için, C_i ve C_j birleştirilmek için seçilsin ve sonuç küme ise $C_t = C_i \cup C_j$ tarafından gösterilsin. Bu durumda, ESS'deki artış

$$\begin{aligned} \Delta\text{ESS}_{ij} &= \text{ESS}(C_t) - \text{ESS}(C_i) - \text{ESS}(C_j) \\ &= \left(\sum_{x \in C^T} xx^T - |C_t| \mu_t \mu_t^T \right) - \left(\sum_{x \in C_i} xx^T - |C_i| \mu_i \mu_i^T \right) \\ &- \left(\sum_{x \in C_j} xx^T - |C_j| \mu_j \mu_j^T \right) \\ &= |C_i| \mu_i \mu_i^T + |C_j| \mu_j \mu_j^T - |C_t| \mu_t \mu_t^T \end{aligned} \tag{2.78}$$

dır. Burada, μ_i, μ_j ve μ_t sırasıyla C_i, C_j ve C_t kümelerinin ortalamalarıdır. $|C_t|\mu_t = |C_i|\mu_i + |C_j|\mu_j$'dır ve bu denklemin her iki tarafının karesi alındığında, $\mu_i \mu_i^T + \mu_j \mu_j^T - (\mu_i - \mu_j)(\mu_i - \mu_j)^T = 2\mu_i \mu_j^T$ olduğundan dolayı,

$$|C_t|^2 \mu_t \mu_t^T = |C_i|^2 \mu_i \mu_i^T + |C_j|^2 \mu_j \mu_j^T + 2|C_i||C_j| \mu_i \mu_j^T$$

ya da

$$
\begin{aligned}
|C_t|^2\mu_t\mu_t^T &= |C_i|^2\mu_i\mu_i^T + |C_j|^2\mu_j\mu_j^T + |C_i||C_j|\Big[\mu_i\mu_i^T + \mu_j\mu_j^T \\
&- (\mu_i-\mu_j)(\mu_i-\mu_j)^T\Big] \\
&= |C_i|^2\mu_i\mu_i^T + |C_j|^2\mu_j\mu_j^T + |C_i||C_j|(\mu_i\mu_i^T + \mu_j\mu_j^T) \\
&- |C_i||C_j|(\mu_i-\mu_j)(\mu_i-\mu_j)^T \\
&= |C_i|\Big[|C_i|+|C_j|\Big]\mu_i\mu_i^T + |C_j|\Big[|C_i|+|C_j|\Big]\mu_j\mu_j^T \\
&- |C_i||C_j|(\mu_i-\mu_j)(\mu_i-\mu_j)^T \qquad (2.79)
\end{aligned}
$$

dir. Denklem (2.79)'in her iki tarafı $|C_t|$'ye bölünürse,

$$
\begin{aligned}
|C_t|\mu_t\mu_t^T &= \frac{|C_i|}{|C_i|+|C_j|}\Big[|C_i|+|C_j|\Big]\mu_i\mu_i^T + \frac{|C_j|}{|C_i|+|C_j|}\Big[|C_i|+|C_j|\Big]\mu_j\mu_j^T \\
&- \frac{|C_i||C_j|}{|C_i|+|C_j|}(\mu_i-\mu_j)(\mu_i-\mu_j)^T \\
&= |C_i|\mu_i\mu_i^T + |C_j|\mu_j\mu_j^T \\
&- \frac{|C_i||C_j|}{|C_i|+|C_j|}(\mu_i-\mu_j)(\mu_i-\mu_j)^T
\end{aligned}
$$

elde edilir ve bu eşitlik Denklem (2.78)'deki $|C_t|\mu_t\mu_t^T$'ye uygulanırsa, ESS'deki değişim

$$\Delta\text{ESS}_{ij} = \frac{|C_i||C_j|}{|C_i|+|C_j|}(\mu_i-\mu_j)(\mu_i-\mu_j)^T \qquad (2.80)$$

olarak elde edilir. Şimdi, C_k ve C_t gruplarının potansiyel birleşmelerinden çıkan sonuçlar olan ESS'deki değişimler düşünüldüğünde, Denklem (2.80)'den

$$\Delta\text{ESS}_{kt} = \frac{|C_k||C_t|}{|C_k|+|C_t|}(\mu_k-\mu_t)(\mu_k-\mu_t)^T \qquad (2.81)$$

olur. Burada, $\mu_k = \mu(C_k)$ C_k grubunun ortalamasıdır. Ayrıca, $\mu_t = \frac{1}{|C_t|}(|C_i|\mu_i + |C_j|\mu_j)$ ve $|C_t| = |C_i| + |C_j|$'dır. Eğer kare öklidyen mesafe kullanılarak $D = \{x_1, x_2, ..., x_n\}$ veri kümesi için benzersizlik matrisi hesaplanmak istenirse, bu durumda, benzersizlik matrisinin (i, j) girdileri

$$\begin{aligned} d_{ij}^2 &= d(x_i, x_j) = (x_i - x_j)(x_i - x_j)^T \\ &= \sum_{l=1}^{d} (x_{il} - x_{jl})^2 \end{aligned}$$

dir. Burada, d veri kümesinin boyutudur. Eğer, Denklem (2.80)'de $C_i = \{x_i\}$ ve $C_j = \{x_j\}$ ise, bu durumda x_i ve x_j'nin birleşimi sonucunda ortaya çıkan ESS'deki değişim aşağıdaki gibidir.

$$\Delta \text{ESS}_{ij} = \frac{1}{2} d_{ij}^2 = \frac{1}{2} d(x_i, x_j) \tag{2.82}$$

Ward kümeleme algoritmasının her aşamasında grup içindeki ESS toplamındaki minimum artışı veren iki grubun birleşimini bulmak olduğu için, minimum kare öklidyen mesafe ile iki nokta ilk aşamada birleştirilir. Varsayalım x_i ve x_j minimum kare öklidyen mesafesine sahip olsunlar. Bu durumda, $C_i = \{x_i\}$ ve $C_j = \{x_j\}$ birleştirilir. C_i ve C_j birleştirildikten sonra, $C_i \cup C_j$ ve diğer noktalar arasındaki mesafeler hesaplanır [23].

Kümeleme süreci boyunca Denklem (2.77) kullanılarak benzersizlik matrisi hesaplanırsa, minimum mesafe ile iki grup birleştirilir. Şekil 2.13'de verilen veri kümeleri göz önüne alınırsa, kare öklidyen mesafe tarafından hesaplanan benzersizlik matrisi Tablo 2.18 ile verilmektedir.

Tablo 2.18: Beş noktalı ikili veri kümelerinin kare öklidyen mesafe kullanılarak elde edilen benzersizlik matrisi.

	x_1	x_2	x_3	x_4	x_5
x_1	0	0.25	5	11.25	9
x_2	0.25	0	6.25	13	9.25
x_3	5	6.25	0	1.25	2
x_4	11.25	13	1.25	0	2.25
x_5	9	9.25	2	2.25	0

Başlangıçta, her bir tek nokta bir küme oluşturur ve $\text{ESS}_0 = 0$'dır. Yukarıdaki tartışmaya göre, Ward metodunun ilk aşamasında x_1 ve x_2 birleştirilir ve x_1 ve x_2'nin birleşiminden

çıkan sonuç ESS'deki artış $\Delta\text{ESS}_{12} = \frac{1}{2}(0.25) = 0.125$'dir. Bu yüzden, ESS

$$\text{ESS}_1 = \text{ESS}_0 + \Delta\text{ESS}_{12} = 0.125$$

olur. Denklem (2.77) kullanılarak, mesafeler

$$D(\{x_1, x_2\}, x_3) = \frac{2}{3}\left(d(x_1, x_3) + d(x_2, x_3)\right) - \frac{1}{3}d(x_1, x_2) = 7.42$$

$$D(\{x_1, x_2\}, x_4) = \frac{2}{3}\left(d(x_1, x_4) + d(x_2, x_4)\right) - \frac{1}{3}d(x_1, x_2) = 16.08$$

$$D(\{x_1, x_2\}, x_5) = \frac{2}{3}\left(d(x_1, x_5) + d(x_2, x_5)\right) - \frac{1}{3}d(x_1, x_2) = 12.08$$

olarak hesaplanır ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0	7.42	16.08	12.08
x_3	7.42	0	1.25	2
x_4	16.08	1.25	0	2.25
x_5	12.08	2	2.25	0

Bu metodun ikinci aşamasında x_3 ve x_4 birleştirilir ve ESS'deki artış sonucu $\Delta\text{ESS}_{34} = \frac{1}{2}(1.25) = 0.625$'dır. Toplam ESS

$$\text{ESS}_2 = \text{ESS}_1 + \Delta\text{ESS}_{34} = 0.125 + 0.625 = 0.75$$

olur. x_3 ve x_4 birleştirildikten sonra, mesafeler

$$D(\{x_3, x_4\}, \{x_1, x_2\}) = \frac{3}{4}(7.42 + 16.08) - \frac{2}{4}(1.25) = 17$$

$$D(\{x_3, x_4\}, x_5) = \frac{2}{3}(2 + 2.25) - \frac{1}{3}(1.25) = 2.42$$

olarak hesaplanır ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0	17	12.08
$\{x_3, x_4\}$	17	0	2.42
x_5	12.08	2.42	0

Üçüncü aşamada $\{x_3, x_4\}$ ve x_5 birleştirilir. ESS'deki artış sonucu $\Delta\text{ESS}_{(34)5} = \frac{1}{2}(2.42) =$

1.21'dir. Bu durumda, toplam ESS

$$\text{ESS}_3 = \text{ESS}_2 + \Delta\text{ESS}_{(34)5} = 0.75 + 1.21 = 1.96$$

olur. Mesafe

$$D(\{x_1, x_2\}, \{x_3, x_4, x_5\}) = \frac{4}{5}(17) + \frac{3}{5}(12.08) - \frac{2}{5}(2.42) = 19.88$$

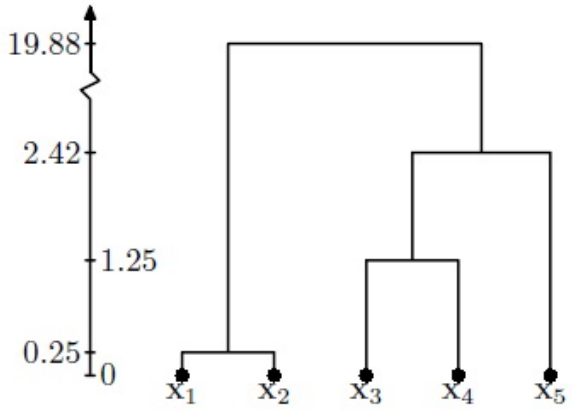
olarak hesaplanır ve benzersizlik matrisi aşağıdaki gibi olur:

	$\{x_1, x_2\}$	$\{x_3, x_4, x_5\}$
$\{x_1, x_2\}$	0	19.88
$\{x_3, x_4, x_5\}$	19.88	0

Tüm veri noktaları tek bir küme oluşturmak için birleştirildiğinde, ESS'deki artış $\Delta\text{ESS}_{(12)(345)} = \frac{1}{2}(19.88) = 9.94$ ve toplam ESS

$$\text{ESS}_4 = \text{ESS}_3 + \Delta\text{ESS}_{(12)(345)} = 1.96 + 9.94 = 11.9$$

olacaktır. Bu kümelemenin bütün süreçleri Şekil 2.20'de gösterilen dendrogram tarafından temsil edilmektedir.



Şekil 2.20: Beş noktalı ikili veri kümelerine Ward metodu uygulanarak üretilen dendrogram.

Ward metodunun sonuçları genellikle optimal değerler vermektedir. Kuiper and Fisher 1975 yılında Monte Carlo metodu kullanarak diğer hiyerarşik kümeleme algoritmalarıyla da Ward metodunun karşılaştırmasını sunmuştur [23].

Hiyerarşik kümeleme algoritmaları pek çok sayıda benzerlik veya uzaklık ölçüsü bulunduğundan dolayı ve bunların seçimlerinin araştırmacının amaçlarına göre değişiklik göstermesinden kaynaklı esnek bir yapıya sahiptir. Bu durum hiyerarşik kümeleme tekniklerinin olumlu yönlerinden sadece biridir. Ayrıca, hiyerarşik olmayan kümeleme tekniklerindeki gibi hiyerarşik kümeleme tekniklerinde kümeleme yapılmadan önce küme sayısının belirlenmesine gerek yoktur. Bu yönüyle de hiyerarşik kümeleme teknikleri avantajlıdır. Ancak, herhangi bir kümeleme tekniğinde olduğu gibi bu hiyerarşik kümeleme tekniklerinde de belli başlı dezavantajlar bulunmaktadır. Örneğin, uygulamada kullanılacak olan veritabanı büyük olsun. Bu durumda, oluşacak olan benzersizlik matrisi de oldukça büyük olacağından dolayı, çoğu istatistiksel bilgisayar yazılımlarından sonuç elde edilemeyebilir. Buna ek olarak, yöntem aykırı değerlerden kolaylıkla etkilenmektedir.

2.3.9. Küme Sayısının Belirlenmesi

Kümeleme analizi üzerinde çalışan araştırmacılar için küme sayısının belirlenmesi önemli bir problemdir. Hiyerarşik kümeleme tekniklerinde, analiz öncesinde küme sayısının belirlenmesi gerekmemektedir. Bunun aksine, hiyerarşik olmayan kümeleme tekniklerinde analiz yapılmadan önce küme sayısının belirlenmesi gerekmektedir [48]. Küme sayısının belirlenmesi için pek çok sayıda yöntem bulunmaktadır. Böyle olmasına rağmen, bu yöntemler kesin bir sonuç ortaya koyamamaktadır. Bundan dolayı da küme sayısının belirlenmesinde araştırmacının bu yöntemler hakkındaki bilgi ve birikiminden ya da uzmanların tecrübelerinden faydalanması genellikle doğru bir yaklaşım olabilir. Küme sayısının belirlenmesinde en çok bilinen ve en pratik yöntem aşağıda verilmektedir.

Küme sayısı k ve gözlem sayısı n olmak üzere, küme sayısı

$$k \cong \sqrt{\frac{n}{2}} \tag{2.83}$$

formülü ile hesaplanmaktadır. Bu formül, boyutu küçük olan örneklemler için kullanılmaktadır. Büyük boyutlu örneklemlerde ise iyi sonuçlar elde edilememektedir. Örneğin; gözlem sayısı 50 ise yaklaşık küme sayısı $k \cong \sqrt{\frac{50}{2}} = 5$'dir.

2.3.9.1. Marriot Yöntemi

Küme sayısının belirlenmesindeki diğer bir yöntem Marriot yöntemidir. Bu yöntem M harfi ile gösterilir ve

$$M = k^2|W| \tag{2.84}$$

olarak tanımlanmıştır. Bu yöntem en küçük M sayısını veren k değeri küme sayısı olarak kabul edilir. Burada, W grup içi kareler toplamı matrisidir ve

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \overline{X_j})(X_{ij} - \overline{X_j})^T \tag{2.85}$$

olarak ifade edilir. Burada, n_j ile j. kümedeki eleman sayısını, k ile küme sayısını, X_{ij} ile j. kümedeki i.birim değerlerini ve $\overline{X_j}$ ile de j. kümenin örneklem ortalama vektörü gösterilmektedir. 1974 yılında Everitt pek çok farklı uygulama ile bu yöntemin diğer yöntemlere göre daha olumlu sonuçlar verdiğini ortaya koymuştur [2].

2.3.9.2. Silhouette Geçerlilik İndeksi

Küme sayısını belirlenmesindeki en önemli yöntemlerden biri de Silhouette indeks yöntemidir. Bu yöntem 1987 yılında Rousseeuw tarafından herhangi bir küme içindeki gözlemlerin o küme içerisinde olup olmamasının uygunluğunu araştırmak için geliştirilmiştir. Bu indekste, $\alpha(x_i)$ ile i. gözlemin bulunduğu küme içerisindeki gözlemlere olan ortalama uzaklığı ve $b(x_i)$ ile de i. gözlemin diğer kümelerdeki bütün gözlemlere olan ortalama uzaklıklarının minimumu gösterilsin. Böylece, i. gözlem için

Silhouette indeksi [2]

$$\text{sil}(x_i) = \frac{b(x_i) - \alpha(x_i)}{\max(\alpha(x_i), b(x_i))} \tag{2.86}$$

olarak tanımlanır ve -1 ve 1 arasında değer alır. Eğer,

$\text{sil}(x_i) \cong 1$ ise i. gözlem doğru sınıflandırılmıştır.

$\text{sil}(x_i) \cong 0$ ise i. gözlem iki küme arasında yer alır.

$\text{sil}(x_i) \cong -1$ ise i. gözlem yanlış sınıflandırılmıştır.

Tüm kümelemenin kalitesi Silhouette değeri ile ölçülmektedir. Doğal ölçü olarakta tüm gözlemler için ortalama Silhouette değeri

$$\text{sil}(C) = \frac{1}{n} \sum_{i=1}^{n} \text{sil}(i) \tag{2.87}$$

olarak hesaplanır. Buna göre, maksimum ortalama Silhouette değerine karşılık gelen küme sayısı uygun küme sayısı olarak alınır [2]. Genel olarak, ortalama Silhouette değeri 0.50'nin üzerinde ise, uygun küme sayısı ve dolayısıyla uygun kümelemeye ulaşıldığı kabul edilir. 6 gözlem ve 4 değişkenli bir örnek göz önüne alındığında, K-ortalamalar tekniği ile k küme sayısı olmak üzere $k = 2, 3$ değerleri için kümeler oluşturulsun. Gözlemlere ilişkin değerler ve kümeleme sonucu oluşmuş öklid mesafesine göre benzersizlik tablosu aşağıdaki gibi olsun.

Tablo 2.19: Gözlemlere ilişkin değerler.

Gözlem	x_1	x_2	x_3	x_4	$k = 2$	$k = 3$
G_1	2	3	1	0	2	2
G_2	3	3	1	0	2	2
G_3	2	2	2	2	1	1
G_4	1	0	5	3	1	3
G_5	1	2	3	4	1	3
G_6	0	1	0	1	2	1

Tablo 2.20: Gözlemlere ilişkin değerler arasındaki öklid uzaklıklar.

Gözlem	G_1	G_2	G_3	G_4	G_5	G_6
G_1	0	1	2,45	5,92	4,7	3,16
G_2	1	0	2,65	6,16	5	3,88
G_3	2,45	2,65	0	3,88	2,45	3,16
G_4	5,92	6,16	3,88	0	3	5,6
G_5	4,7	5	2,45	3	0	4,5
G_6	3,16	3,88	3,16	5,6	4,5	0

$k = 2$ durumunda, gözlemlere ilişkin değerlere K-ortalamalar algoritması uygulandığında, gözlemler $\{G_3, G_4, G_5\}$ ve $\{G_1, G_2, G_6\}$ olarak kümelere ayrılır. Bu durumda, $k = 2$ için G_1 gözlemine ait Silhoutte değeri

$\alpha(G_1) = \frac{1}{2}(1 + 3,16) = 2,08$
$b(G_1) = \frac{1}{3}(2,45 + 5,92 + 4,7) = 4,35$
$\text{sil}(G_1) = \frac{4,35-2,08}{\max(2,08,4,35)} = 0,521$

olarak hesaplanır. Benzer şekilde, G_2, G_3, G_4, G_5, G_6 gözlemleri içinde işlemler yapılırsa,

$\text{sil}(G_2) = \frac{4,603-2,44}{\max(2,44,4,603)} = 0,4699$
$\text{sil}(G_3) = \frac{2,753-3,165}{\max(3,165,2,753)} = -0,13$
$\text{sil}(G_4) = \frac{5,893-3,44}{\max(3,44,5,893)} = 0,416$
$\text{sil}(G_5) = \frac{4,733-2,725}{\max(2,725,4,733)} = 0,424$
$\text{sil}(G_6) = \frac{4,42-3,52}{\max(3,52,4,42)} = 0,203$

Silhouette değerleri bulunur. Bu durumda, $k = 2$ küme sayısı için Silhouette değeri,

$$\text{sil}(k = 2) = \frac{1}{6}(0,521 + 0,4699 - 0,13 + ... + 0,203) = \frac{1}{6}(1,9039) = 0,3173$$

olarak hesaplanır.

$k = 3$ durumunda, gözlemlere ilişkin değerlere K-ortalamalar algoritması uygulandığında, gözlemler $\{G_3, G_6\}$, $\{G_1, G_2\}$ ve $\{G_4, G_5\}$ olarak kümelere ayrılır. Bu

durumda, $k = 3$ için G_1 gözlemine ait Silhouette değeri

$\alpha(G_1) = \frac{1}{1}(1) = 1$
$b(G_1) = \min\{\frac{1}{2}(2,45 + 3,16), \frac{1}{2}(5,92 + 4,7)\} = \min\{2,805, 5,31\} = 2,805$
$\text{sil}(G_1) = \frac{2,805-1}{\max(1,2,805)} = 0,643$

olarak hesaplanır. Benzer şekilde, G_2, G_3, G_4, G_5, G_6 gözlemleri içinde işlemler yapılırsa,

$\text{sil}(G_2) = \frac{3,265-1}{\max(1,3,265)} = 0,693$
$\text{sil}(G_3) = \frac{2,55-3,16}{\max(3,16,2,55)} = -0,193$
$\text{sil}(G_4) = \frac{3,44-3}{\max(3,3,44)} = 0,127$
$\text{sil}(G_5) = \frac{3,475-3}{\max(3,3,475)} = 0,136$
$\text{sil}(G_6) = \frac{3,52-3,16}{\max(3,16,3,52)} = 0,102$

şeklinde Silhouette değerleri bulunur. Bu durumda, $k = 3$ küme sayısı için Silhouette değeri yukarıda elde edilen $\text{sil}(G_i)$ değerlerinin ortalamasıdır.

$$\text{sil}(k = 3) = \frac{1}{6}\sum_{i=1}^{6} \text{sil}(G_i) = \frac{1}{6}(1,508) = 0,251$$

olarak hesaplanır. $\text{sil}(k = 2)$ ve $\text{sil}(k = 3)$ durumlarında elde edilen sonuçlar göz önüne alındığında, $\text{sil}(k = 2)$ değerinin $\text{sil}(k = 3)$ değerinden daha büyük olduğu ve 1 değerine daha yakın olduğu görülmektedir. Bu durumda, gözlemlerin K-ortalamalar algoritmasına göre küme sayısının $k = 2$ olması daha uygun görülür.

2.3.9.3. Calinski-Harabazs Geçerlilik İndeksi

Küme sayısının belirlenmesindeki diğer önemli bir yöntem ise, Calinski-Harabazs indeksidir. Calinski ve Harabazs bir veritabanına ait verilerin hiyerarşik olmayan yöntemlerle keyfi şekilde sınıflandırıldığında en uygun küme sayısını belirlemek amacıyla geliştirdikleri bir yöntemdir. Bu yöntemde, n gözlem sayısı ve k oluşturalacak olan küme sayısı olmak üzere, Calinski-Harabazs indeksi [2]

$$\text{CH}(k) = \frac{\text{BSS}(k)/(k-1)}{\text{WSS}(k)/(n-k)} \tag{2.88}$$

olarak tanımlanır. Burada, $\text{BSS}(k)$ kümeler içi kareler toplamı ve $\text{WSS}(k)$ ise kümeler arası kareler toplamını ifade etmektedir ve sırasıyla

$$\text{BSS}(k) = \frac{1}{2}\sum_{l=1}^{k}\sum_{i,j\in C_l} d(i,j) \tag{2.89}$$

ve

$$\text{WSS}(k) = \frac{1}{2}\sum_{l=1}^{k}\sum_{i\in C_l,\, j\notin C_l} d(i,j) \tag{2.90}$$

olarak tanımlanır. Bu durumda, oluşturulan küme sayılarının uygun olup olmadığını karşılaştırmak istendiğinde, hangi küme sayısının Calinski-Harabazs indeks değeri daha büyükse, o küme sayısının diğer küme sayısına göre uygun olduğu söylenmektedir. Bu durumu daha iyi anlayabilmek adına bir örnek verilecek olursa, aşağıdaki tabloda 7 adet gözlem ve bu gözlemlerle ilişkili 4 adet değişken verilmiştir.

Tablo 2.21: Gözlemlere ilişkin değerler.

Gözlem	x_1	x_2	x_3	x_4	$k=2$	$k=3$
D_1	1	1	1	0	1	1
D_2	2	3	1	1	2	3
D_3	4	2	0	1	2	3
D_4	3	3	2	1	2	3
D_5	0	0	0	0	1	2
D_6	1	0	1	0	1	2
D_7	1	1	1	1	1	1

Bu veritabanına karesel öklidyen mesafe kullanılarak aşağıdaki benzersizlik matrisi elde edilir:

Tablo 2.22: Gözlemlere ilişkin değerler arasındaki karesel öklid uzaklıkları.

Gözlem	D_1	D_2	D_3	D_4	D_5	D_6	D_7
D_1	0	6	12	10	3	1	1
D_2	6	0	6	2	15	11	5
D_3	12	6	0	6	21	15	11
D_4	10	2	6	0	23	15	9
D_5	3	15	21	23	0	2	4
D_6	1	11	15	15	2	0	2
D_7	1	5	11	9	4	2	0

$k = 2$ durumunda, Şekil 2.21'de verilen veritabanına K-ortalamalar algoritması uygulanırsa, gözlemler (D_1, D_5, D_6, D_7) ve (D_2, D_3, D_4) olarak sınıflara ayrılmaktadır. Böylelikle, $k = 2$ olduğunda Calinski-Harabazs indeksini hesaplamak için öncelikle BSS$(k = 2)$ ve WSS$(k = 2)$ hesaplanması gerekir. Bu durumda,

$$\begin{aligned}
\text{WSS}(k=2) &= \frac{1}{2}\Big[\sum_{i=1,5,6,7}\sum_{j=1,5,6,7} d(i,j) + \sum_{i=2,3,4}\sum_{j=2,3,4} d(i,j)\Big] \\
&= \frac{1}{2}\Big[(3+1+1)+(3+2+4)+(1+2+2)+(1+4+2)\Big] \\
&+ \frac{1}{2}\Big[(6+2)+(6+6)+(2+6)\Big] \\
&= \frac{1}{2}\Big[(5+9+5+7)+(8+12+8)\Big] \\
&= \frac{1}{2}\cdot 54 \\
&= 27
\end{aligned}$$

ve

$$\begin{aligned}
\text{BSS}(k=2) &= \frac{1}{2}\Big[\sum_{i=1,5,6,7}\sum_{j=2,3,4} d(i,j)\Big] \\
&= \frac{1}{2}\Big[(6+12+10)+(15+21+23)+(11+15+15)+(5+11+9)\Big] \\
&+ \frac{1}{2}\Big[(6+15+11+5)+(12+21+15+11)+(10+23+15+9)\Big] \\
&= \frac{1}{2}\Big[(28+59+41)+(25+37+59+57)\Big] = 153
\end{aligned}$$

olarak hesaplanır. Böylelikle, $k = 2$ durumu için Calinski-Harabazs indeks değeri

$$\text{CH}(k=2) = \frac{153/(2-1)}{27/(7-2)} = \frac{153}{5,4} = 28,33$$

olarak bulunur. $k = 3$ durumunda ise Şekil 2.21'de verilen veritabanına K-ortalamalar algoritması uygulanırsa, gözlemler (D_1, D_7), (D_5, D_6) ve (D_2, D_3, D_4) olarak sınıflara ayrılmaktadır. Böylelikle, $k = 3$ için Calinski-Harabazs indeksini hesaplamak için öncelikle $\text{BSS}(k=3)$ ve $\text{WSS}(k=3)$ hesaplanması gerekir. Bu durumda,

$$\begin{aligned}
\text{WSS}(k=3) &= \frac{1}{2}\Big[\sum_{i=1,7}\sum_{j=1,7} d(i,j) + \sum_{i=5,6}\sum_{j=5,6} d(i,j) + \sum_{i=2,3,4}\sum_{j=2,3,4} d(i,j)\Big] \\
&= \frac{1}{2}\Big[(1+1)+(2+2)+(6+2)+(6+6)+(2+6)\Big] \\
&= \frac{1}{2}(2+4+8+12+8) \\
&= \frac{1}{2}\cdot 34 \\
&= 17
\end{aligned}$$

ve

$$\begin{aligned}
\text{BSS}(k=3) &= \frac{1}{2}[\sum_{i=1,7}\sum_{i=5,6} d(i,j) + \sum_{i=5,6}\sum_{i=1,7} \\
&+ \sum_{i=1,7}\sum_{i=2,3,4} d(i,j) + \sum_{i=2,3,4}\sum_{i=1,7} d(i,j) \\
&+ \sum_{i=5,6}\sum_{i=2,3,4} d(i,j) + \sum_{i=2,3,4}\sum_{i=5,6} d(i,j)] \\
&= \frac{1}{2}[(3+1)+(4+2)+(3+4)+(1+2) \\
&+ (6+12+10)+(5+11+9)+(6+5)+(12+11) \\
&+ (10+9)+(15+21+23)+(11+15+15) \\
&+ (15+11)+(21+15)+(23+15)] \\
&= \frac{1}{2}[4+6+7+3+28+25+11+23+19+59+41+26+36+38] \\
&= \frac{1}{2}\cdot 326 \\
&= 163
\end{aligned}$$

olarak elde edilir. Böylelikle, $k = 3$ durumu için Calinski-Harabazs indeks değeri

$$\text{CH}(k = 3) = \frac{163/(3-1)}{17/(7-3)} = \frac{81,5}{3,4} = 23,97$$

olarak bulunur. $\text{CH}(k = 2) > \text{CH}(k = 3)$ olduğundan, Calinski-Harabazs indeks değeri kriterine göre küme sayısının $k = 2$ olarak alınması $k = 3$ alınmasına göre daha uygun görülmektedir.

Küme sayısının belirlenmesinde yukarıda bahsedilenlerin dışında araştırmacıların amaçlarına ve tecrübelerine dayalı tercih edebilecekleri Lewis-Thomas yöntemi, Wilk Lamda İstatistiği, Krzanowski-Lai İndeksi gibi yöntemler de bulunmaktadır. Fakat, bu tez çalışmasında bu konulara odaklanmayacağız.

3. MALZEME VE YÖNTEM

Bu bölümde, öncelikle tez çalışmasının uygulama kısmında kullanılan veri tabanına ait değişkenlerin belli başlı özelliklerinden bahsedilmiş, daha sonra da çalışmanın değişik aşamalarını gerçekleştirmek için kullanılan yöntemler açıklanmaya çalışılmıştır.

3.1. MALZEME

Çok zengin sayılabilecek bir biyo çeşitliliğe sahip olan ülkemiz, florasında barındırdığı çok sayıda tür nedeniyle önemli bitki zenginliği olan ülkelerden birisidir. Bu zenginlik birçok familya ve cinste olduğu gibi Lamiaceae familyası ve bu familya içinde 42 taksonla temsil edilen Teucrium cinsi içinde geçerlidir. Türkiye Lamiaceae familyasının önemli bir gen merkezi konumunda olup, bu familyaya ait 45 cins, 546 tür ve diğer alt birimlerle birlikte toplam 731 takson ile temsil edilir. Türkiye'nin en zengin üçüncü familyası konumundadır. Familyanın karakteristik özelliklerinden bazıları; gövde dört köşeli, yapraklar çoğu zaman basit, bazen parçalıdır. Uçucu yağları; sapı tek, başı sekiz hücreli pul şeklindeki Labiatae tipi salgı tüylerindedir. Çiçeklerde kaliks beş loblu, kalıcı bazen bilabiat; korolla bilabiat, üst dudak bazen eksiktir. Stamenler genellikle dört tane olup çoğu zaman didinamdır, bazen de iki stamen bulunur. Ovaryum iki karpelden meydana gelmiş dört gözlü ve üst durumludur, her gözde bir ovül bulunur. Meyve dört nukstan meydana gelen bir şizokarptır [49].

Teucrium L. cinsi çoğunluğu Akdeniz bölgesinde yayılış gösteren ve dünyada yaklaşık 260 civarında türe sahip olan çok büyük, çok farklı şekilleri bulunan ve yeryüzünde oldukça geniş alanlara yayılmış farklı yetişme koşullarında varlığını sürdürebilen bir türdür. Avrupa'nın bitki örtüsünde, Teucrium cinsi 49 çeşit ile yedi bölüme ayrılmaktadır. Türkiye'de ise Teucrium 45 taksonla 34 tür içerir ve bu taksonların 16'sı bulunduğu bölgenin ekolojik şartlarından dolayı yalnızca belirli bölgede yetişebilen bitki türleridir. Türkiye bitki örtüsünde türler 13 taksonla Teucrium Benth., 13 taksonla Chamaedrys (Mill.) Schreb., 2 taksonla Polium (Mill.) Schreb, 10 taksonla Isotriodon Boiss, 1 taksonla Scrodonia (Hill) Schreb, 3 taksonla Stachybotrys Benth, 3 taksonla Scordium (Mill.) Benth ve 1 taksonla Spinularia Boiss olmak üzere sekiz

bölümde sınıflandırılır. Bunlar çanak şekli ve çiçeklenme yapısı bakımından birbirlerinden ayırt edilebilir [50, 51].

Teucrium türleri çalı ve bodur çalılardır. Bu Teucrium türleri açık, kuru, kayalık yerlerde, yamaçlarda ve dağınık alanlarda yetişen çok yıllık veya senelik bitkilerdir. Teucrium türleri bitki tacının üst dudağının yetersizliği ve temel jine olmayışı bakımdan diğer Lamiaceae familyasının diğer üyelerinden ayırt edilebilir.

Öte yandan, doğadaki birçok bitki türünün faydası olduğu gibi Teucrium cinsinin de faydaları bulunmaktadır. Bunlardan birkaçı, Teucrium cinsine ait çok sayıdaki tıbbi tür halk hekimliğinde ve eczacılıkta kullanılmaktadır. Bazı Teucrium türleri ateş düşürücü, ağrı kesici, hemeroid, migren, yara iyileştirici, mide hastalıkları gibi hastalıklarda kullanılmaktadır. Ayrıca, sindirim tedavisinde, solunum bozukluklarında, apse oluşumunda, gut ve konjonktivit, yağ uyarılması ve selüloit bozulması durumunda, antioksidan, antibiyotik, anti-diyabetik ve anti-helmintik etkilere sahip durumlarda en çok kullanılan Teucrium türleri T.chamedrys, T.montamum ve T.poliumdur. Ancak bunların en önemli tedavi edici etkisi sindirim sistemindeki bazı sorunları yok etmesidir [52].

Bunların yanısıra, Teucrium cinsinin türleri çok güçlü biyolojik aktiviteyle meyve ve sebzelerin kendilerine özgü buruk tadını ve renklerini veren fenolik bileşikler bakımından çok zengindir ve ayrıca Anadolu'nun bazı bölgelerinde bazı Teucrium türlerinin kanseri önleyici bir etkisinin var olduğuna da inanılmaktadır. Son zamanlardaki çalışmalarda Teucrium cinsinden elde edilen bitki özleri ve izole bileşiklerin anti-kanser aktivitesine sahip olduğunu göstermektedir. Çalışmaların çoğu anti-kanserle ilgili Teucrium cinsinin türlerinde fenolik bileşiklerin önemini göstermektedir. Teucrium cinslerine ait türlerin araştırılması tamamlanmadığından dolayı, bazı Teucrium türlerinin kanser faaliyetleri ve potansiyel ilaç yapımı ile ilgili bilgiler bulunmamaktadır [52].

Bu tezin uygulamasında kullanılan Teucrium cinsine ait bazı türleri daha iyi tanıyabilmek için bu türlerin mikromorfolojik özelliklerinin karşılaştırılması aşağıdaki Tablo 3.1'de verilmiştir.

Tablo 3.1: Teucrium cinsine ait bazı türlerin özellikleri.

Özellikler/Türler	T.creticum	T.sandrasicum	T.brevifolium	T.pestalozzae	T.ekimii	T.alyssifolium
Türkiye'deki İsmi	Akpüren	Ülper	Vererik	Oğlanotu	Erkurtaran	Gül Mahmut
Türkiye'de Yetişme Alanı	Mersin-K.Maraş	Muğla	Muğla	Antalya-Burdur	Antalya	Muğla
Çiçeklenmesi	5-9	4-8	3-5	5-5	5-6	6-6
Yapısı	Çalı	Çalı	Çalı	Çalı	Çalımsı	Odunsu Ot
Rengi	Leylak rengi	Koyu Kahverengi	Mavi-Leylak	Mavi	Mavi	Mavi-Pembe
Şekli	Eliptik	Dikdörtgenimsi	Eliptik-Dikdörtgen	Eliptik	Eliptik	Yusyuvarlak

Özellikler/Türler	T.pseudaroanium	T.multicaule	T.orientale orien.	T.orientale puber.	T.orientale glabr.	T.pruinosum
Türkiye'deki İsmi	Tepeotu	Haptutan	Kirveotu-1	Kirveotu-2	Kirveotu-3	Puslu Mahmut
Türkiye'de Yetişme Alanı	Antalya	Sivas	Konya	Erzurum	Bayburt	Ankara
Çiçeklenmesi	4-6	4-7	6-9	6-9	6-8	5-7
Yapısı	Çalı	Odunsu Ot	Ot	Ot	Ot	Ot
Rengi	Açık Kahverengi	Soluk Leylak-Pembe	Mavi-Lila	Mavi	Mor-Mavi	Eflatun-Kırmızı
Şekli	Dikdörtgenimsi	Eliptik-Dikdörtgen	Dikdörtgen	Eliptik	Dikdörtgen-Eliptik	Yusyuvarlak

Özellikler/Türler	T.parviflorum	T.scordium scordioides	T.melissoides	T.chamaedrys cham.	T.chamaedrys lydium	T.chamaedrys trap.
Türkiye'deki İsmi	Koyun Otu	Kurtluca	Tetre	Kısa Mahmut	Bodur Mahmut	Dalak Otu
Türkiye'de Yetişme Alanı	Ankara	Tokat	Hakkari-Şırnak	Adana-Bolu	Kütahya-Çanakkale	Trabzon-Artvin
Çiçeklenmesi	5-8	5-9	3-8	6-8	6-7	6-8
Yapısı	Ot	Ot	Odunsu Ot	Ot	Ot	Ot
Rengi	Mor-Mavi	Leylak rengi	Morumsu-Pembe	Kırmızımsı-Mor	Kırmızımsı	Mor-Mavi
Şekli	Dikdörtgenimsi Oval	Dikdörtgenimsi Oval	Dikdörtgenimsi	Dikdörtgenimsi	Dikdörgenimsi Oval	Dikdörtgen

Özellikler/Türler	T.chamaedrys sysp.	T.chamaedrys sinu.	T.divaricatum diva.	T.divaricatum gra.	T.flavum ssp. hele.	T.leucophyllum
Türkiye'deki İsmi	Sıcak Otu	Sancı Otu	Mürcü Otu	Böce Otu	Sarı Yavşan	Buldumcuk
Türkiye'de Yetişme Alanı	Malatya	Hakkari-Muş	Antalya-İzmir	Hatay	Aydın-Çanakkale	Bitlis-Bingöl
Çiçeklenmesi	6-8	6-8	4-5	4-5	3-5	2-4
Yapısı	Ot	Ot	Odunsu Ot	Odunsu Ot	Odunsu Ot	Odunsu Ot
Rengi	Morumsu	Morumsu	Mor-Pembe	Morumsu-Pembe	Sarı	Eflatun-Kırmızı
Şekli	Dikdörtgenimsi Oval	Oval	Dikdörtgenimsi Oval	Dikdörtgenimsi Oval	Dikdörtgenimsi	Yuvarlak

Özellikler/Türler	T.microphyllum	T.montanum ssp. mont.	T.polium ssp. pol.	T.montbretii ssp. montb.	T.montbretti ssp. yıld.	T.lamifolium ssp. stach.
Türkiye'deki İsmi	Ada Yavşanı	Dağdalak	Acı Yavşan	Fatmacık Otu	Sürmeli Fatmacık Otu	Kumacı Otu
Türkiye'de Yetişme Alanı	Muğla-Antalya	Bursa-Çanakkale	Mardin-Bitlis	Kırklareli-İstanbul	Antalya	İstanbul-Zonguldak
Çiçeklenmesi	3-5	7-8	6-9	6-7	4-8	6-7
Yapısı	Odunsu Ot	Çalı	Odunsu Ot	Odunsu Ot	Odunsu Ot	Ot
Rengi	Pembe-Beyaz	Soluk Sarı	Beyazımtrak	Leylak	Beyazımtrak	Yeşilimsi-Beyaz
Şekli	Eliptik Oval	Dört Köşeli	Dört Köşeli	Dikdörtgenimsi Oval	Dikdörtgenimsi Oval	Dikdörgenimsi Oval

Şekil 3.1: Teucrium türlerine ait görüntüler.

Bu çalışmada faktör analizi ve kümeleme analizi 32 değişkene uygulanmış ve bu 32 değişkenin 14 özelliği göz önünde bulundurulmuştur. Ele alınan bu özellikler alfabetik harflerle isimlendirilmiştir. İsimlendirilen bu 14 özellik aşağıdaki gibi tanımlanmıştır:

Tablo 3.2: Teucrium cinsine ait bazı türlerin özellikleri ve tanımları.

Özellikler	Tanım
ch1	Meyvede longuitidal ridgeslerin var olup olmadığıdır
ch2	Meyve yüzeyinin özellikleridir
ch3	Meyvede B tipi salgı tüyünün var olup olmadığıdır
ch4	Meyvede A tipi salgı tüyünün var olup olmadığıdır
ch5	Meyvede örtü tüyünün var olup olmadığıdır
ch6	Meyvede ince çeperli örtünün var olup olmadığıdır
ch7	F2 tipi tüyün var olup olmadığıdır
ch8	F4 tipli tüyün var olup olmadığıdır
ch9	Yaprakta B tipi salgı tüyünün var olup olmadığıdır
ch10	Yaprakta A1 tipi tüyün var olup olmadığıdır
ch11	Yaprakta A2 tipi tüyün var olup olmadığıdır
ch12	Yaprakta ince çeperli örtünün var olup olmadığıdır
ch13	Yaprakta kalın çeperli tüy örtüsünün olup olmadığıdır
ch14	G2 tipli tüyün olup olmadığıdır

3.2. YÖNTEM

Bu çalışmada, çok değişkenli istatistiksel analiz yöntemlerinden faktör analizi ve kümeleme analizi olmak üzere iki yöntem kullanılmıştır. Araştırmada öncelikle Türkiye üzerinde yetişen Teucrium cinsine ait bazı türlerin sınıflandırması amacıyla faktör analizinde faktör türetme yöntemlerinden biri olan temel bileşenler analizi veri tabanına uygulanmıştır. Aynı zamanda, veri tabanına faktör analizinin uygulabilirliğini araştırmak için Kaiser-Meyer-Olkin örneklem yeterliği ölçüsü kullanılmıştır. Faktörlerin sayılarını belirlemek için hem özdeğer ölçütü 1'den büyük olan faktörlere bakılmış hem de yamaç grafiği incelenmiştir. Ayrıca, faktör analizinin son aşamalarından biri olan faktör yükleri ve faktör skorlarının hesaplanması yapılıp, ilgili faktör skorları dikkate alınarak kümeleme analizi yapılmasına kolaylık sağlanmaktadır. Daha sonraki aşamada ise, öklidyen mesafe kullanılarak kümeleme analizinin hiyerarşik kümeleme tekniklerinden olan tek bağlantılı, tam bağlantılı ve Ward kümeleme algoritmaları uygulanmıştır.

Faktör analizi ve kümeleme analizi teknikleri ile kümelendirme yapılırken Statistica 20,

PAST 3.09 ve Mathematica 9.01 paket programlarından faydalanılmıştır.

4. BULGULAR

Bu bölümde, önceki bölümlerde teoriksel olarak verilen benzerlik ya da uzaklık ölçüleri, faktör analizi ve kümeleme analizi Bölüm 3'de bahsedilen Teucrium türlerine ait verilere uygulanır. Böylece, Teucrium türleri arasındaki akrabalık ilişkileri saptanarak, onların sınıflandırılmasına çalışılır: Elde edilen simülasyon sonuçları gerek tıp, kimya ve eczacılık gibi bilim alanlarında hem ilaç yapımında hem de diğer sanayilerde kullanılabilecektir. 32 adet Teucrium cinsinin türlerine ve karakteristik özelliklerine ait veri tabanı Tablo 4.1'de verilmiştir. Tablo 4.1'e bakıldığında, hangi Teucrium türlerinin birbiriyle aynı sınıfta olduğu doğrudan görülmemektedir.

Kümeleme veya sınıflandırma analizi literatürde, çok boyutlu ve karmaşık yapıda bulunan verilere ait türlerin hangilerinin birbirleriyle ilişkili olduğunu az da olsa zihinde canlandırmak amacıyla benzerlik veya uzaklık ölçülerini kullanmaktadır. Bu sayede, türler arasındaki uzaklıklar veya benzerlikler ortaya konulur. Tablo 4.1'de bulunan Teucrium türlerinin birbirlerine göre öklidyen uzaklıklarını içeren matris Tablo 4.2 ile verilmiştir. Bu uzaklık matrisi incelendiğinde, hangi türlerin küme oluşturacağı hakkında herhangi bir çıkarım yapılamamaktadır. Bu sorunu gidermek için, bu veri tabanına sırasıyla çok değişkenli istatistiksel analiz yöntemlerinden kümeleme analizinin tek bağlantı, tam bağlantı ve Ward kümeleme teknikleri ve daha sonra faktör analizinin faktör türetme yöntemi olan temel bileşenler analizi uygulanmıştır.

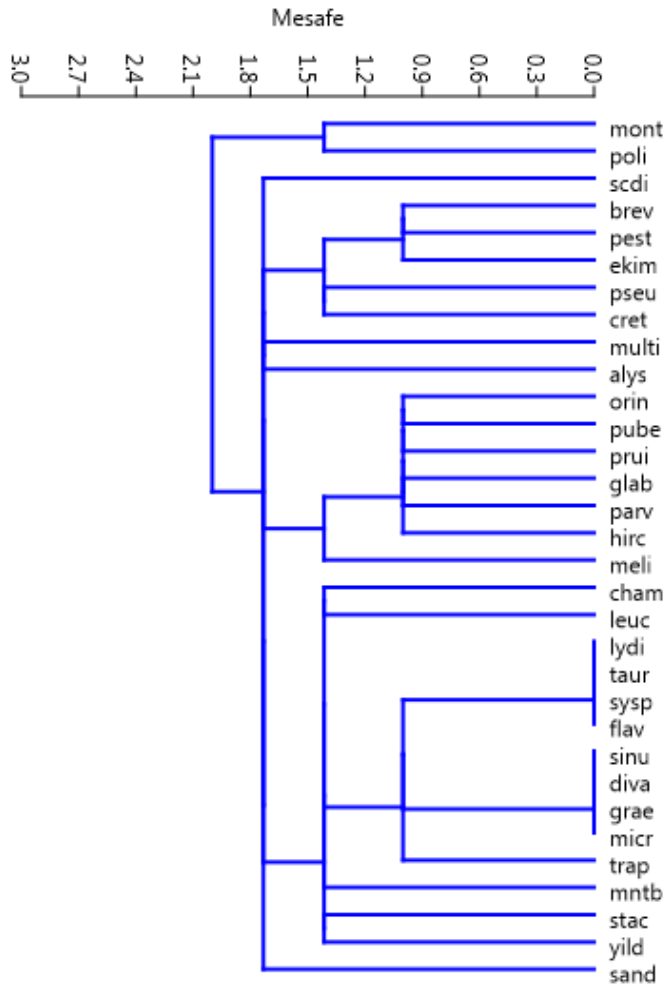
Tablo 4.1: Teucrium Türlerine Ait Veri Matrisi.

	ch1	ch2	ch3	ch4	ch5	ch6	ch7	ch8	ch9	ch10	ch11	ch12	ch13	ch14
cret	1	1	1	0	1	1	1	0	1	1	0	1	1	0
sand	1	1	1	0	0	0	0	0	0	1	0	0	1	0
brev	1	1	1	0	1	1	0	1	0	1	0	0	1	1
pest	1	1	1	0	1	1	0	1	1	1	0	0	1	1
ekim	1	1	1	0	1	1	0	1	1	1	0	0	1	0
alys	0	2	1	0	1	1	1	0	1	1	0	0	1	0
pseu	1	1	1	0	1	1	1	0	1	1	0	0	1	1
multi	0	3	1	0	1	1	1	1	1	1	0	0	1	1
orin	0	3	1	0	1	1	1	1	1	1	0	1	0	0
pube	0	3	1	0	1	1	0	1	1	1	0	1	0	0
glab	0	4	1	1	1	1	1	0	1	1	0	1	0	0
prui	0	4	1	0	1	1	1	1	1	1	0	1	0	0
parv	0	4	1	1	1	1	1	1	1	1	0	1	0	0
scdi	0	5	1	0	0	0	0	0	1	0	1	1	0	0
meli	0	5	1	0	1	1	1	0	1	1	0	1	0	0
cham	0	2	0	0	0	0	0	0	1	0	1	0	1	0
lydi	0	2	0	0	0	0	0	0	1	1	0	0	1	1
trap	0	2	1	0	0	0	0	0	1	1	0	0	1	0
taur	0	2	0	0	0	0	0	0	1	1	0	0	1	1
sysp	0	2	0	0	0	0	0	0	1	1	0	0	1	1
sinu	0	2	1	0	0	0	0	0	1	1	0	0	1	1
diva	0	2	1	0	0	0	0	0	1	1	0	0	1	1
grae	0	2	1	0	0	0	0	0	1	1	0	0	1	1
flav	0	2	0	0	0	0	0	0	1	1	0	0	1	1
leuc	0	2	1	0	0	0	0	0	1	0	1	0	1	1
micr	0	2	1	0	0	0	0	0	1	1	0	0	1	1
mont	0	6	0	0	0	0	0	0	1	1	0	0	1	0
poli	0	6	0	0	0	0	0	0	1	0	1	0	1	0
mntb	0	3	1	0	1	0	0	0	1	1	0	0	1	1
yild	0	3	1	0	0	0	0	0	1	1	1	0	1	0
stac	0	4	1	0	0	0	0	0	1	0	1	0	1	0
hirc	0	4	1	0	1	1	0	1	1	1	0	1	0	0

Tablo 4.2: Veri Matrisinin Uzaklık Ölçülerinin Matrisi.

Türler	cret	sand	brev	pest	ekim	alys	pseu	multi	orin	pube	glab	prui	parv	scdi	meli	cham	lydi	trap	taur	sysp	sinu	diva	grae	flav	leuc	micr	mont	poli	mntb	yild	stac	hirc
cret	0.	2.24	2.24	2.	1.73	1.73	1.41	2.83	2.64	2.83	3.46	3.46	3.60	4.79	4.24	3.	2.83	2.45	2.83	2.83	2.64	2.64	2.64	2.83	3.	2.64	5.57	5.74	3.	3.16	4.	3.6
sand	2.24	0.	2.	2.24	2.	2.45	2.24	3.32	3.46	3.32	4.12	4.12	4.24	4.69	4.79	2.45	2.24	1.73	2.24	2.24	2.	2.	2.	2.24	2.45	2.	5.29	5.48	2.83	2.64	3.6	4.
brev	2.24	2.	0.	1.	1.41	2.45	1.73	2.64	3.16	3.	4.12	3.87	4.	5.1	4.79	3.16	2.64	2.64	2.64	2.64	2.45	2.45	2.45	2.64	2.83	2.45	5.66	5.83	2.83	3.32	4.12	3.74
pest	2.	2.24	1.	0.	1.	2.24	1.41	2.45	3.	2.83	4.	3.74	3.87	5.	4.7	3.	2.45	2.45	2.45	2.45	2.24	2.24	2.24	2.45	2.64	2.24	5.57	5.74	2.64	3.16	4.	3.6
ekim	1.73	2.	1.41	1.	0.	2.	1.73	2.64	2.83	2.64	3.87	3.6	3.74	4.9	4.58	2.83	2.64	2.24	2.64	2.64	2.45	2.45	2.45	2.64	2.83	2.45	5.47	5.66	2.83	3.	3.87	3.46
alys	1.73	2.45	2.45	2.24	2.	0.	1.73	1.73	2.	2.24	2.64	2.64	2.83	4.	3.32	2.45	2.24	1.73	2.24	2.24	2.	2.	2.	2.24	2.45	2.	4.47	4.69	2.	2.24	3.	2.83
pseu	1.41	2.24	1.73	1.41	1.73	1.73	0.	2.45	3.	3.16	3.74	3.74	3.87	5.	4.47	3.	2.45	2.45	2.45	2.45	2.24	2.24	2.24	2.45	2.64	2.24	5.57	5.74	2.64	3.16	4.	3.87
multi	2.83	3.32	2.64	2.45	2.64	1.73	2.45	0.	1.73	2.	2.45	2.	2.24	3.6	2.83	3.	2.45	2.45	2.45	2.45	2.24	2.24	2.25	2.45	2.64	2.24	3.87	4.12	1.73	2.45	2.83	2.24
orin	2.64	3.46	3.16	3.	2.83	2.	3.	1.73	0.	1.	1.73	1.	1.42	3.16	2.24	3.16	3.	2.64	3.	3.	2.83	2.83	2.83	3.	3.16	2.83	4.	4.24	2.45	2.64	3.	1.41
pube	2.83	3.32	3.	2.83	2.64	2.24	3.16	2.	1.	0.	2.	1.41	1.73	3.	2.45	3.	2.83	2.45	2.83	2.83	2.64	2.64	2.64	2.83	3.	2.64	3.87	4.12	2.24	2.45	2.83	1.
glab	3.46	4.12	4.12	4.	3.87	2.64	3.74	2.45	1.73	2.	0.	1.41	1.	2.64	1.41	3.6	3.46	3.16	3.46	3.46	3.32	3.32	3.32	3.46	3.6	3.32	3.32	3.6	2.64	2.83	2.83	1.73
prui	3.46	4.12	3.87	3.74	3.6	2.64	3.74	2.	1.	1.41	1.41	0.	1.	2.64	1.41	3.6	3.46	3.16	3.46	3.46	3.32	3.32	3.32	3.46	3.6	3.32	3.32	3.6	2.64	2.83	2.83	1.
parv	3.6	4.24	4.	3.87	3.74	2.83	3.87	2.24	1.41	1.73	1.	1.	0.	2.83	1.73	3.74	3.6	3.32	3.6	3.6	3.46	3.46	3.46	3.6	3.74	3.46	3.46	3.74	2.83	3.	3.	1.41
scdi	4.79	4.7	5.1	5.	4.9	4.	5.	3.6	3.16	3.	2.64	2.64	2.83	0.	2.24	3.46	3.87	3.6	3.87	3.87	3.74	3.74	3.74	3.87	3.46	3.74	2.45	2.	3.16	2.64	1.73	2.45
meli	4.24	4.79	4.79	4.7	4.58	3.32	4.47	2.83	2.24	2.45	1.41	1.41	1.73	2.24	0.	4.12	4.	3.74	4.	4.	3.87	3.87	3.87	4.	4.12	3.87	2.64	3.	3.	3.16	2.83	1.73
cham	3.	2.45	3.16	3.	2.83	2.45	3.	3.	3.16	3.	3.6	3.6	3.74	3.46	4.12	0.	1.73	1.73	1.73	1.73	2.	2.	2.	1.73	1.41	2.	4.24	4.	2.45	1.73	2.24	3.46
lydi	2.83	2.24	2.64	2.45	2.64	2.24	2.45	2.45	3.	2.83	3.46	3.46	3.6	3.87	4.	1.73	0.	1.41	0.	0.	1.	1.	1.	0.	1.73	1.	4.12	4.36	1.73	2.	2.83	3.32
trap	2.45	1.73	2.64	2.45	2.24	1.73	2.45	2.45	2.64	2.45	3.16	3.16	3.32	3.6	3.74	1.73	1.41	0.	1.41	1.41	1.	1.	1.	1.41	1.73	1.	4.12	4.36	1.73	1.41	2.45	3.
taur	2.83	2.24	2.64	2.45	2.64	2.24	2.45	2.45	3.	2.83	3.46	3.46	3.6	3.87	4.	1.73	0.	1.41	0.	0.	1.	1.	1.	0.	1.73	1.	4.12	4.36	1.73	2.	2.83	3.32
sysp	2.83	2.24	2.64	2.45	2.64	2.24	2.45	2.45	3.	2.83	3.46	3.46	3.6	3.87	4.	1.73	0.	1.41	0.	0.	1.	1.	1.	0.	1.73	1.	4.12	4.36	1.73	2.	2.83	3.32
sinu	2.64	2.	2.45	2.24	2.45	2.	2.24	2.24	2.83	2.64	3.32	3.32	3.46	3.74	3.87	2.	1.	1.	1.	1.	0.	0.	0.	1.	1.41	0.	4.24	4.47	1.41	1.73	2.64	3.16
diva	2.64	2.	2.45	2.24	2.45	2.	2.24	2.24	2.83	2.64	3.32	3.32	3.46	3.74	3.87	2.	1.	1.	1.	1.	0.	0.	0.	1.	1.41	0.	4.24	4.47	1.41	1.73	2.64	3.16
grae	2.64	2.	2.45	2.24	2.45	2.	2.24	2.24	2.83	2.64	3.32	3.32	3.46	3.74	3.87	2.	1.	1.	1.	1.	0.	0.	0.	1.	1.41	0.	4.24	4.47	1.41	1.73	2.64	3.16
flav	2.83	2.24	2.64	2.45	2.64	2.24	2.45	2.45	3.	2.83	3.46	3.46	3.6	3.87	4.	1.73	0.	1.414	0.	0.	1.	1.	1.	0.	1.73	1.	4.12	4.36	1.73	2.	2.83	3.32
leuc	3.	2.45	2.83	2.64	2.83	2.45	2.64	2.64	3.16	3.	3.6	3.6	3.74	3.46	4.12	1.41	1.73	1.73	1.73	1.73	1.41	1.41	1.41	1.73	0.	1.41	4.47	4.24	2.	1.73	2.24	3.46
micr	2.64	2.	2.45	2.24	2.45	2.	2.24	2.24	2.83	2.64	3.32	3.32	3.46	3.74	3.87	2.	1.	1.	1.	1.	0.	0.	0.	1.	1.41	0.	4.24	4.47	1.41	1.73	2.64	3.16
mont	5.57	5.29	5.67	5.57	5.48	4.47	5.57	3.87	4.	3.87	3.32	3.32	3.46	2.45	2.64	4.24	4.12	4.12	4.12	4.12	4.24	4.24	4.24	4.12	4.47	4.24	0.	1.41	3.46	3.32	2.64	3.16
poli	5.74	5.48	5.83	5.74	5.66	4.7	5.74	4.12	4.24	4.12	3.6	3.6	3.74	2.	3.	4.	4.36	4.36	4.36	4.36	4.47	4.47	4.47	4.36	4.24	4.47	1.41	0.	3.74	3.32	2.24	3.46
mntb	3.	2.83	2.83	2.64	2.83	2.	2.64	1.73	2.45	2.24	2.64	2.64	2.83	3.16	3.	2.45	1.73	1.73	1.73	1.73	1.41	1.41	1.41	1.73	2.	1.41	3.46	3.74	0.	1.73	2.24	2.45
yild	3.16	2.64	3.32	3.16	3.	2.24	3.16	2.44	2.64	2.45	2.83	2.83	3.	2.64	3.16	1.73	2.	1.41	2.	2.	1.73	1.73	1.73	2.	1.73	1.73	3.32	3.32	1.73	0.	1.41	2.64
stac	4.	3.6	4.12	4.	3.87	3.	4.	2.83	3.	2.83	2.83	2.83	3.	1.73	2.83	2.24	2.83	2.45	2.83	2.83	2.64	2.64	2.64	2.83	2.24	2.64	2.64	2.24	2.24	1.41	0.	2.64
hirc	3.6	4.	3.74	3.6	3.46	2.83	3.87	2.24	1.41	1.	1.73	1.	1.41	2.45	1.73	3.46	3.32	3.	3.32	3.32	3.16	3.16	3.16	3.32	3.46	3.16	3.16	3.46	2.45	2.64	2.64	0.

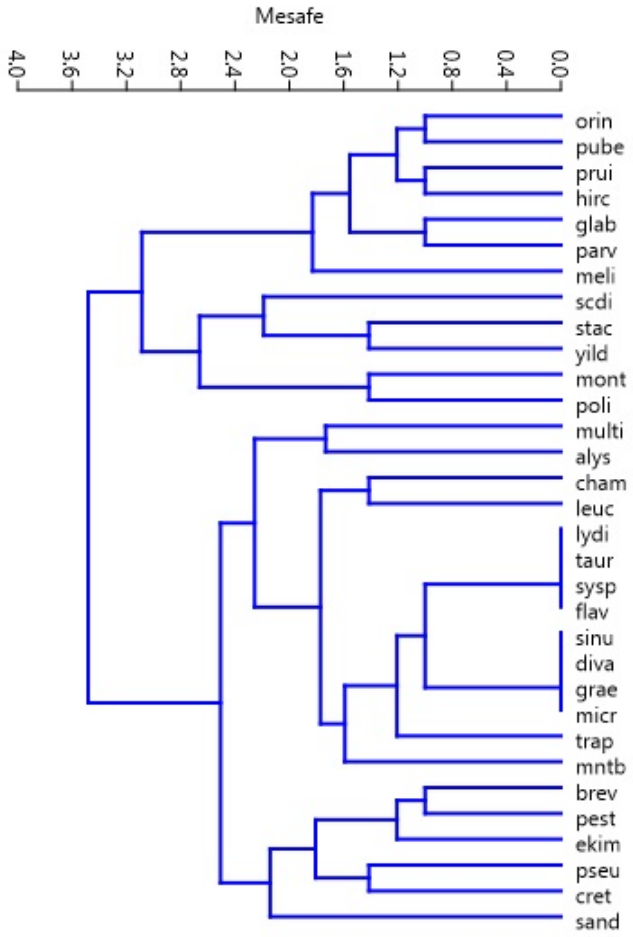
Bu veri tabanına Bölüm 2.1.5.1'de bahsedilen öklidyen mesafe baz alınarak çeşitli kümeleme teknikleri uygulanır. Bu kümeleme tekniklerinden birincisi olarak hiyerarşik kümeleme tekniklerinden biri olan tek bağlantı metodu uygulanmıştır. Tek bağlantı metoduna göre kümeleme sonuçlarının dendrogramı Şekil 4.1 ile verilmiştir.



Şekil 4.1: Teucrium Türlerinin Tek Bağlantı Metoduna Göre Dendrogramı.

Şekil 4.1'de görüldüğü gibi, 32 adet Teucrium türü 8 kümeye ayrılmış olarak düşünülebilir. Bu kümeler lydi, taur, sysp, flav, sinu, diva, grae, micr cham, leuc, mntb, stac, yild, trap türleri 1.4 birimlik mesafe ile, orin, pube, prui, glab, parv, hirc, meli türleri 1.4 birimlik mesafe ile, brev, pest, ekim, cret türleri 1.4 birimlik mesafe ile, mont ve poli türleri 1.4 birimlik mesafe ile küme oluşturmuşlar ve daha sonra bunlarla scdi, multi, alys ve sand türleri bağımsız kümeler oluşturarak 1.7 birimlik mesafe ile küme oluşturmuşlardır. Gruplayıcı hiyerarşik kümeleme algoritmasının en son adımı olarak bütün türler birleşip, 2 birimlik mesafeyle tek bir küme oluşturmuştur.

Şekil 4.2'de, Teucrium verilerine hiyerarşik kümeleme tekniği olan tam bağlantı metodunun uygulanması sonucu elde edilen bilgilerin dendrogram ile gösterilmesi verilmiştir.

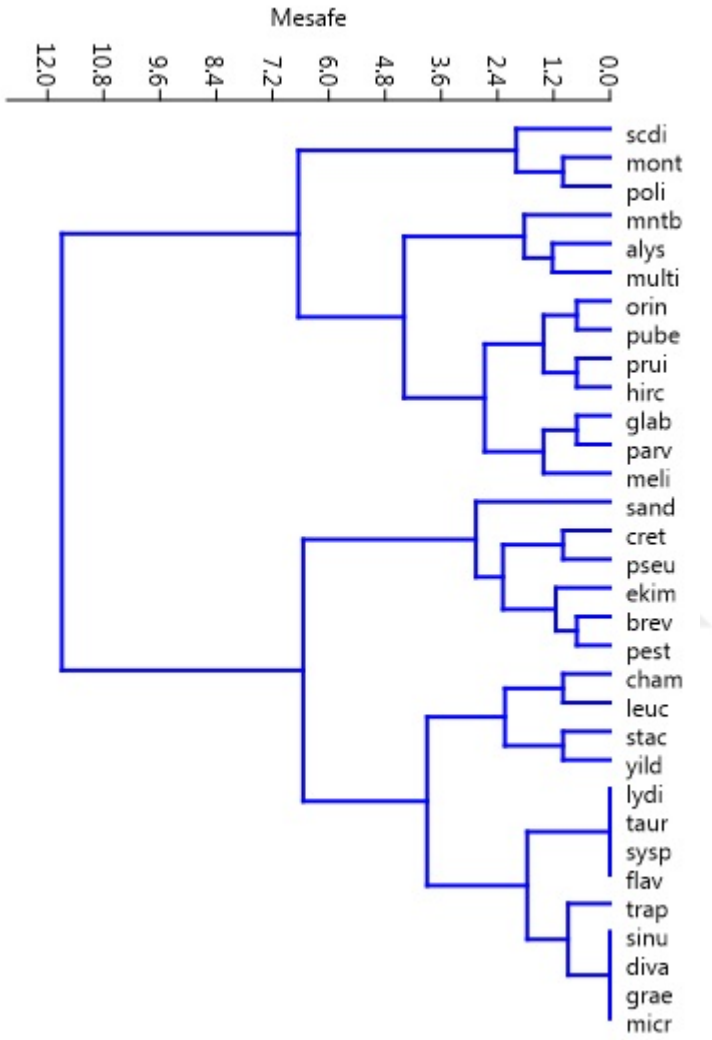


Şekil 4.2: Teucrium Türlerinin Tam Bağlantı Metoduna Göre Dendrogramı.

Şekil 4.2'ye göre, Teucrium türlerinin 7 kümeye ayrıldığı söylenebilir. Bu kümeler orin, pube, prui, hirc, glab, parv, meli türleri 1.8 birimlik mesafe ile, scdi, stac, yild türleri 2.2 birimlik mesafe ile, mont, poli türleri 1.4 birimlik mesafe ile, multi, alys türleri 1.8 birimlik mesafe ile, cham, leuc türleri 1.4 birimlik mesafe ile, lydi, taur, sysp, flav, sinu, diva, grae, micr, trap, mntb türleri 1.6 birimlik mesafe ile, brev, pest, ekim, pseu, cret, sand türleri 2.2 birimlik mesafe ile küme oluşturmuşlardır. Daha sonrasında hiyerarşik olarak kümeleme devam ettiğinde orin, pube, prui, hirc, glab, parv, meli, scdi, stac, yild, mont, poli türlerinin bulunduğu küme 3 birimlik mesafe ile, geriye kalan türlerin bulunduğu küme ise 2.4 birimlik mesafe ile birer küme oluşturmuşlardır. En sonunda, bir önceki

adımda oluşan iki küme 3.4 birimlik mesafe ile tek bir küme oluşturur.

Teucrium türlerine ait veri tabanına uygulanacak olan en son kümeleme tekniği ise Ward kümeleme tekniğidir. Teucrium türlerinin Ward kümeleme tekniğine göre dendrogramı Şekil 4.3'de verilmiştir.



Şekil 4.3: Teucrium Türlerinin Ward Kümeleme Metoduna Göre Dendrogramı.

Ward kümeleme tekniğine göre sınıflandırılan Teucrium türlerinin 8 kümeye ayrıldığı söylenebilir. Bu kümeler scdi, mont, poli türleri 1.8 birimlik mesafe ile, mntb, alys, multi türleri 1.8 birimlik mesafe ile, orin, pube, prui, hirc türleri 1.6 birimlik mesafe ile, glab, parv, meli türleri 1.6 birimlik mesafe ile, sand, cret, pseu, ekim, brev, pest türleri 2.8 birimlik mesafe ile cham, leuc, stac, yild türleri 3 birimlik mesafe ile, lydi, taur, sysp, flav türleri 0 birimlik mesafe ile, trap, sinu, diva, grae, micr türleri 0.9 birimlik

mesafe ile küme oluşturdukları söylenebilir. Daha sonrasında hiyerarşik olarak kümelemeye devam edildiğinde scdi, mont, poli, mntb, alys, multi, orin, pube, prui, hirc, glab, parv, meli türleri 6.4 birimlik mesafe ile, geriye kalan türler ise yine aynı şekilde 6.4 birimlik mesafe ile birer küme oluşturmaktadırlar. En sonunda ise bu iki küme hiyerarşik kümelemenin son adımı olan tek bir küme oluşturmayı 11.6 birimlik mesafe ile gerçekleştirmektedir.

Her üç kümeleme tekniğine göre, Teucrium türlerinin küme sayılarının aşağı yukarı aynı olduğu fakat kümelerin elemanlarının az da olsa bazı farklılıklar sergilediği saptanmıştır. Bu durumda, hangi yöntemin daha kullanışlı olduğunu veya hangisinin daha doğru bir teknik olduğunu anlamak gibi bir problem ortaya çıkmaktadır. Bu tür durumlar karşısında ya araştırmacının amaçları doğrultusunda hangi kümeleme algoritmasının kullanılacağına karar verilir ya da başka bir çok değişkenli istatistiksel analiz tekniği kullanılarak, her iki analiz yönteminin sonuçları karşılaştırılır.

Teucrium türlerine ait verilerin kümeleme algoritmaları sonucu ortaya çıkan kümeleri hem karşılaştırmak hem de kümeleme sonucu ortaya çıkan kümeleri iki boyutlu düzlem üzerinde gösterebilmek maksadıyla, çok değişkenli istatistiksel analiz yöntemi olan faktör analizi Teucrium türlerine ait veri tabanına uygulansın.

Bunun için öncelikle faktör analizinin önemli bir noktası olan veri tabanına Bölüm 2.2.2.5 de bahsedilen Kaiser-Meyer-Olkin örneklem yeterliği ölçüsü uygulanarak faktör analizinin Teucrium verilerine uygulanıp uygulanmayacağına karar verilir. Elde edilen Teucrium türlerine ait KMO değeri 0.7 olarak bulunmuştur. Bu değer veri tabanının örneklem yeterliliği bakımından orta düzeyde olduğu bilgisini ortaya çıkarmaktadır. Buradan bu veri tabanına faktör analizinin uygulanabileceği sonucuna varılmaktadır. Bu durumda, 14 adet karakterin lineer ilişkisini gösteren ve faktör analizine ait önemli bilgiler ortaya koyan korelasyon matrisi oluşturulabilir. Bu korelasyon matrisi Tablo 4.3'de verilmektedir.

Tablo 4.3: Veri Matrisinin Korelasyon Matrisi.

	ch1	ch2	ch3	ch4	ch5	ch6	ch7	ch8	ch9	ch10	ch11	ch12	ch13	ch14
ch1	1.	-0.581704	0.254196	-0.124035	0.350966	0.383311	0.0556484	0.233723	-0.537484	0.206725	-0.230769	-0.122426	0.27735	0.0605228
ch2	-0.581704	1.	-0.158108	0.233071	-0.0344683	-0.0471551	0.123951	-0.0229539	0.312656	-0.327816	0.320819	0.417761	-0.521163	-0.490968
ch3	0.254196	-0.158108	1.	0.136626	0.49705	0.466667	0.331006	0.331006	-0.136626	0.188674	-0.13315	0.331006	-0.305505	-0.142857
ch4	-0.124035	0.233071	0.136626	1.	0.274874	0.29277	0.412759	0.125622	0.0666667	0.111111	-0.124035	0.412759	-0.447214	-0.22771
ch5	0.350966	-0.0344683	0.49705	0.274874	1.	0.938872	0.665942	0.665942	-0.016169	0.404226	-0.451243	0.52666	-0.470016	-0.197242
ch6	0.383311	-0.0471551	0.466667	0.29277	0.938872	1.	0.709299	0.709299	-0.03253	0.379517	-0.423659	0.569191	-0.509175	-0.269841
ch7	0.0556484	0.123951	0.331006	0.412759	0.665942	0.709299	1.	0.227053	0.161515	0.269191	-0.300501	0.536232	-0.441415	-0.27146
ch8	0.233723	-0.0229539	0.331006	0.125622	0.665942	0.709299	0.227053	1.	-0.125622	0.269191	-0.300501	0.381643	-0.441415	-0.131352
ch9	-0.537484	0.312656	-0.136626	0.0666667	-0.016169	-0.03253	0.161515	-0.125622	1.	-0.111111	0.124035	0.161515	-0.149071	-0.03253
ch10	0.206725	-0.327816	0.188674	0.111111	0.404226	0.379517	0.269191	0.269191	-0.111111	1.	-0.895806	0.0777663	-0.0496904	0.206023
ch11	-0.230769	0.320819	-0.13315	-0.124035	-0.451243	-0.423659	-0.300501	-0.300501	0.124035	-0.895806	1.	-0.122426	0.09245	-0.262265
ch12	-0.122426	0.417761	0.331006	0.412759	0.52666	0.569191	0.536232	0.381643	0.161515	0.0777663	-0.122426	1.	-0.922958	-0.551677
ch13	0.27735	-0.521163	-0.305505	-0.447214	-0.470016	-0.509175	-0.441415	-0.441415	-0.149071	-0.0496904	0.09245	-0.922958	1.	0.509175
ch14	0.0605228	-0.490968	-0.142857	-0.22771	-0.197242	-0.269841	-0.27146	-0.131352	-0.03253	0.206023	-0.262265	-0.551677	0.509175	1.

Korelasyon matrisi incelendiğinde, ch5-ch6 arasında 0.938872 ile, ch6-ch7 arasında 0.709299 ile, ch6-ch8 arasında 0.709299 ile ch10-ch11 arasında -0.895806 ile ch12-ch13 arasında -0.922958 ile pozitif ve negeatif yönlü yüksek bir ilişki bulunmaktadır. Yani, buradaki ikili karakterler arasında ortak bir özellik vardır ve ch5, ch6, ch7, ch8, ch10, ch11, ch13 karakterlerinin toplam varyansın büyük bir kısmını açıklayan faktörler civarında yer alması öngörülür. Kalan diğer çoğu karakterler arasında ise orta dereceli bir ilişki olduğu görülmektedir.

Faktör analizinin uygulamasında, verilerin korelasyon matrisine ait özvektörler ve özdeğerler göz önüne alınır. Çünkü, özvektörler faktör yüklerinin hesaplanmasında kullanılırken, özdeğerler ise veri tabanındaki değişimi göstermektedir. Bu sebeplerden dolayı, veri tabanına ait korelasyon matrisinin özvektörleri ve özdeğerleri hesaplanır. Teucrium türlerine ait veri matrisinin korelasyon matrisinin özvektörlerine ve özdeğerlerine ilişkin değerler aşağıdaki Tablo 4.4'de verilmektedir.

Tablo 4.4: Kolerasyon Matrisinin Özvektör-Özdeğer Matrisi.

Değişken	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ch1	0.083	0.397	0.419	-0.068	0.152	-0.281	-0.043	-0.217	0.195	-0.53	-0.373	-0.023	0.073	0.192
ch2	0.063	-0.471	-0.06	-0.104	-0.212	-0.124	-0.055	0.377	-0.429	-0.56	-0.216	-0.009	-0.012	-0.082
ch3	0.243	0.101	0.256	0.183	0.237	0.792	-0.179	0.004	-0.284	-0.104	-0.094	0.105	0.006	-0.05
ch4	0.212	-0.152	-0.161	-0.497	0.451	0.064	0.594	-0.27	-0.132	-0.062	0.014	0.011	-0.032	-0.031
ch5	0.401	0.126	0.048	0.23	0.024	-0.154	0.085	0.119	-0.116	-0.216	0.65	-0.086	-0.328	0.349
ch6	0.413	0.113	0.088	0.199	0.022	-0.229	0.086	0.044	-0.026	0.01	0.163	-0.061	0.505	-0.646
ch7	0.335	-0.038	-0.128	0.12	0.495	-0.257	-0.155	0.441	0.009	0.357	-0.388	0.029	-0.064	0.2
ch8	0.302	0.107	0.147	0.186	-0.539	-0.018	0.459	-0.113	-0.213	0.282	-0.388	0.067	-0.19	0.091
ch9	0.004	-0.271	-0.424	0.631	0.144	-0.049	0.003	-0.5	0.0292	-0.204	-0.14	0.041	0.05	0.076
ch10	0.191	0.331	-0.44	-0.215	-0.14	0.066	-0.287	-0.16	-0.215	0.008	-0.116	-0.648	0.001	0.017
ch11	-0.206	-0.334	0.448	0.21	0.167	0.013	0.189	0.013	-0.004	0.115	-0.02	-0.721	-0.011	0.019
ch12	0.352	-0.253	0.023	-0.096	-0.097	0.106	-0.158	-0.065	0.566	-0.082	-0.081	-0.076	-0.528	-0.363
ch13	-0.334	0.294	0.013	0.137	0.237	-0.216	-0.004	-0.046	-0.378	0.001	-0.038	0.056	-0.554	-0.47
ch14	-0.179	0.312	-0.321	0.208	0.005	0.255	0.464	0.479	0.326	-0.268	-0.101	-0.139	0.001	-0.065
Özdeğerler	4.903	3.147	1.509	0.903	0.866	0.701	0.639	0.404	0.386	0.246	0.122	0.086	0.047	0.033

Tablo 4.4 incelendiğinde, en büyük değişimi 4.903 özdeğeri ile ch1'in gerçekleştirdiği ve en az değişimi ise 0.033 özdeğeri ile ch14'ün gerçekleştirdiği görülmektedir. Böylelikle, 4.903 özdeğeri birinci faktöre ait olup bu yapı için anlamlı bir değere sahip olurken, 0.033 özdeğeri ise en sonuncu faktöre ait olup bu yapı için anlamsız bir değere sahip olmaktadır. Tablo 4.4'deki korelasyon matrisine ait özvektörler ve özdeğerler kullanılarak, Bölüm 2.2.7'de bahsedilen faktör yüklerinin ve değişimin yüzdeleri hesaplanarak Tablo 4.5'de verilir.

Tablo 4.5: Faktör Yükleri Matrisi.

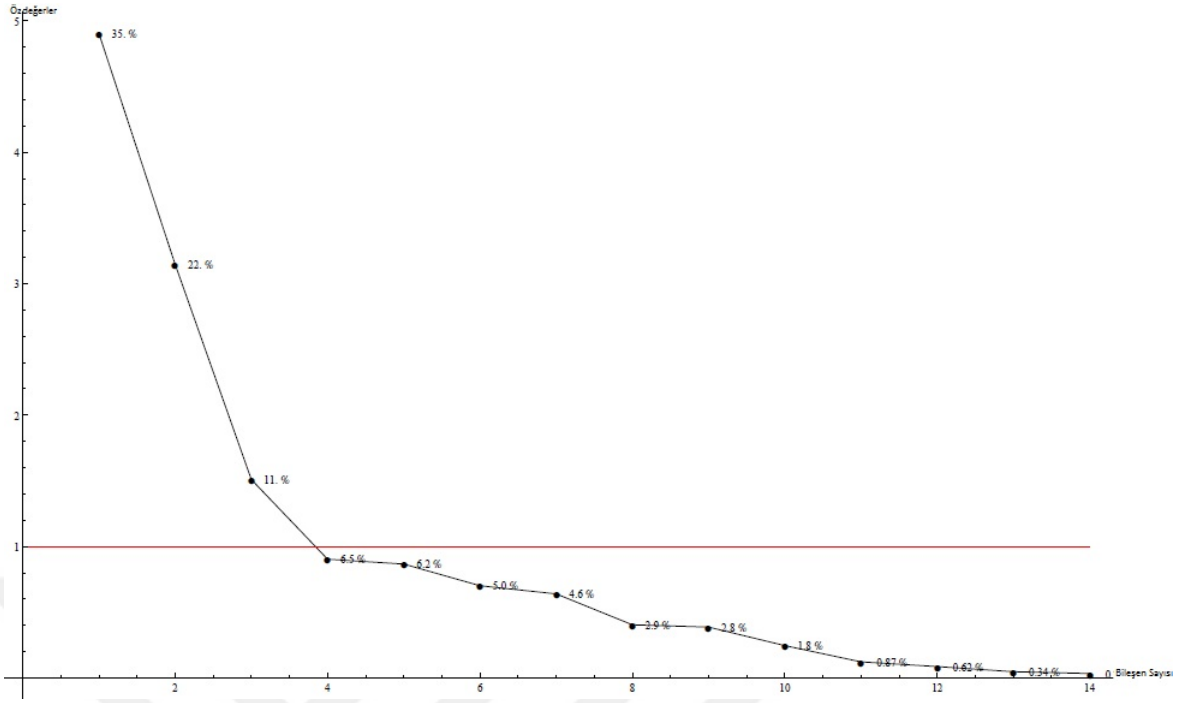
	Faktör1	Faktör2	Faktör3	Faktör4	Faktör5	Faktör6	Faktör7	Faktör8	Faktör9	Faktör10	Faktör11	Faktör12	Faktör13	Faktör14
ch1	0.184	-0.704	0.515	-0.065	0.142	-0.235	-0.034	-0.138	0.121	0.263	-0.13	0.006	0.0159	0.035
ch2	0.139	0.837	-0.074	-0.099	-0.197	-0.103	-0.044	0.239	-0.267	0.278	-0.075	0.002	-0.002	-0.015
ch3	0.539	-0.18	0.315	0.174	0.221	0.663	-0.143	0.0028	-0.176	0.051	-0.033	-0.031	0.001	-0.0091
ch4	0.47	0.27	-0.198	-0.472	0.419	0.054	0.475	-0.171	-0.082	0.0309	0.004	-0.003	-0.007	-0.005
ch5	0.889	-0.224	0.059	0.219	0.022	-0.129	0.068	0.075	-0.072	0.107	0.227	0.0254	-0.071	0.063
ch6	0.915	-0.2005	0.108	0.19	0.021	-0.192	0.068	0.028	-0.016	-0.0051	0.057	0.017	0.11	-0.118
ch7	0.743	0.067	-0.157	0.114	0.461	-0.215	-0.124	0.280	0.0058	-0.177	-0.135	-0.008	-0.014	0.0366
ch8	0.67	-0.189	0.181	0.177	-0.501	-0.015	0.367	-0.072	-0.132	-0.14	-0.135	-0.0197	-0.0415	0.0167
ch9	0.01	0.482	-0.521	0.6	0.134	-0.041	0.002	-0.318	0.018	0.101	-0.0492	-0.012	0.011	0.014
ch10	0.424	-0.587	-0.541	-0.204	-0.13	0.055	-0.23	-0.101	-0.133	-0.0041	-0.0407	0.19	0.00021	0.0031
ch11	-0.456	0.592	0.551	0.2	0.155	0.0113	0.151	0.0084	-0.0029	-0.057	-0.007	0.212	-0.002	0.003
ch12	0.78	0.45	0.0291	-0.0912	-0.09	0.0891	-0.126	-0.041	0.352	0.041	-0.028	0.022	-0.115	-0.0666
ch13	-0.741	-0.522	0.0162	0.1302	0.221	-0.181	-0.003	-0.029	-0.235	-0.0005	-0.0136	-0.0164	-0.12	-0.0861
ch14	-0.398	-0.554	-0.394	0.198	0.005	0.214	0.371	0.304	0.203	0.133	-0.0356	0.0409	0.0003	-0.0119
Özdeğerler	4.903	3.147	1.509	0.903	0.866	0.701	0.639	0.404	0.386	0.246	0.122	0.086	0.047	0.033
%	35.021	22.478	10.7785	6.45	6.1857	5.007	4.564	2.888	2.757	1.757	0.8714	0.6142	0.3357	0.235
Birikimli %	35.021	57.499	68.27775	74.7275	80.93635	85.94855	90.51785	93.40585	96.16995	97.93745	98.81135	99.42905	99.76865	100

Tablo 4.5'e göre elde edilen özdeğerlerin toplamı, yaklaşık olarak 14 olup bu toplam varyansı ifade eder. Tablo 4.5 incelendiğinde, Faktör1 toplam varyansın $\%35.021$'sini $(4.903/14)$, Faktör2 toplam varyansın $\%22.478$'ini $(3.147/14)$, Faktör 3 ise toplam varyansın $\%10.778$'unu $(1.509/14)$ açıkladığı görülmektedir. Birikimli $\%$ satırına bakıldığında ise, ilk üç faktörün toplam varyansın $\%68.277$'sini açıkladığı anlaşılmaktadır.

Faktör yükleri, ilgili değişkenin o faktör üzerindeki ağırlığı olarak tanımladığından, ch3, ch4, ch5, ch6, ch7, ch8, ch12, ch13 karakterlerinin Faktör1 üzerindeki; ch1, ch2, ch11, ch14 karakterlerinin Faktör2 üzerindeki ve ch9, ch10 karakterlerinin ise Faktör3 üzerindeki ağırlıkları diğer faktörlere göre daha fazla olduğu bilgisine ulaşılmıştır. Birikimli $\%$ satırında ilk üç faktöre ait $\%68.292$ değeri her ne kadar istatistiksel olarak yeterli olmasa bile, bütün karakterlerin ilk üç faktör içinde bulunmasından dolayı bu değer yeterli olarak görülmektedir. Dolayısıyla, Teucrium türlerine ait 14 karakterin 3 faktörle büyük bir kısmının açıklanacağı bilgisine ulaşılmıştır.

Tablo 4.5'te ilk üç faktör dışındaki diğer geriye kalan faktörlerin sahip oldukları faktör yüklerinin değerlerinin istatistiksel olarak anlamlı değerler olmadığı görülmektedir. Bundan dolayı, ilk üç faktör dışındaki geriye kalan diğer faktörler faktör analizinde genelde önemli faktörler olarak görülmezler.

Diğer yandan, özdeğer ölçütünün sınırı olarak 1 alındığından itibaren eğimin değişmezliğe ulaştığı ya da çok az azalan değerlere ulaştığı faktörün Faktör3 olduğu Şekil 4.4'deki yamaç grafiğinden de görülmektedir. Dolayısıyla, bu grafik ile Teucrium türlerine ait veri tabanının, faktör analizi sonucu oluşturulan ilk 3 faktör tarafından açıklanabileceği istatistiksel olarak yeterli görülmekte ve yukarıda bahsedilen bilgileri doğrulamaktadır. Ancak, bu çalışmada Teucrium türlerini iki boyutlu düzlem üzerinde gösterebilmek amacı ile yamaç grafiğinde en çok eğimin olduğu ilk iki faktör baz alınmıştır.



Şekil 4.4: Yamaç grafiği.

Diğer önemli bir nokta ise, faktörlerin Teucrium türlerine ait karakterlerinde bulunan toplam varyansın kaçta kaçını açıkladığını belirlemektir. Bu duruma ait bilgiler aşağıda Tablo 4.6'de verilmektedir.

Tablo 4.6: İlk İki Faktör Yükleri ve Ortak Varyanslar.

	Faktör1	Faktör2	h_1^2	h_2^2	h_{1+2}^2
ch1	0.184796	-0.704955	0.0341495	0.496962	0.531111
ch2	0.139643	0.837183	0.0195002	0.700876	0.720376
ch3	0.539923	-0.180775	0.291516	0.0326798	0.324196
ch4	0.470252	0.270327	0.221137	0.0730768	0.294214
ch5	0.889874	-0.224623	0.791876	0.0504556	0.842331
ch6	0.915768	-0.2005	0.838631	0.0402001	0.878831
ch7	0.743244	0.0677153	0.552411	0.004585	0.556997
ch8	0.670137	-0.189991	0.449083	0.0360967	0.48518
ch9	0.0100792	0.482321	0.000101591	0.232633	0.232735
ch10	0.424519	-0.587675	0.180216	0.345362	0.525578
ch11	-0.456198	0.592862	0.208116	0.351486	0.559602
ch12	0.780511	0.450209	0.609198	0.202688	0.811886
ch13	-0.741108	-0.52231	0.549242	0.272807	0.822049
ch14	-0.398367	-0.554606	0.158696	0.307588	0.466284

Tablo 4.6'ya göre, Faktör1 en fazla ch6 karakteri ile toplam varyansın $(0.915768)^2 = 0.838631$'sını, en az ch9 karakteri ile toplam varyansın $(0.0100792)^2 = 0.000101591$'sını açıklamaktadır. Faktör2 ise en fazla ch2 karakteri ile

toplam varyansın $(0.837183)^2 = 0.700876$'sını, en az ch7 karakteri ile toplam varyansın $(0.0677153)^2 = 0.004585$'sını açıklamaktadır. Bunlara ek olarak, ilk iki faktörün açıkladığı ortak varyansın en düşük miktarı $\%0.232735$ ile ch9 karakterinin varyansı, en fazla miktarı ise $\%0.878831$ ile ch6 karakterinin varyansı olduğu Tablo 4.6'da görülür.

Faktör analizinin son aşamalarından biri de faktör skorunun hesaplanmasıdır. Faktör skorunun hesaplanabilmesi için öncelikle orijinal veri matrisinin standartlaştırılması gerekmektedir. Daha sonra, standartlaştırılmış veri matrisi ile faktör yükleri matrisinin çarpılması sonucunda faktör skoru elde edilir.

Orijinal veri matrisinin standartlaştırılması için öncelikle Teucrium türlerine ait karakterlerin aritmetik ortalaması ve standart sapmasına ilişkin temel istatistiksel bilgiler elde edilir. Ardından, temel istatistiksel bilgiler doğrultusunda bu çalışma için $Z = \frac{x-\mu}{s}$ dönüşümü kullanılarak orijinal veri matrisi standartlaştırılır. Bu bağlamda, temel istatistiksel bilgilere ve standartlaştırılmış veri matrisine ait değerler Tablo 4.7 ve Tablo 4.8'de verilmiştir.

Tablo 4.7: Karakterlere Ait Temel İstatistiksel Bilgiler.

Karakterler	Aritmetik Ortalama	Standart Sapma
ch1	0.1875	1.681134
ch2	2.71875	1.442095
ch3	0.78125	0.4200134
ch4	0.0625	0.2459347
ch5	0.46875	0.5070073
ch6	0.4375	0.5040161
ch7	0.28125	0.4568034
ch8	0.28125	0.4568034
ch9	0.9375	0.245934
ch10	0.84375	0.368902
ch11	0.1875	0.3965578
ch12	0.28125	0.4568034
ch13	0.75	0.4399413
ch14	0.4375	0.5040161

Tablo 4.8: Teucrium Veri Matrisinin Standartlaştırılmış Hali.

	Zch1	Zch2	Zch3	Zch4	Zch5	Zch6	Zch7	Zch8	Zch9	Zch10	Zch11	Zch12	Zch13	Zch14
cret	0.483	-1.191	0.52	-0.254	1.047	1.116	1.573	-0.615	0.254	0.423	-0.472	1.573	0.568	-0.868
sand	0.483	-1.191	0.52	-0.254	-0.924	-0.868	-0.615	-0.615	-3.8119	0.423	-0.472	-0.615	0.568	-0.868
brev	0.483	-1.191	0.52	-0.254	1.047	1.116	-0.615	1.573	-3.8119	0.423	-0.472	-0.615	0.568	1.116
pest	0.483	-1.191	0.52	-0.254	1.047	1.116	-0.615	1.573	0.254	0.423	-0.472	-0.615	0.568	1.116
ekim	0.483	-1.191	0.52	-0.254	1.047	1.116	-0.615	1.573	0.254	0.423	-0.472	-0.615	0.568	-0.868
alys	-0.111	-0.498	0.52	-0.254	1.047	1.116	1.573	-0.615	0.254	0.423	-0.472	-0.615	0.568	-0.868
pseu	0.483	-1.191	0.52	-0.254	1.047	1.116	1.573	-0.615	0.254	0.423	-0.472	-0.615	0.568	1.116
multi	-0.111	0.195	0.52	-0.254	1.047	1.116	1.573	1.573	0.254	0.423	-0.472	-0.615	0.568	1.116
orin	-0.111	0.195	0.52	-0.254	1.047	1.116	1.573	1.573	0.254	0.423	-0.472	1.573	-1.704	-0.868
pube	-0.111	0.195	0.52	-0.254	1.047	1.116	-0.615	1.573	0.254	0.423	-0.472	1.573	-1.704	-0.868
glab	-0.111	0.888	0.52	3.811	1.047	1.116	1.573	-0.615	0.254	0.423	-0.472	1.573	-1.704	-0.868
prui	-0.111	0.888	0.52	-0.254	1.047	1.116	1.573	1.573	0.254	0.423	-0.472	1.573	-1.704	-0.868
parv	-0.111	0.888	0.52	3.811	1.047	1.116	1.573	1.573	0.254	0.423	-0.472	1.573	-1.704	-0.868
scdi	-0.111	1.581	0.52	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	-2.287	2.048	1.573	-1.704	-0.868
meli	-0.111	1.581	0.52	-0.254	1.047	1.116	1.573	-0.615	0.254	0.423	-0.472	1.573	-1.704	-0.868
cham	-0.111	-0.498	-1.86	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	-2.287	2.048	-0.615	0.568	-0.868
lydi	-0.111	-0.498	-1.86	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	0.423	-0.472	-0.615	0.568	1.116
trap	-0.111	-0.498	0.52	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	0.423	-0.472	-0.615	0.568	-0.868
taur	-0.111	-0.498	-1.86	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	0.423	-0.472	-0.615	0.568	1.116
sysp	-0.111	-0.498	-1.86	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	0.423	-0.472	-0.615	0.568	1.116
sinu	-0.111	-0.498	0.52	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	0.423	-0.472	-0.615	0.568	1.116
diva	-0.111	-0.498	0.52	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	0.423	-0.472	-0.615	0.568	1.116
grae	-0.111	-0.498	0.52	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	0.423	-0.472	-0.615	0.568	1.116
flav	-0.111	-0.498	-1.86	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	0.423	-0.472	-0.615	0.568	1.116
leuc	-0.111	-0.498	0.52	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	-2.287	2.048	-0.615	0.568	1.116
micr	-0.111	-0.498	0.52	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	0.423	-0.472	-0.615	0.568	1.116
mont	-0.111	2.275	-1.86	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	0.423	-0.472	-0.615	0.568	-0.868
poli	-0.111	2.275	-1.86	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	-2.287	2.048	-0.615	0.568	-0.868
mntb	-0.111	0.195	0.52	-0.254	1.047	-0.868	-0.615	-0.615	0.254	0.423	-0.472	-0.615	0.568	1.116
yild	-0.111	0.195	0.52	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	0.423	2.048	-0.615	0.568	-0.868
stac	-0.111	0.888	0.52	-0.254	-0.924	-0.868	-0.615	-0.615	0.254	-2.287	2.048	-0.615	0.568	-0.868
hirc	-0.111	0.888	0.52	-0.254	1.047	1.116	-0.615	1.573	0.254	0.423	-0.472	1.573	-1.704	-0.868

Orijinal veri matrisi standartlaştırıldığında ch karakterleri Zch standartlaştırılmış karakter adlarını aldığı Tablo 4.8'de belirtilmiştir. Teucrium türüne ait orijinal veri matrisinden elde edilen bu standartlaştırılmış veri matrisi ile faktör yüklerinin çarpılmasıyla ortaya çıkan faktör skorları matrisini Tablo 4.9'da, bu matrisin kovaryans matrisi ise Tablo 4.10'da verilmiştir. Tablo 4.9'de görüldüğü gibi Teucrium türlerine ait değişimlerin büyük bir çoğunluğunu (yaklaşık $\%58$) 1. ve 2. faktörlerden kaynaklandığı tespit edilmiştir. Diğer yandan, $i \neq j$ olmak üzere, $\text{cov}(FS_i, FS_j) = 0$ olduğu yani, farklı indisli faktör skorlarının kovaryansının sıfır olduğu Tablo 4.10'da görülür.

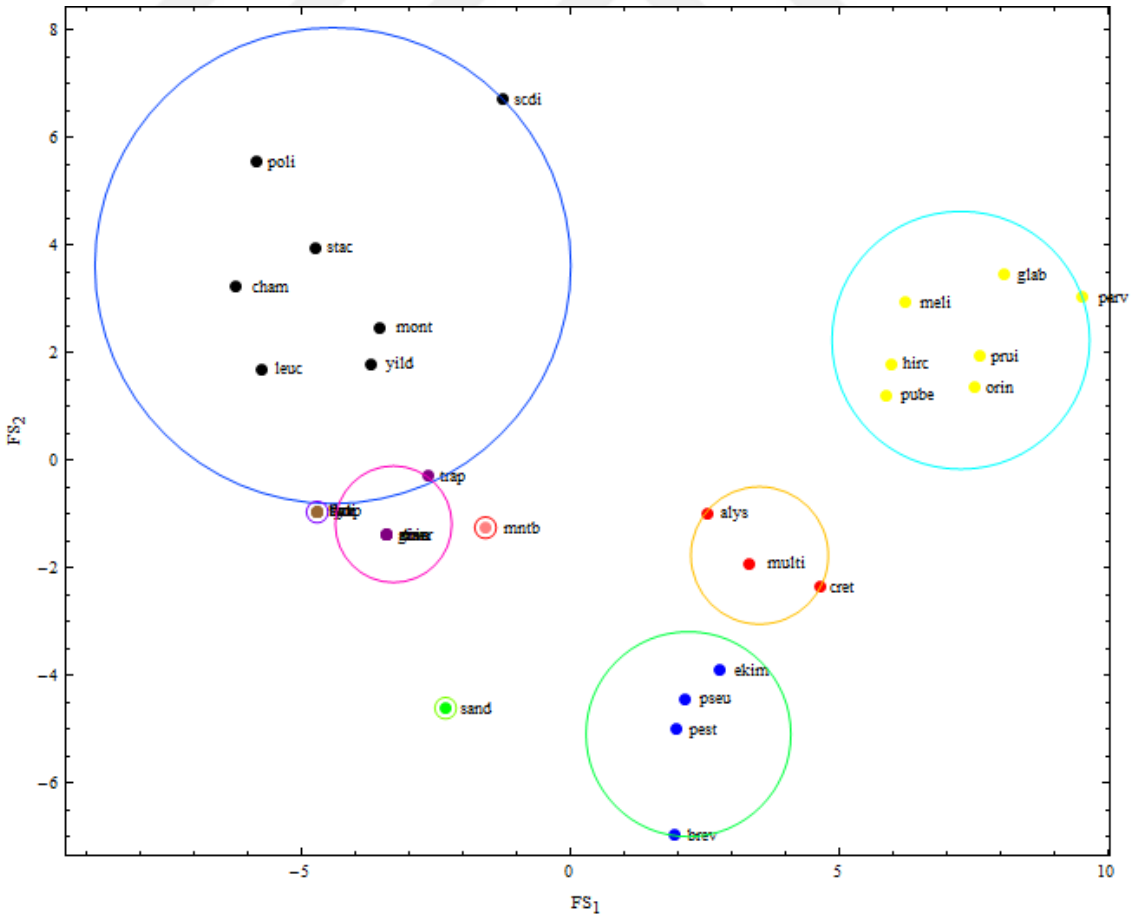
Tablo 4.9: Faktör skorları.

	FS_1	FS_2	FS_3	FS_4	FS_5	FS_6	FS_7	FS_8	FS_9	FS_{10}	FS_{11}	FS_{12}	FS_{13}	FS_{14}
cret	4.63599	-2.35429	0.959174	0.437663	1.50137	-0.847756	-1.1797	-0.401346	0.687282	0.141216	-0.0723416	0.00793949	-0.155966	-0.071236
sand	-2.31278	-4.60843	3.02871	-2.86318	0.056032	0.233243	-0.913061	0.164819	0.00506361	-0.174865	-0.0740322	-0.0579347	0.004233	0.0462495
brev	1.93594	-6.96556	2.9758	-1.27175	-0.945372	-0.0135752	0.898675	0.817168	-0.0587118	-0.0156095	0.119988	0.0660289	-0.00823059	-0.0498566
pest	1.97693	-5.00438	0.854447	1.16875	-0.398038	-0.180531	0.910669	-0.478155	0.0152252	0.397226	-0.0803289	0.0161145	0.0366103	0.00737225
ekim	2.76731	-3.90401	1.63744	0.775458	-0.408038	-0.605511	0.174262	-1.08323	-0.387542	0.133197	-0.00962551	-0.0651602	0.0358258	0.0310334
alys	2.55819	-0.981636	-0.456713	0.732706	1.20395	-0.52107	-0.844697	0.204112	-0.57707	-0.419837	0.267634	-0.0562577	0.0538484	-0.0247735
pseu	2.13697	-4.44023	0.112265	1.03077	1.70975	-0.618044	-0.165507	0.294809	0.318472	0.315469	-0.0803525	0.0402486	0.0968702	0.0510203
multi	3.33165	-1.91739	-0.894363	1.44622	-0.0215095	-0.202824	0.665093	0.817359	-0.650239	-0.269841	-0.152966	-0.0162234	-0.0383161	-0.0222148
orin	7.51523	1.35577	-0.0844388	0.556964	-0.732514	-0.0206603	-0.341236	0.188455	0.254427	-0.442951	-0.114005	-0.0110416	-0.0161864	0.0514085
pube	5.88818	1.20753	0.260529	0.305991	-1.74186	0.451492	-0.0692985	-0.426299	0.24167	-0.0543486	0.183444	0.00799658	0.0145703	-0.0289251
glab	8.05715	3.4514	-1.34006	-1.82252	1.93661	0.162366	0.757886	-0.185787	0.0232741	0.182425	0.151014	0.0204853	0.0442304	-0.0190782
prui	7.61207	1.9363	-0.136312	0.48823	-0.869535	-0.0927542	-0.372091	0.354833	0.0689803	-0.25014	-0.166476	-0.00910956	-0.0181189	0.0409429
parv	9.52417	3.03549	-0.94285	-1.43356	0.838168	0.127727	1.56212	-0.343998	-0.267216	-0.12442	-0.146411	-0.0226871	-0.0467861	0.0176074
scdi	-1.25842	6.71341	2.28488	0.0314595	-0.257291	0.856769	-0.199863	0.156439	0.692821	0.303142	-0.0942946	-0.0121124	0.0171307	0.0231504
meli	6.24189	2.93275	-0.5854	0.0305414	0.0918828	-0.130209	-1.20719	0.679422	0.174024	0.249517	0.0784773	0.0359949	0.0709651	-0.00620828
cham	-6.22761	3.22943	1.66267	0.319047	0.327888	-1.11394	0.503962	-0.325761	0.363114	-0.489951	0.173554	-0.0302763	-0.00323372	0.0264571
lydi	-4.71683	-0.959002	-1.9785	-0.348727	-0.409325	-0.566621	0.233893	-0.0179319	0.410379	-0.0930152	0.010923	0.0322825	0.00437349	0.00223407
trap	-2.64096	-0.289032	-0.444611	-0.327446	0.107565	0.588242	-0.84385	-0.616122	-0.413774	-0.233305	0.00293407	-0.123081	0.00683669	0.00402334
taur	-4.71683	-0.959002	-1.9785	-0.348727	-0.409325	-0.566621	0.233893	-0.0179319	0.410379	-0.0930152	0.010923	0.0322825	0.00437349	0.00223407
sysp	-4.71683	-0.959002	-1.9785	-0.348727	-0.409325	-0.566621	0.233893	-0.0179319	0.410379	-0.0930152	0.010923	0.0322825	0.00437349	0.00223407
sinu	-3.43134	-1.38941	-1.2276	0.0658497	0.117565	1.01322	-0.107444	-0.0110417	-0.0110074	0.0307239	-0.0677693	-0.0418062	0.00762123	-0.0196378
diva	-3.43134	-1.38941	-1.2276	0.0658497	0.117565	1.01322	-0.107444	-0.0110417	-0.0110074	0.0307239	-0.0677693	-0.0418062	0.00762123	-0.0196378
grae	-3.43134	-1.38941	-1.2276	0.0658497	0.117565	1.01322	-0.107444	-0.0110417	-0.0110074	0.0307239	-0.0677693	-0.0418062	0.00762123	-0.0196378
flav	-4.71683	-0.959002	-1.9785	-0.348727	-0.409325	-0.566621	0.233893	-0.0179319	0.410379	-0.0930152	0.010923	0.0322825	0.00437349	0.00223407
leuc	-5.7325	1.69865	1.63058	1.12692	0.864777	0.890887	0.899032	0.286209	0.344494	-0.102183	0.0241588	-0.0230904	0.000798554	-0.019076
micr	-3.43134	-1.38941	-1.2276	0.0658497	0.117565	1.01322	-0.107444	-0.0110417	-0.0110074	0.0307239	-0.0677693	-0.0418062	0.00762123	-0.0196378
mont	-3.53911	2.4635	-1.403	-1.01696	-0.96741	-1.27998	-0.625935	0.0425013	-0.734173	0.414202	-0.12826	-0.0412638	-0.00414102	-0.0159672
poli	-5.84027	5.55156	1.45518	0.0441107	-0.220197	-1.40231	0.380541	0.339752	-0.378671	0.281295	-0.0363315	-0.022548	-0.0109637	-0.0154054
mntb	-1.57936	-1.25191	-1.16264	0.429296	0.0256737	0.686149	-0.003945	0.305095	-0.339298	0.43574	0.328759	0.0103712	-0.135418	0.0961061
yild	-3.69452	1.78652	0.894645	0.10938	0.363288	0.544764	-0.492432	-0.428342	-0.606697	-0.184685	-0.0680468	0.414274	-0.0013247	0.00272158
stac	-4.74845	3.96009	2.30982	0.596156	0.580735	0.321719	0.100915	0.0138858	-0.429165	0.0194112	-0.0100808	-0.100501	-0.00385097	-0.016346
hirc	5.98501	1.78807	0.208656	0.237257	-1.87889	0.379399	-0.100154	-0.259921	0.0562237	0.138463	0.130972	0.00992865	0.0126378	-0.0393907

Tablo 4.10: Faktör Skorları Matrisinin Kovaryans Matrisi.

	FS_1	FS_2	FS_3	FS_4	FS_5	FS_6	FS_7	FS_8	FS_9	FS_{10}	FS_{11}	FS_{12}	FS_{13}	FS_{14}
FS_1	24.048	0	0	0	0	0	0	0	0	0	0	0	0	0
FS_2	0	9.906	0	0	0	0	0	0	0	0	0	0	0	0
FS_3	0	0	2.278	0	0	0	0	0	0	0	0	0	0	0
FS_4	0	0	0	0.816	0	0	0	0	0	0	0	0	0	0
FS_5	0	0	0	0	0.75	0	0	0	0	0	0	0	0	0
FS_6	0	0	0	0	0	0.492	0	0	0	0	0	0	0	0
FS_7	0	0	0	0	0	0	0.409	0	0	0	0	0	0	0
FS_8	0	0	0	0	0	0	0	0.163	0	0	0	0	0	0
FS_9	0	0	0	0	0	0	0	0	0.149	0	0	0	0	0
FS_{10}	0	0	0	0	0	0	0	0	0	0.06	0	0	0	0
FS_{11}	0	0	0	0	0	0	0	0	0	0	0.014	0	0	0
FS_{12}	0	0	0	0	0	0	0	0	0	0	0	0.007	0	0
FS_{13}	0	0	0	0	0	0	0	0	0	0	0	0	0.002	0
FS_{14}	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001

Dolayısıyla, faktörlerin birbirlerinden bağımsız olduğu söylenmektedir. Teucrium türlerine ait toplam değişimin büyük miktarını ilk iki faktör temsil ettiğinden dolayı, türlerin bu faktörlere göre saçılım grafiği ve sınıflandırılması Şekil 4.5 ile verilmiştir.



Şekil 4.5: Türlerin faktör skorlarına göre saçılımı ve sınıflandırılması.

Şekil 4.5 incelendiğinde, Teucrium türlerinin 8 kümeye ayrıldığı kolaylıkla görülmektedir. Bu sınıflandırmalara göre poli, scdi, stac, cham, mont, yild, leuc türlerinin, meli, glab, parv, hirc, prui, pube, orin türlerinin, ekim, pseu, pest, brev türlerinin alys, multi, cret türlerinin, lydi, taur, sysp, flav türlerinin, sinu, diva, grae, micr türlerinin birer küme oluşturdukları görülmektedir. Öte yandan, mntb ve sand türleri tek başlarına birer küme oluşturmuşlardır. Faktör analizine göre sınıflandırma, kümeleme analizinin hiyerarşik kümeleme tekniği olan tek bağlantı, tam bağlantı ve Ward kümeleme tekniklerine göre aynı küme sayısına sahip oldukları bilgisine ulaşılmıştır. Öte yandan, faktör analizi ile tek bağlantı kümeleme tekniğinde çıkan kümelerin neredeyse hemen hemen birbirlerine benzer olduğu kanısına varılmıştır. Diğer yandan, faktör analizi ile oluşturulan kümelerin Tam bağlantı ve Ward kümeleme teknikleriyle oluşturulan kümeler ile bazı farklılıklar gösterebileceği saptanmıştır.

5. TARTIŞMA VE SONUÇ

Literatürde sıkça kullanılan çok değişkenli veri analizi yöntemlerinden kümeleme analizi ve faktör analizi, karmaşık bir şekilde bulunan veri tabanlarını basit hale getirerek anlaşılır bir yapı oluşturmaktadır. Oluşan bu yapılar kümeleme analizi için *dendrogram* adı verilen ağaç diyagramları ile, faktör analizi için ise *saçılım grafikleri* ile yorumlanabilmektedir. Bu tezde, Türkiye'nin en zengin üçüncü bitki familyası içerisinde yer alan Lamiaceae ailesinin bir üyesi olan ve çoğunluğu Akdeniz bölgesinde yetişen Teucrium cinsine ait 32 adet türün 14 karakteristik özelliğini içeren bir veri tabanı ile çalışma gerçekleştirilmiştir. Bu uygulamanın amacı, 32 adet Teucrium türünün kümeleme analizi ve faktör analizi vasıtasıyla sınıflandırıp, karşılaştırmasını yapmaktır.

Bu veri tabanına literatürde sıkça kullanılan ve tercih edilen öklidyen mesafesi uygulanarak, kümeleme analizinin hiyerarşik kümeleme tekniği olan tek bağlantı metodu, tam bağlantı metodu ve Ward kümeleme metodu uygulanmıştır. Bu yöntemlerin dendrogramları detaylı olarak incelendiğinde, 32 tane türün tek bağlantı metoduna göre $\{lydi, taur, sysp, flav, sinu, diva, grae, micr, cham, leuc, mntb, stac, yild, trap\}$, $\{orin, pube, prui, glab, parv, hirc, meli\}$, $\{brev, pest, ekim, cret\}$, $\{mont, poli\}$, $\{scdi\}$, $\{multi\}$, $\{alys\}$, $\{sand\}$ şeklinde 8 kümeye, tam bağlantı metoduna göre $\{orin, pube, prui, hirc, glab, parv, meli\}$, $\{scdi, stac, yild\}$, $\{mont, poli\}$, $\{multi, alys\}$, $\{cham, leuc\}$, $\{lydi, taur, sysp, flav, sinu, diva, grae, micr, trap, mntb\}$, $\{brev, pest, ekim, pseu, cret, sand\}$ şeklinde 7 kümeye, Ward bağlantı metoduna göre $\{scdi, mont, poli\}$, $\{mntb, alys, multi\}$, $\{orin, pube, prui, hirc\}$, $\{glab, parv, meli\}$, $\{sand, cret, pseu, ekim, brev, pest\}$, $\{cham, leuc, stac, yild\}$, $\{lydi, taur, sysp, flav\}$, $\{trap, sinu, diva, grae, micr\}$ şeklinde 8 kümeye ayrıldığı söylenebilmektedir. Ancak, her ne kadar küme sayıları yaklaşık olarak birbirine yakın olsa da, her bir metoda göre oluşan kümelerin birbirine benzemediği görülmüştür. Bu durum akıllara hangi kümeleme tekniğinin daha kullanışlı olduğu ile ilgili bir soru getirmektedir. Bu sorunun cevabını kısmen belirlemek için, Teucrium türlerine ait veri tabanına başka bir istatistiksel analiz yöntemi olan faktör analizinin faktör türetme yöntemi olan temel bileşenler analizi uygulanmıştır. Bu doğrultuda, kümeleme analizi ile temel bileşenler analizinin sonuçları karşılaştırılmıştır. Böylelikle, hem bu sorunun

cevabı kısmen netleştirilecek hem de analizler doğrultusunda çıkan kümeleme sonuçları iki boyutlu düzlem üzerinde görsel olarak gösterilecektir.

Bu veri tabanına öncelikle faktör analizinin uygulanabilirliğini araştıran Kaiser -Meyer-Olkin örneklem yeterlilik ölçütü uygulanmıştır. Bu hesaplama sonucunda KMO değeri 0.7 olarak hesaplanmış ve böylelikle bu veri tabanına faktör analizinin uygulanabileceği bilgisine ulaşılmıştır. Daha sonra, orijinal veri matrisinin korelasyon matrisi bulunarak, korelasyon matrisinin özdeğer-özvektör matrisi oluşturulmuştur. Korelasyon matrisi incelendiğinde, $ch5, ch6, ch7, ch8, ch10, ch11, ch13$ karakterlerinin toplam varyansın büyük bir kısmını açıklayan faktörler civarında yer alacağı öngörülmüş ama henüz kesin bir sonuca varılamamıştır. Ardından, özdeğerler ve özvektörler kullanılarak gerekli cebirsel işlemler yapılması sonucunda, faktör yükleri matrisi elde edilmiştir. Faktör yükleri matrisi gözlemlendiğinde, oluşan Faktör1 toplam varyansın $\%35.021$'ini, Faktör2 toplam varyansın $\%22.478$'ini, Faktör3 toplam varyansın $\%10.778$'ini, Faktör4 toplam varyansın $\%6.45$'ini açıkladığı görülmüştür. Ardından, elde edilen faktörlerden kaç tanesi ile çalışılması gerektiğini belirlemek için yamaç grafiği ile özdeğer ölçütü olan 1 değeri aynı anda gözlemlenmiş ve bu inceleme sonucunda toplam varyansın $\%68.27775$'ini açıklayan ilk üç faktörle çalışılması gerektiğine karar verilmiştir. Ancak, bu çalışma sonucunda ortaya çıkmış olan kümeleme sonuçlarını iki boyutlu düzlem üzerinde gösterebilmek amacıyla, toplam varyansın $\%57.499$'sını açıklayan ilk iki faktör ele alınarak çalışma yapılmıştır. Ele alınan ilk iki faktörün açıkladığı en az varyans miktarı $\%0.232735$ ile ch9 karakterinin varyansı, en fazla ise $\%0.878831$ ile ch6 karakterinin varyansı olduğu görülmüştür. Elde edilen bu bilgiler doğrultusunda faktör skoru tablosu oluşturulmuş ve ilk iki faktöre göre, Teucrium cinsine ait 32 türün $\{$poli, scdi, stac, cham, mont, yild, leuc$\}$, $\{$meli, glab, parv, hirc, prui, pube, orin$\}$, $\{$ekim, pseu, pest, brev$\}$, $\{$alys, multi, cret$\}$, $\{$lydi, taur, sysp, flav$\}$, $\{$sinu, diva, grae, micr$\}$, $\{$mntb$\}$, $\{$sand$\}$ şeklinde 8 kümeye ayrıldığı görülmüştür. Diğer yandan, bu tezin teori ve uygulama kısımlarında faktör döndürme yöntemlerine değinilmemiştir. Faktör döndürme yöntemleri, üretilen faktörlerden bilgi elde edilmesi zor olduğu durumlarda kullanılmaktadır. Böylelikle, faktör döndürme yöntemlerinden herhangi biri kullanılarak, orijinal faktör matrisi daha anlaşılır ve basit hale gelmiş olur. Bu uygulamada kullanılan orijinal verilerden elde edilen faktör yükleri

sonucunda elde edilen bilgilerin yeterli olmasından dolayı, faktör döndürme yöntemlerinin kullanılmasına ihtiyaç duyulmamıştır.

Bu uygulamada faktör analizi ile kümeleme analizi karşılaştırıldığında, kümeleme analizinin tek bağlantı metodunda ortaya çıkan kümeler ile faktör analizinde ortaya çıkan kümelerin aşağı yukarı benzer oldukları görülmüştür. Bundan dolayı, bu çalışmada kümeleme analizi olarak tek bağlantı metodunun amacımızı karşıladığı kısmen söylenebilmektedir. Bu çalışma ile poli, scdi, stac, cham, mont, yild, leuc türleri arasında, meli, glab, parv, hirc, prui, pube, orin türlerinin, ekim, pseu, pest, brev türleri arasında, alys, multi, cret türleri arasında, lydi, taur, sysp, flav türleri arasında, sinu, diva, grae, micr türleri arasında bir akrabalık ilişkisi olduğu söylenebilir. Geriye kalan mntb ve sand türlerinin ise, tek elemanlı kümeler oluşturduğu ve Teucrium türlerine ait veri tabanındaki diğer türlerle herhangi bir akrabalık ilişkisinin olmadığı bilgisine ulaşılmıştır. Elde edilen bu bilgilerin ileri ki dönemlerde biyoloji, eczacılık gibi laboratuvar ortamlarında yapılacak çalışmalara katkı sağlaması beklenmektedir. Ayrıca, bu tezde kümelendirme, faktör analizi ve temel bileşenler analizi konuları detaylı incelenerek, bu alanlarda karşılaştıkları problemlerin çözümünü isteyen araştırmacılar için yapılan tez çalışmasının bir kaynak olması düşünülmüştür.

KAYNAKLAR

[1]. Anderberg, M.R., 1973, *Cluster Analysis for Applications*, Academic Press, New York.

[2]. Günay Atbaş, A.C., 2008, *Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma*, Yüksek Lisans, Ankara Üniversitesi Fen Bilimleri Enstitüsü.

[3]. Ray, J., 1704, *Historia Plantarum Species Hactenus Editas Aliasque Insuper Multas Noviter Inventas and Descriptas Complectens*, Mariae Clark, London, UK.

[4]. Williams, R.H., Zimmerman, D.W., Zumbo, B.D., Ross, D., 2003, Charles Spearman: British Behavioral Scientist, *Human Nature Review*, 3, 114-118.

[5]. Child, D., 1975, *The Essentials of Factor Analysis*, Billing and Sons Ltd., London.

[6]. Koç, N., Ülkü, S., 1979, Faktör Analizi-Yetenekleri Sınıflama (Çeviri), *Ankara Üniversitesi Eğitim Fakültesi Dergisi*, 10(1-2), 24-32.

[7]. Kline, P., 1994, *An Easy Guide to Factor Analysis*, Routledge, London-USA.

[8]. Sangün, L., 2007, *Temel Bileşenler Analizi, Ayırma Analizi, Kümeleme Analizleri ve Ekolojik Verilere Uygulanması Üzerine Bir Araştırma*, Doktora, Çukurova Üniversitesi Fen Bilimleri Enstitüsü.

[9]. Tekin, M., 2004, 1987-1996 Yılları Arasında İllerin GSYH'ya Katkısının Faktörlere Ayrılması ve Bulunan Faktörler Açısından Sıralanması(Faktör Analizi Uygulaması), *İstanbul Üniversitesi İktisat Fakültesi İktisat Mecmuası*, 54(1), 167-194.

[10]. Karabayır, M.E., Doğanay, M., 2010, Kümeleme Analizi İle Portföy Seçimi: İmkb-100 Endeksi Üzerine Bir Çalışma , *Gazi Üniversitesi Ticaret ve Turizm Eğitim Fakültesi Dergisi*, 2, 160-179.

[11]. Akın, H.B., Eren, Ö., 2012, OECD Ülkelerinin Eğitim Göstergelerinin Kümeleme Analizi ve Çok Boyutlu Ölçekleme Analizi İle Karşılaştırmalı Analizi, *Marmara Üniversitesi Öneri Dergisi*, 10(37), 175-181.

[12]. Sanguansat, P., 2012, *Principal Component Analysis-Multidisciplinary Applications*, InTech, Croatia, ISBN:978-953-51-0129-1.

[13]. Çelik, Ş., 2012, Türkiyede İllerin Bitkisel Üretiminin Faktör Analizi ile İncelenmesi, *Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi*, 22(2), 69-76.

[14]. Kaya, M.F., 2013, Sürdürülebilir Kalkınmaya Yönelik Tutum Ölçeği Geliştirme Çalışması, *Marmara Coğrafya Dergisi*, 22, 175-193.

[15]. Yıldırım, B., 2015, Fen Bilimleri Öğrenme Kaygı Ölçeği: Geçerlilik ve Güvenirlik Çalışması, *Muş Alparslan Üniversitesi Sosyal Bilimler Dergisi*, 3(1), 33-43.

[16]. Ross, G.S., Foran, L.M., Barbot, B., Sossin, K.M., Perlman, J.M., 2016, Using Cluster Analysis to Provide New Insights into Development of Very Low Birthweight (VLBW) Premature Infants, *Early Human Development*, 92, 45-49.

[17]. Bandyopadyay S., Saha, S., 2013, *Unsupervised Classification-Similarity Measures, Classical and Metaheuristic Approaches and Applications*, Springer, Berlin ISBN: 978-3-642-32451-2.

[18]. Krebs, C.J., 2014, *Ecological Methodology*, Addison-Wesley Educational Publishers, Inc.

[19]. Ergüt, Ö., 2011, *Uzaklık ve Benzerlik Ölçülerinin Kümeleme Sonuçlarına Etkisi*, Yüksek Lisans, Marmarma Üniversitesi Sosyal Bilimler Enstitüsü.

[20]. Hasan, N., Adam, M.B., Mustapha, N., Abu Bakar, M.R., 2012, Similarity Measure Exercise for Classification Trees Based on the Classification Path, *Applied Mathematics and Computational Intelligence*, 1, 33-41.

[21]. Teknomo, K., 2015, Similarity Measurements - Kendall Distance, http://people.revoledu.com/kardi/tutorial/Similarity/KendallDistance.html, [Ziyaret Tarihi:29.03.2016].

[22]. Teknomo, K., 2015, Similarity Measurements - Cayley Distance, http://people.revoledu.com/kardi/tutorial/Similarity/CayleyDistance.html, [Ziyaret Tarihi:29.03.2016].

[23]. Gan, G., Ma, C., Wu, J., 2007, *Data Clustering Theory Algorithms, and Applications*, ASA-SIAM, Philadelphia, USA, ISBN: 978-0-898716-23-8.

[24]. Albayrak, A.S., 2006, *Uygulamalı Çok Değişkenli İstatistik Teknikleri*, Asil Yayın Dağıtım, Ankara.

[25]. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.L., 2006, *Multivariate Data Analysis*, Pearson Prentice Hall, USA.

[26]. Alpar, R., 2013, *Çok Değişkenli İstatistiksel Yöntemler*, Detay Yayıncılık, Ankara, ISBN:978-605-5437-42-8.

[27]. Genç, I., Özhatay, N., Cevri, M., 2013, A karyomorphological study of the genus Allium (sect. Melanocrommyum) from Turkey, *International Journal of Cytology, Cytosystematics and Cancer Genetics And Cytogenetics*, 1, 31-40.

[28]. Kenett D.Y., Huang, X., Vodenska, I., Havlin, S., Stanley, H.E., 2015, Partial Correlation Analysis: Applications for Financial Markets, *Quantitative Finance*, 15(4), 569-578.

[29]. Tatlıdil, H., 2002, *Uygulamalı Çok Değişkenli İstatistiksel Analiz*, Akademi Matbaası, Ankara.

[30]. Atan, M., Göksel, A., Karpat, G., 2002, Üniversite öğrencilerinin başarılarını etkileyen faktörlerin çok değişkenli istatistiksel analiz yöntemleri ile tespiti, *XI. Eğitim Bilimleri Kongresi*,23-26 Ekim 2002 Yakın Doğu Üniversitesi, 23-26.

[31]. Ersungur Ş.M., Kızıltan, A., Polat, Ö., 2007, Türkiye'de Bölgelerin Sosyo-Ekonomik Gelişmişlik Sıralaması: Temel Bileşenler Analizi, *İktisadi ve İdari Bilimler Dergisi*, 21(27), 55-66.

[32]. Polat, Y., 2012, *Faktör Analizi Yöntemlerinin Karşılaştırılmalı Olarak İncelenmesi ve Hayvancılık Denemesine Uygulanışı*, Doktora, Çukurova Üniversitesi Fen Bilimleri Enstitüsü.

[33]. Büyüköztürk, Ş., 2002, Faktör Analizi: Temel Kavramlar ve Ölçek Geliştirmede Kullanımı, *Kuram ve Uygulamada Eğitim Blimleri*, Sayı :32, 470-483.

[34]. Khalaf, K., 2007, *Faktör Analizi ve Bir Uygulaması*, Yüksek Lisans, Gazi Üniversitesi Fen Bilimleri Enstitüsü.

[35]. Özdamar, K., 2004, *Paket Programlar İle İstatistiksel Veri Analizi-1*, Kaan Kitabevi, Eskişehir.

[36]. Emin, S.M., 1984, *Çok Boyutlu Verilerin Bazı İstatistiksel Analiz Yöntemleri ve Uygulamaları*, Doktora, Gazi Üniversitesi Fen Bilimleri Enstitüsü.

[37]. Alkin, E., Yıldırım, K., Özer M., 2003, *İktisata Giriş*, Anadolu Üniversitesi, Eskişehir.

[38]. Jain, A.K., Murty, M.N., Flynn, P.J., 1999, Data clustering: A review, *ACM Computing Survey*, 31(3), 264–323.

[39]. Kaya, H., Köymen, K., 2008, Veri Madenciliği Kavramı ve Uygulama Alanları, *Doğu Bölgesi Araştırmaları*.

[40]. Özdamar, K., 2010, *Paket Programlar ile İstatistiksel Veri Analizi- 2 (Çok Değişkenli Analizler)*, Kaan Kitabevi, Eskişehir.

[41]. Sarıman, G., 2011, ,Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması, *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 15(3), 192-202.

[42]. Koldere Akın, Y., 2008, *Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi*, Doktora, Marmara Üniversitesi Sosyal Bilimler Enstitüsü.

[43]. Vattani, A., 2011, K-Means Requires Exponential Many Iterations Even in The Plane, *Discrete Computer Genom* Cilt: 45, 596-616.

[44]. Han, J., Kamber, M., 2000, *Data Mining : Concepts and Techniques*, Morgan Kaufmann Publishers, Kanada.

[45]. Ketchen, D.J., Shook, C.L., 1996, *The Application of Cluster Analysis in Strategic Management Reserch: An Analysis and Critique*, Strategic Management Journal, 17(6), 441-458.

[46]. Koltan Yılmaz, Ş., Patır S., 2011, Kümeleme Analizi ve Pazarlamada Kullanımı, *Akademik Yaklaşımlar Dergisi*, 2(1), 91-113.

[47]. Everitt, B., 1993, *Cluster analysis, (3rd edition)*, Halsted Press, New York, USA.

[48]. Çelik, H.,C., 2004, *Çok Değişkenli İstatistiksel Yöntemlerden Kümeleme Yöntemi ve Kronik Sigara İçiciler Üzerine Bir Çalışma*, Doktora, Dicle Üniversitesi Sağlık Bilimleri Enstitüsü.

[49]. Doğan, M., Yüce, E., Doğan, G., Bağcı, E., 2008, Teucrium polium L. (Lamiaceae) Türünün Morfolojik Varyasyonu Üzerine Bir Araştırma, *Fırat Üniversitesi Fen ve Mühendislik Bilimleri Dergisi*, 20(3), 389-402.

[50]. Vural, M., Duman, H., Dirmenci, T., Özcan, T., 2015, A new species of Teucrium sect. Stachyobotrys (Lamiaceae) from the south of Turkey, *Turkish Journal of Botany*, 39, 318-324.

[51]. Özcan, T., Dirmenci, T., Coşkun, F., Akçiçek, E., Güner, Ö., 2015, A New Species of Teucrium Sect. Scordium (Lamiaceae) from SE of Turkey, *Turkish Journal of Botany*, 39, 310-317.

[52]. Stankovic, M.S., Curcic, M.G., Zizic, J.B., Topuzovic, M.D., Solujic, S.R., Markovic, S.D., 2011, Teucrium Plant Species as Natural Sources of Novel Anticancer Compounds: Antiproliferative, Proapoptotic and Antioxidant Properties, *International Journal of Molecular Sciences*, 12, 4190-4205.

[53]. Ecevit Genç, G., Özcan, T., Dirmenci, T., 2015, Micromorphological Characters on Nutlet and Leaf Indumentum of Teucrium Sect. Teucrium (Lamiaceae) in Turkey, *Turkish Journal of Botany*, 39, 439-448.

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı	Fatih Fırat
Uyruğu	T.C.
Doğum yılı, Yeri	1991, Darende
Telefon	05418381591
E-mail	fatihfirat@hotmail.com

Eğitim

Derece	Kurum/Anabilim Dalı/Programı	Yılı
Lisans	İ.Ü. Fen Fakültesi/Matematik Bölümü	2013
Lise	Kadri Yörükoğlu Lisesi	2009