TIBBİ VERİ ANALİZİNDE KÜME YAKLAŞIMLARI

Dr. Öğr. Üyesi Çağlar CENGİZLER

yaz
yayınları

Tıbbi Veri Analizinde Küme Yaklaşımları

Temel Kavramlar ve Uygulamalar

Dr. Çağlar Cengizler

yaz yayınları

2025

yaz
yayınları

Tıbbi Veri Analizinde Küme Yaklaşımları

Yazar: Dr. Öğr. Üyesi Çağlar CENGİZLER

ORCID No: 0000-0002-6699-5683

E_ISBN 978-625-8508-71-0

Ekim 2025 – Afyonkarahisar

Dizgi/Mizanpaj: YAZ Yayınları

Kapak Tasarım: YAZ Yayınları

YAZ Yayınları. Yayıncı Sertifika No: 73086

M.İhtisas OSB Mah. 4A Cad. No:3/3
İscehisar/AFYONKARAHİSAR

www.yazyayinlari.com

yazyayinlari@gmail.com

info@yazyayinlari.com

Desteğini asla esirgememiş ailem ve Sonat için.

Önsöz

Duyularımız çevremizi saran hayata ve bu hayat içinde olup biten olaylara dair ilk bilgi kaynağımız. Dış dünyayı anlamlandıracak yeni aletler geliştirmek ise insan olmanın önemli ve güzel gereklerinden birisi. Bir dürbün normalde göremeyeceğimiz mesafeden fırtınanın yaklaşmakta olduğu bilgisini bize kazandırarak geleceğe dair kestirim yapmamızı sağlayabilir. Bir steteskop ise hastanın kalp ritmine dair bilgi sağlayarak onu tedavi etmemizi kolaylaştırabilir. Değişen dünya ve gelişen teknoloji duyularımızın çok ötesinde bize bilgi sağlamaya başladığından beri dünyayı ve yaşamı anlamlandırmak, yorumlayarak katkıda bulunmak pek çok farklı boyut kazandı. Gökyüzüne baktığımızda gördüğümüz yıldızların ötesinde kozmik aktörlerin artık birbirleri ile olan ilişkilerini, onları etkileyen fiziksel güçlerin ve olayların etkilerini bilebiliyoruz. Aynı şekilde artık hastaların iç yapılarına, demografilerine ve onları karakterize eden özniteliklerine bakmakla kalmıyoruz. İlişkiler kurarak örüntüleri öğreniyor, bu örüntüleri yorumlayarak kestirimlerde bulunuyoruz. Bu araştırmalarımızı olduğu kadar yaşantımızı da aydınlatma becerisi kazandırarak bizi olduğumuzdan daha ileri götürüyor. Bu kitapta teknolojik imkanların ve tıp pratiğinin kazandırdığı veriyi daha iyi tanımaya, araştırmaya ve anlamlandırmaya yarayabilecek temel bilgilerin okuyucuya kazandırılması hedeflenmiştir. Küme ve kümelenme analizleri perspektifinden veri madenciliğine nazik bir giriş yapılması ve okuyucunun daha önce sahip olmadığı bir perspektifi kazanarak kestirim yapabilmeye başlaması hedeflenmiştir. Kitabın daha önce keşfedilmemiş her köşeye bakmayı şiar edinmiş bütün araştırmacılar ve dünyayı daha anlamlı kılmayı uman tüm okuyucular için yeni perspektifler aydınlatacak bir ışık kaynağı olmasını diliyorum.

İçindekiler

Birlikte sayısız veri işlediğimiz, kıymetli destekleriyle katkılarını esirgemeyen hocalarım Prof. Dr. Memduha Gülhal Bozkır ve Arş. Gör. Dr. Ayşe Gül Kabakcı'ya teşekkür ederim.

Engin tecrübeleri ve yönlendirmeleriyle güzel çalışmalara imza attığımız Prof. Dr. Mehmet Satar hocama, ayrıca kıymetli hekim dostlarım Dr. Şerif Hamitoğlu, Dr. Mustafa Özdemir ve Dr. Tugay Tepe'ye ilgileri, destekleri ve yardımları için gönülden teşekkür ederim.

Beni sınırsız yeni proje ve veriyle buluşturan yenilikçi hocam Prof. Dr. Deniz Bolat'a ve radyolojik veri analizlerimizi mümkün kılan kıymetli uzman hekim arkadaşım Dr. Ebru Hasbay'a teşekkür ederim.

Kendimi bir parçası olarak görmekten büyük gurur duyduğum HİTES Ar-Ge ekibine ve çok şey öğrendiğim Levent Kasımay'a yürekten teşekkür ederim.

Kıymetli dostum, değerli akademisyen Dr. Çağatay Cebeci'ye eşsiz desteği için teşekkür ederim.

Kitabın yazım sürecinde ve tüm eğitim hayatım boyunca sonsuz desteğiyle her zaman yanımda olan Prof. Dr. İbrahim Cengizler'e özel olarak teşekkür etmek isterim.

Son olarak, başta Prof. Dr. M. Kerem Ün ve Prof. Dr. Selim Büyükkurt hocalarım olmak üzere, eğitimime katkıda bulunarak beni sabırla işleyip veri analizini ve tıp uygulamalarını kavramamı sağlayan tüm değerli akademisyen hocalarıma müteşekkirim.

Bölüm 1

Giriş

Hekim pratiği tıp biliminin kendini göstermesinden bu yana insanlığın gelişimini ve teknolojik ilerleyişini takip ederek kendini güncelleyen dinamik bir koleksiyondur. Farklı bilim dallarında su yüzüne çıkan her yeni buluntu, icat veya bilgi kırıntısı hekim pratiğini değiştirmeye aday olabilir. Buna göre bilgisayar teknolojisinin seyri uzun yıllardır hem günlük pratiğe hem de tıp camiasının kolektif vizyonuna katkılarda bulunabilmiştir. Bu katkı günümüzde enstrümantasyondan eğitim tekniklerine, mikro algıdan genetik çözümlemeye kadar çok geniş bir spektrumda kendini göstermektedir. Bilgisayarların ve bilgisayar destekli tıp teknolojilerinin gelişimi ile veri kavramı insan için en önemli mesleklerden birisi olan hekimliğin jargonu içinde kendine has bir yer edinmiştir. Tıbbi veri kaynaklarının sayı ve çeşitliliğindeki artış elbette yıllar içinde bilim insanlarının bu değerli cevhere bakışını değiştirmeye başlamıştır. Özellikle işlem gücü artan ve yaygınlaşan bilgisayarlar insan ölçeğinde işlenerek yorumlanması çok zor olan büyüklüklerde verileri anlamlandırmaya başladıkça eldeki cevherin de değeri artmaya başlamıştır. Tıbbi veri onu işleyebilecek nitelikteki bir araçla insan gözü için görünmez olan değerli örüntüleri bilgiye dönüştürme potansiyeline sahiptir. Bu hastalıkların nedenlerinin araştırılmasında, ilaçların etkenliğinin artırılmasında hatta kişiselleştirilmiş yaklaşımlar ile tedavi süreçlerinin optimize edilmesinde etkili olabilecek bir bilgi kaynağına sahip olmak demektir. İnsana dair olan, insan sağlığını niteleyebilen veya tanımlayabilen

sayısal veri aynı zamanda çoğu zaman insan algısına şaşırtıcı derecede yabancıdır. Modern teknolojinin bir hediyesi olarak zevkle dinlediğimiz sayısallaştırılmış müzik kayıtları ses duyumuza hitap edecek şekilde yeniden oluşturulmak yerine yazılı veri olarak parşömenlere bastırılsaydı kimsenin ne olduğunu anlamayacağı sayfalar dolusu işlevsiz veri yığınlarına dönüşürlerdi. Bu noktada şu soruyu sormak modern mühendislerin önemli sorumluluklarından birisidir: Tıbbi veri insan algı aralığı dışında kalan örüntüler ve ilişkiler taşıyor ise bu gömülü bilginin su yüzüne çıkarılması mümkün olabilir mi? Elbette bu soru uzun yıllardır pek çok veri analizi görevi sırasında sorulmuş ve bilim insanları tarafından onlarca farklı analiz tekniği geliştirilerek cevaplanmıştır. Günümüzde büyük veri (Big Data) ve veri madenciliği (Data Mining) kavramları yaygınlaşmış, gelişen teknoloji hem tıbbi verilerin toplanmasında hem de bu veri yığınlarından anlamlı örüntülerin çıkarılmasında kullanılmaya başlanmıştır. İstatistik, bu noktada en temel araçlardan biridir. İstatistiksel yöntemler veriler arasındaki ilişkilerin farklı açılardan incelenmesine ve örüntülerin yakalanmasına olanak tanımaktadır. Zaman içinde gelişen makine öğrenmesi yaklaşımları ise çok boyutlu verilerin analizinde önemli ilerlemeler sağlamıştır. Bu bağlamda, kümeleme yöntemleri veri setlerini oluşturan gözlemler arasındaki benzerlikleri incelemek ve doğal grupları keşfetmek için kullanılabilmekte, böylece verilerden sınıflandırma, öngörü ve farklı analiz görevleri için yararlı yapılar ortaya çıkarılabilmektedir.

1.1 Küme Yaklaşımlarının Tıpta Kullanım Alanları

Küme yaklaşımları veya kümeleme yöntemleri sayısal veri yığını içinde ilişkili gözlemlerin bilgisayar destekli analizi için güçlü araçlar olarak gösterilebilir. Bu bağlamda küme yaklaşımları tanı ve sınıflandırma görevlerinde sıklıkla kullanılmıştır. Tıbbi uygulama örneklerinin başında hasta gruplarının oluşturulması ve hastalık etiyolojisinin belirlenmesi gelmektedir. 1970'li yıllarda yapılan erken dönem çalışmalarından birisinde, psikiyatrik hastaların sınıf-

landırılmasında kümeleme tekniklerinin kullanılabileceği gösterilmiştir (Strauss vd., 1973). Örnek çalışmalardan birisinde ise bir hastaneye gelen hasta profilleri kümelenerek hastalık eğilim örüntüleri araştırılmış bu sayede hastanenin hizmet önceliklerini önceden planlanabilmesi hedeflenmiştir (Silitonga, 2017). Bir başka yakın dönem çalışmasında, görece olarak daha karmaşık hastalık profillerine sahip hasta gruplarının kümelenerek klinik açıdan anlamlı profiller elde edilebileceği ve bu yaklaşımın kaynak yönetimi ile tedavi planlama sürecinin optimizasyonu için önemli olabileceği ortaya konmuştur (Grant vd., 2020). Benzer bir çalışmada hastalar farklı kümeleme yaklaşımları ile hastanede kalış sürelerine göre gruplandırılmış ve klinik açıdan daha anlamlı kümeler üretilebileceği ortaya konmuştur (El-Darzi vd., 2009). Elektronik tıbbi kayıtlar üzerinde yapılan yine bir yakın dönem çalışmasında, veriler klinik açıdan anlamlı topluluklar halinde kümelenmiş ve bu sayede sınıflayıcı başarısının artırılabileceği gösterilmiştir (Huang vd., 2019).

Kümeleme yaklaşımları ile örüntü gruplama şüphesiz tıbbi verinin pek çok tipinde yaygın şekilde uygulanmıştır. Önemli veri tiplerinden birisi de sayısal görüntülerdir. İnsan gözü için kolaylıkla ayırt edilebilen veya edilemeyen klinik açıdan anlamlı olabilecek herhangi örüntünün sayısal görüntüler üzerinde sınıflandırılması tıbbi pratiğe ciddi katkı sağlama potansiyeline sahip kritik bir görevdir. Bu noktada kümeleme yaklaşımları da makine öğrenimi algoritmaları ile birlikte etkin şekilde uygulanmıştır. Erken dönem çalışmalardan birisinde, manyetik rezonans görüntü dilimleri üzerinde denetimsiz kümeleme yöntemleri uygulanarak beyin bölgelerinin otomatik olarak bölütlenmesi hedeflenmiştir (Clark vd., 2002). Bir başka çalışmada ise ultrason görüntülerinde kümeleme yöntemleri kullanılarak lezyon sınırlarının otomatik olarak belirlenmesi amaçlanmıştır (Shan vd., 2012). Literatürde farklı radyolojij görüntüleme yöntemleri ile elde edilmiş görüntülerin kümelenerek analiz edilmesine dair pek çok çalışma bulunmaktadır. Alternatif araştırmalardan birisinde dinamik pozitron emisyon tomografisi görüntülerinde doku bölütleme işlemi için kümeleme yaklaşımı uygulanmış ve yaklaşımın aynı zamanda girişimsel olmayan

görüntü tabanlı modelleme için bir önişleme basamağı olarak uygulanabileceği vurgulanmıştır (Wong vd., 2002). X-ray görüntüleri üzerinde yapılan bir araştırmada ise spektral kümeleme gürültü temizleme için bir önadım olarak uygulanmış sonrasında kemik bölgesinin bölütlenmesi gerçekleştirilmiştir (Wu ve Mahfouz, 2016). X-ray üzerine yapılmış bir başka çalışmada ise çene görüntüleri üzerinde kümeleme yaklaşımı uygulanarak dişle ilgili yapılar bölütlenmiştir (Tuan vd., 2017).

Küme yaklaşımlarının uygulama alanları arasında biyopotansiyel sinyal analizi de dikkat çeken bir başlık olarak öne çıkmaktadır. Biyosinyal çözümleme süreçlerinde kullanılan veri setlerinin yapısı ve özellik uzayları oldukça çeşitlilik gösterebilmektedir. Bununla birlikte, hem sınıflama hem de tanılama görevlerinde kümeleme yöntemlerinin etkin şekilde uygulandığını raporlayan çok sayıda çalışma mevcuttur. Elektrokardiyografi (EKG) sinyalleri, bu alanda en yoğun biçimde incelenmiş başlıca veri kaynaklarından biridir. Nitekim, bir çalışmada EKG sinyallerindeki artefaktların belirlenmesi için zaman serisi kümeleme tabanlı bir yaklaşım uygulanmış ve yöntemin yüksek doğrulukla sınıflama başarısı sağladığı raporlanmıştır (Rodrigues vd., 2017). Bir başka EKG çalışmasında ise, kümeleme yaklaşımı eğitim verilerinin gruplandırılmasında kullanılarak yapay sinir ağı tabanlı sınıflayıcının eğitim süreci optimize edilmiştir (Özbay vd., 2006). Bir başka sık çalışılmış biyopotansiyel veri kaynağı ise eletromiyografi (EMG) sinyalleridir. Örnek olabilecek kümeleme tekniği uygulanmış bir çalışmada EMG sinyallerinin spektral özellikleri gruplanarak normal ve kas hasarı olan bireylerin birbirinden ayrılması sağlanmıştır (Zhang vd., 1991). Yine bir başka EMG çalışmasında fiziksel hareketlerin kümeleme yaklaşımı ile protez kontrolü için sınıflandırılması hedeflenmiştir (Asanza vd., 2018). Farklı biyopotansiyel sinyal işleme örneklerinden birinde ise epilepsi tespiti amacıyla elektroensefalografi (EEG) sinyalleri yapay sinir ağı tabanlı bir modelle sınıflandırılmış, kümeleme yaklaşımı ise özniteliklerin önişleme aşamasında düzenlenmesi için kullanılmıştır (Orhan vd., 2011).

Gözlemler arasındaki ilişkilerin incelendiği ve bu ilişki örüntülerinin klinik açıdan önem taşıyabildiği araştırma alanlarından

birisi de özellikle genetik verilerin analizi ile öne çıkan biyoinformatiktir. Donanım tabanlı örnek bir çalışmada bilgisayarlardan bağımsız çalışabilecek elektronik bir donanımın genom analizi için kümeleme yapabilmesi sağlanmış, sistemin yazılım modeline göre çok daha hızlı çözümleme yapabildiği raporlanmıştır (Hussain vd., 2011). Bir biyoinformatik derleme çalışmasında ise yüksek boyutlu verilerin küme analizi için derin öğrenme tabanlı yaklaşımların klasik yöntemlere kıyasla daha başarılı sonuçlar verebileceği ortaya konmuştur (Karim vd., 2021). Alternatif araştırmalardan birinde spektral kümeleme yaklaşımının biyoinformatik görevlerinde etkili bir şekilde kullanılabileceği raporlanmıştır (Higham vd., 2007). Mikrodizi analizi yapan bir çalışmada ise çok basamaklı bir yaklaşım uygulanarak farklı kümeleme yöntemleri karşılaştırılmıştır (Ciaramella vd., 2008). Yapılan araştırmalar, kümeleme yaklaşımlarının farmakogenetik çalışmalarda da etkin biçimde kullanılabileceğini ortaya koymuştur (Shannon vd., 2003).

Literatürden örnek olarak gösterilmiş tüm bu çalışmalar, biyopotansiyel sinyallerden genetik verilere kadar tıbbi pratiğin pek çok farklı alanından gelen verilerin kümeleme yaklaşımları ile analizinin mümkün olduğunu ortaya koymaktadır. Önceki çalışmalar kümeleme yaklaşımlarının hem keşifsel analizlerde hem de klinik uygulamalarda önemli bir araç haline geldiğini göstermektedir.

1.2 Kitabın Amacı ve Kapsamı

Tıbbi verinin sağlanmasında kilit rol oynayan hekimin bakışı, tanı, tedavi ve keşif süreçlerinde de büyük önem taşır. Bu bağlamda, gözlem ve bilgi sağlayan enstrümanların modernliği ve etkinliği tıbbi pratiğin başarısını nasıl belirliyorsa, verinin analizinde kullanılacak araç ve yöntemler de keşif ve anlamlandırma sürecinin etkinliğini belirler. Hekimin veriyi tanıyarak veri yapısına ve doğasına hakim olması ve gözlem uzayını anlamlandırabilmesi önemli ilişkileri ve yeni bilgiyi ortaya çıkarabilecek önemli bir yetenektir. Bu yeteneği sağlayacak ve ileri götürecek pek çok bilgisayar destekli yaklaşım ve istatistiksel analiz yöntemi bulunmaktadır. Küme analizleri de bu bağlamda yaygın kullanım bulmuş önemli

araçlardan birisidir.

Bu kitap yalnızca hekimler için değil; biyomedikal mühendisliği, sağlık bilimleri ve veri bilimi alanında çalışan araştırmacılar için de yol gösterici bir kaynak olarak tasarlanmıştır. Okuyucunun veri ve veri madenciliği kavramlarıyla yeni bir başlangıç yapması hedeflenmektedir. Böylece, altyapısından bağımsız olarak özellik uzayını, gözlemleri ve gözlemler arasındaki ilişkileri küme yaklaşımlarıyla inceleyebilmesi için gerekli temel nosyon ve teknik bilgi kazandırılacaktır. Bu doğrultuda kitabın sonraki bölümlerinde temel kavramlar ele alınacak, veri kümelerinin doğasını belirleyen özellikler incelenecektir. Ayrıca, bahsi geçen yöntem ve metrikler sentetik veri kümeleri üzerinden örneklendirilerek okuyucunun temel analiz yaklaşımlarını uygulayabilmesi amaçlanmıştır.

Bölüm 2

Temel Kavramlar

2.1 Veri Kümeleri

Veri kümesi kavramını açıklamak için çok geniş bir spektruma yayılmış multidisipliner bir tanım yapmak gerekebilir. Bu nedenle kesin sınırları netleştirilmeden bilimsel ve teknik literatürde sıkça kullanılagelmiş bir kavramdır. Genel kabul gören anlayışa göre bir veri kümesi; belirli bir gruplama mantığıyla bir araya getirilmiş, belli bir içeriğe sahip, kendi içinde ilişkili olan ve belirli bir amaç doğrultusunda kullanılan veriler bütünüdür. Bu yönüyle veri kümesi, yalnızca verilerin bir araya gelmiş hali değil, aynı zamanda gözlemler arasındaki olası örüntülerin bilimsel bilgi yaratacak şekilde ortaya çıkarılmasına hizmet eden bir organizasyondur (Renear vd., 2010).

Veri kümeleri sayısal görüntülerden analog sinyallere, hasta demografisinden anket sonuçlarına kadar çok çeşitli veri tipleri ile oluşturulmuş olabilirler. Veri kümelerini niteleyecek ve tanımlanmalarına yardımcı olacak pek çok faktör bulunmaktadır. Bu faktörler arasında verinin kaynağı (ölçüm, gözlem, sensör çıktısı, kayıt), formatı ve yapısı (sayısal, metinsel, görsel, işitsel), kapsamı (örneklem büyüklüğü, zaman aralığı, coğrafi kapsam) ve bağlamı (verinin hangi amaçla ve hangi koşullarda toplandığı) öne çıkmaktadır.

Veri kümelerinin kullanımına dair sahip olunması gereken önemli farkındalıklardan birisi de fikri mülkiyet hakları ve etik

değerlere uygun kullanım bilincidir. Bu noktada analizi yapılacak verinin nasıl ve hangi şartlarda toplandığından, verinin gerçek sahibinin kim olduğuna kadar dikkat edilmesi gereken pek çok husus bulunmaktadır. Bu nedenle araştırmacıların, veri kümelerini kullanırken yalnızca teknik doğruluk ve bilimsel katkıyı değil, aynı zamanda etik sorumlulukları da gözetmeleri gerekir. Verinin anonimleştirilmesi, yetkisiz erişimin engellenmesi ve gerekli izinlerin alınması, veri kümeleri ile çalışılırken mutlaka dikkat edilmesi gereken hususlardır. Unutulmamalıdır ki bu bağlamda yasal izinler kadar etik kurulların onayı da kritik önem arz etmektedir. Literatürde, yasa dışı yollarla elde edilmiş veri kümelerinin kullanımına dair etik ve hukuki sorunların sıklıkla göz ardı edildiği vurgulanmaktadır (Thomas vd., 2017). Araştırmacılar kullanacakları veri kümesinin fikri mülkiyet hakkının kimde olduğunu, verinin ne gibi izinlerle toplandığını ve yapacakları araştırmaların bu izinlerin dahilinde olup olmadığını mutlaka netleştirmelidir. Devlet hastaneleri gibi kamu kurumlarından toplanacak hasta orijinli verinin ancak ve muhtemelen pek çok izinle kullanılabileceği unutulmamalıdır. Bu çerçevede veri kümelerinin (açık erişimli veriler dahil) etik kullanımını değerlendirmek için araştırmacılar öncelikle verilerin yasal yollarla ve izinli biçimde toplanmış olduğundan emin olmalıdır. Buna ek olarak gerekli durumlarda etik kurul onaylarını, rıza belgelerini sağlamalı, yasal uygunluğu mutlaka öncelikli olarak denetlemelidir. Anonimleştirme, şifreleme, erişim kontrolü ve güvenli saklama gibi önlemlerin gerekliliği değerlendirilmelidir. Bu konudaki farkındalık bilimsel çalışmanın güvenilirliğini arttırdığı gibi aynı zamanda toplumla kurulan güven ilişkisini de pekiştirmeye adaydır.

2.1.1 Gözlemler ve Özellikleri

Bir veri kümesini oluşturan her bir temel birim gözlemdir. Küme analizleri gözlemler arasındaki olası örüntüleri ortaya çıkartmaya yarayan yaklaşımlardan birisidir. Buna göre veri kümesi olarak gördüğümüz yığın aslında bir grup gözlemden ibarettir. Ancak bu veri yığınının temel yapısını belirleyen başat etkenlerden birisi de gözlemlerin nasıl temsil edildiğidir. Bir gözlemi niteleyen her farklı

değişken o gözlemin bir özelliğidir. Aşağıda on gözlemden oluşan sentetik olarak oluşturulmuş bir örnek veri seti verilmiştir.

Tablo 2.1: Sentetik olarak üretilmiş mini demografik veri seti

ID	Boy (m)	Kilo (kg)	Açlık şekeri	Aile öyküsü	BMI	Obez
1	1.78	64	89	0	20.1	0
2	1.70	60	86	0	20.6	0
3	1.79	76	86	0	23.6	0
4	1.77	63	86	0	20.1	0
5	1.69	63	85	0	22.0	0
6	1.74	103	127	1	33.9	1
7	1.66	92	114	1	33.4	1
8	1.79	101	127	1	31.6	1
9	1.69	90	124	1	31.6	1
10	1.63	86	125	1	32.5	1

Verilen tabloda her bir satır bir gözlemi göstermekteyken her bir sütun ise bir özelliği gösteriyor olabilir. Ancak burada analizin bağlamına dikkat etmek gerekir. Hastanın obezite durumu bir özellik olarak görülebilir ancak tabloda verilen küme seti hastaların obezite durumlarına göre iki kümede incelenmek istenseydi obezite sütunu artık bir özellik sütunu değil etiket sütunu haline gelecekti. Yani veri setimizdeki her bir hastanın hangi gruba dahil olduğunu belirten sütun olacaktı. Böyle bir durumda özellik uzayımız altı değil beş özellik barındırıyor olacaktı. Aynı veride obezite durumu değil de aile öyküsü etiket olarak kullanılsaydı bu sefer kalan beş özelliğin ailede şeker hastalığı geçmişinin olup olmadığı ile ilgisi araştırılmış olurdu.

Dikkat edilmesi gereken hususlardan birisi de özelliklerin kümeler üzerindeki ayırt ediciliğidir. Bu noktada bağlam mutlaka gözönünde bulundurulmalıdır. Özellikler gözlemleri tanımlayarak onları karakterize etmektedir ancak özelliklerin gözlemleri sınıflara ayırma konusunda güçleri eşit olmadığı gibi bağlama yani çıktı değişkenine bağlı olarak ayırt edicilikleri değişiklik gösterebilir. Örneğin vücut kitle indeksi veya açlık şekeri değerleri obezite durumunu belirlemede güçlü ayırt edici özelliklerdir; ancak boy tek başına bu ayrımı yapmakta zayıf kalabilir. Bu nedenle, özelliklerin ayırt ediciliği kavramı kümeleme analizlerinde kritik bir rol oynar.

Unutulmamalıdır ki çok sayıda farklı özellikten oluşan veri kümeleri söz konusu olabilir. Bu tür veri kümelerinin küme analizinde anlamlı örüntülerin ortaya çıkarılması, çıktı değişkenine göre en yüksek ayırt ediciliğe sahip özelliklerin belirlenmesiyle mümkün olabilir. Tersinden düşünüldüğünde, mevcut özelliklerin en yüksek ayırt ediciliğe sahip olacakları çıktı değişkeninin bulunması da anlamlı örüntülerin keşfinde anahtar rol oynayabilir.

2.1.2 Veri Kümelerinin Görselleştirilmesi

Veri kümelerinin görselleştirilerek incelenmesi, küme analizinde gözlemler arasındaki ilişkilerin daha anlaşılır biçimde ortaya konulmasında sıkça kullanılan yöntemlerden birisidir. Görselleştirme sayesinde gözlemlerin dağılımı, uç örneklerin varlığı ve sınıflar arasındaki potansiyel ilişkiler daha net değerlendirilebilmektedir. Ayrıca, sınıfların birbirinden ayrılma derecesi hakkında fikir edinebilmek adına farklı görselleştirme tekniklerinden faydalanmak mümkündür. Bununla birlikte, özellik sayısı arttıkça veri kümelerinin görselleştirilerek incelenmesi daha karmaşık bir problem haline gelebilmektedir.

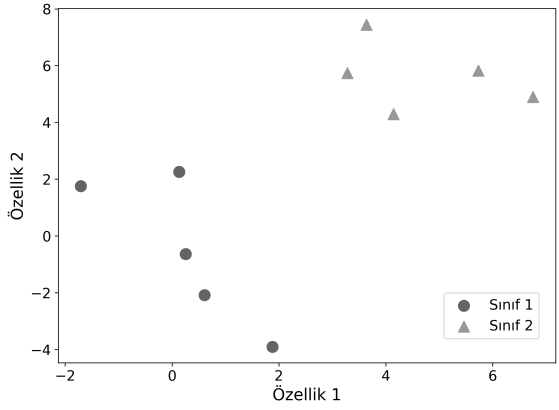
2.1.2.1 Dağılım Grafikleri

Temel görselleştirme yaklaşımlarından biri de dağılım grafikleridir. Bu grafikler, özellik uzayının iki veya üç boyutla incelenebildiği durumlarda gözlemlerin dağılımını incelemek için en sık kullanılan yöntemler arasındadır. Bu grafikler sayesinde kümelerin birbirine yakınlığı, yoğunluk bölgeleri ve uç değerlerin varlığı kolaylıkla fark edilebilir (Mayorga ve Gleicher, 2013). Dağılım grafiklerini görselleştirmeye örnek olması için on gözlem ve üç farklı özellikten oluşan iki sınıflı bir sentetik veri kümesi üretilmiştir. Tablo 2.2'de üretilen örnek veri seti görülmektedir.

Tablo 2.2: İki sınıflı veri kümesi dağılım grafiği örneği için sentetik olarak oluşturulmuştur.

ID	Özellik 1	Özellik 2	Özellik 3	Sınıf
1	0.609	-2.080	1.501	0
2	1.881	-3.902	-2.604	0
3	0.256	-0.632	-0.034	0
4	-1.706	1.759	1.556	0
5	0.132	2.254	0.935	0
6	3.281	5.738	3.082	1
7	6.757	4.900	4.630	1
8	3.638	7.445	4.691	1
9	4.143	4.296	6.065	1
10	5.731	5.825	5.862	1

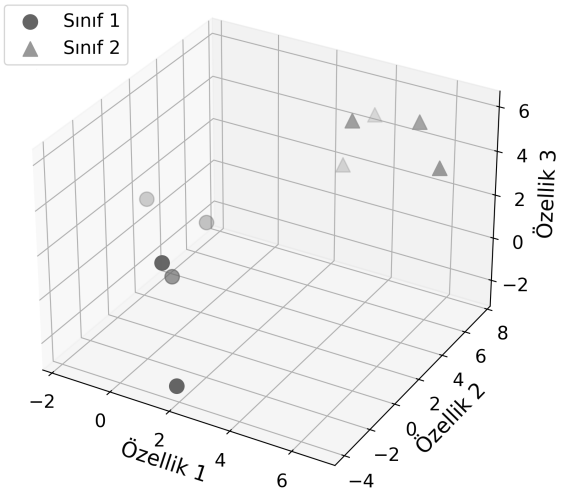
Tablo 2.2'de verilen sayısal özelliklerin iki sınıftan birine ait olduğu görülmekte. Buna göre bu veride sınıf kümelerinin birbiri ile olan ilişkisi pek çok farklı dağılım grafiği alternatifi ile gözlemlenebilir. Örnek olarak özellik 1 ve özellik 2 kullanılarak oluşturulmuş dağılım grafiği Şekil 2.1'de verilmiştir.



Şekil 2.1: Örnek veri setindeki gözlemlerin özellik 1 ve özellik 2 eksenlerinde dağılım grafiği

Şekil 2.1'de sınıf kümelerinin yüksek düzeyde ayrıştıkları görülüyor. Gözlemleri karakterize eden başka özellik kombinasyonları ile

oluşturulmuş bir dağılm grafiği yeni bir perspektif sunacağı gibi benzer ayırt ediciliğin görülemeyebileceği unutulmamalıdır. Şekil 2.2'de aynı gözlemler üç özellikle birlikte temsil edilerek dağılım grafiği tekrar oluşturulmuştur.



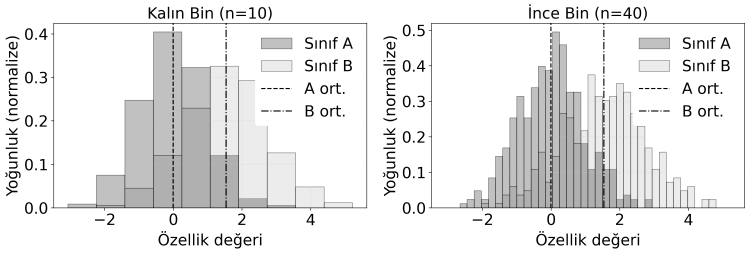
Şekil 2.2: Örnek veri setindeki gözlemlerin üç özellikle birlikte oluşturulmuş dağılım grafiği

Aynı özellik eksenleri ile farklı bir sınıf değişkeni kullanılırsa yine ilişki örüntülerinin ve ayırt ediciliğin değişebileceği not edilmelidir. Buna örnek olarak bir hasta grubundaki obezite sınıflarının gözlenmesi verilebilir. Bu örnek uzayında bireyler Kilo, BMI ve gözlük kullanım yılı olmak üzere üç özellik ile karakterize edilsin. Bu noktada kilo ve BMI özellikleri obezite sınıfını belirlemede güçlü ayırt edici özellikler iken, gözlük kullanım yılı bu ilişkiyi yansıtmayabilir. Bu durum, özellik seçiminde dikkatli olunması gerektiğini ve farklı kombinasyonların incelenmesinin önemli olduğunu göstermektedir.

2.1.2.2 Histogramlar

Histogram grafikleri kümelerin ve kümeler arası ilişki karakterinin görsel olarak incelenmesi için yaygın olarak kullanılan bir başka görselleştirme aracıdır. Histogramlar gözlemlerin spesifik değerlere

veya aralıklara ne sıklıkta dağıldığını gösteren grafiklerdir (Scott, 2010). Bu sayede farklı sınıflara ait gözlemlerin hangi değerlerde yoğunlaştığını ve ne kadar saçıldığını görmek mümkün olmaktadır. Histogramların yalnızca tek değişkeni gösterebildiği unutulmamalıdır. Dolayısı ile büyük miktarlarda özellik içeren çok boyutlu uzayların analizinde kullanımları zor olabilir ancak her bir değişken için sınıflar arası ilişkilerin ortaya çıkarılmasında etkin rol oynayabilmektedirler. Mesela sınıfların üst üste bindiği veya yoğunluk merkezilerinin birbirinden ayrık olduğu durumlarda ilgili özelliğin ayırt ediciliğine dair tahmin yürütmek mümkün olabilir. Ayrıca yine sınıf dağılımlarının simetrisi veya çarpıklığı gözlemlerin dağılım eğilimlerine dair fikir verebilmektedir. Sınıf kümeleri histogramlar ile incelenirken dikkat edilmesi gereken hususlardan birisi de histogramlar çizilirken tercih edilen değerler veya aralıklardır. Bin olarak da adlandırılan aralıkların seçimi histogramın çözünürlüğünü belirlemektedir. Mesela ilgili özellik 0 ile 10 arasında değişen değerler alıyorsa ve bin sayısı beş olarak alındıysa bin kenarları [0, 2, 4, 6, 8, 10] şeklinde olacaktır. Her bin kendi aralığına düşen gözlem sayısını göstermektedir. Buna göre bin genişliği çok büyük seçilirse dağılım daha yumuşak geçişler yapacaktır ancak bu sınıflar arası küçük ayrımların görülmesini zorlaştırabilir. Çok küçük seçilmiş bin aralıkları ise keskin geçişlere ve gürültüye sebep olarak yorum yapmayı zorlaştırabilir. Tıbbi veri incelenirken küçük farklılıkların incelenmesi önemli olabilir ancak klinik anlamı olan aralığın bulunması ve aynı özelliğin farklı sınıflara ait histogramlarının aynı bin aralıklarıyla çizilmesi analizin etkinliği açısından kritik olabilir. Bin aralığı seçiminin görsel etkisi şekil 2.3'te gösterilmiştir.



Şekil 2.3: İki sınıfın aynı özelliğe ait histogramlarında bin genişliği seçiminin etkisi görülmektedir.

2.1.3 Etiketler ve Sınıflar

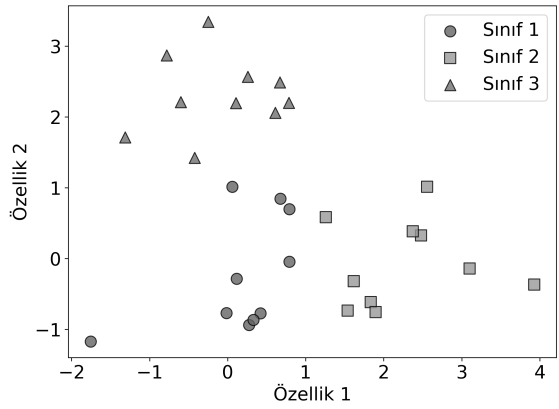
Kümeleme veya sınıflandırma problemin karakterini belirleyen önemli bir kavram da etikettir. Buna göre etiketler her bir gözlemin ait olduğu sınıfı ya da kategoriyi belirten değişkendir. Bu noktada hangi gözlemin hangi kümeye ait olduğu biliniyorsa etiket değişkeni altın standart, gerçek değer veya doğruluk ölçütü olarak adlandırılabilir (Färber vd., 2010). Ancak gözlemler bir makine öğrenimi yaklaşımı ile otomatik olarak bir kümeye atandılarsa bu durumda etiket değişkeni artık tahmin veya kestirim olarak adlandırılabilir. Kestirim değişkeni ile gerçek değer değişkeni arasındaki benzerlik sınıflamanın ne kadar başarılı gerçekleştirildiğini göstermektedir ve bu benzerlik farklı metriklerle sayısal olarak ölçülebilir.

Başka bir ifadeyle, etiket değişkeni veri kümesindeki gözlemleri önceden tanımlanmış gruplara ayırmaya yarayan bilgiyi sağlamaktadır. Gözlemlerin ait oldukları sınıfları önceden bilmemiz küme analizini ilgilendiren bazı durumlarda elzem olabilir. Mesela etiketi bilinmeyen bir gözlemin hangi sınıfa ait olabileceğini kestirmek için önceden sınıfı bilinenen gözlemlerin özellikleri kullanılabilir. Buna ek olarak kılavuzlu öğrenime ihtiyaç duyulacağı herhangi senaryoda makine öğrenimi modelinin probleme adaptasyonunu sağlamak için mutlaka bilinen etiketlerle bir eğitim verisi hazırlamak gerekmektedir. Unutulmamalıdır ki hangi sınıfa ait oldukları bilinmeden de örnekler kılavuzsuz yaklaşımlar ile kümelenebilirler. Etiket değişkeninin varlığı ya da yokluğu, uygulanabilecek istatistiksel yöntemlerin ve makine öğrenmesi algoritmalarının seçiminde

belirleyici rol oynamaktadır.

2.1.3.1 İkili ve Çok Sınıflı Özellik Uzayları

Özellik uzayının sadece iki alternatif sınıftan ibaret olduğu durumlarda etiket değişkeni sadece iki değer almaktadır kümeleme problemi ise iki sınıflı (binary) olarak adlandırılır. Mesela bütün örnekler hasta/sağlıklı, normal/abnormal, test pozitif/test negatif şeklinde etiketlendiyse özellik uzayı ikilik olacaktır. Ancak elbette ikiden fazla sınıfla temsil edilen özellik uzayları mevcuttur ve bunlar çok sınıflı (multi-class) olarak adlandırılır. Örneğin tümör örneklerinin benign, malign veya prekanseröz olarak sınıflandırılması çok sınıflı özellik uzayına bir örnek olarak gösterilebilir. Aşağıda çok iki özellikli ve çok sınıflı bir sentetik özellik uzayının dağılım grafiği (Şekil 2.4) gösterilmiştir.



Şekil 2.4: İki özellikli, çok sınıflı dağılım grafiği

2.1.3.2 Sınıf Dengesi

Özellik uzayında her bir sınıfa ait olan gözlemlerin sayısının birbirine olan yakınlığı sınıf dengesi kavramı ile nitelenmektedir. Buna göre dengeli bir veri kümesinde sınıflar benzer sayıda gözlem bulundururken dengesiz bir veri kümesinde sınıflardan bazıları baskın şekilde daha fazla veya çok az sayıda örnek ile temsil ediliyor olabilir. Bu yapılacak analizin doğruluğunu etkileyebilecek bir hu-

sustur (Jan ve Verma, 2020). Bir sınıf eğer sadece bir gözlemden oluşuyor ise tekil (singleton) sınıf olarak adlandırılmaktadır. Tekil sınıfların varlığı söz konusu ise özellikle istatistiki kestirimlerin doğruluğunun etkilenebileceği unutulmamalıdır.

Sınıflardan birisinin çok az sayıda gözlem içerdiği durumlarda sınıf dengesini sentetik olarak iyileştiren çeşitli yaklaşımlar bulunmaktadır (Yang vd., 2024). Bunlardan birisi aşırı örneklemedir. Rastgele aşırı örneklemede azınlık sınıfa ait rastgele seçilmiş örnekler hedeflenen sınıf oranına ulaşana kadar kopyalanır (aynı gözlem birden çok kez kopyalanabilir). En pratik yöntemlerden biri olmakla birlikte sınıf içindeki gürültünün etkisini artırma ihtimali bulunduğundan dikkatle uygulanmalıdır. Alternatif olarak SMOTE (Synthetic Minority Over-sampling Technique) yöntemi gösterilebilir. Bu yaklaşımda azınlık sınıfına ait yeni örnekler komşu gözlemler arasında kurulan doğrusal ilişkiye dayanılarak üretilir. Bu sayede sınıfın içerdiği gözlem miktarı sentetik ama kopya olmayan gözlemler ile genişletilmektedir (Chawla vd., 2002). Böylece birbirinin aynısı kopya gözlemlerden kaynaklanması muhtemel analiz problemleri kısmen azaltılabilir. Tablo 2.3'de iki aşırı örnekleme yaklaşımı ile sentetik bir dengesiz dağılımlı veri setinden üretilmiş yeni gözlemler gösterilmiştir.

Tablo 2.3: Orijinal veri, rastgele aşırı örnekleme ve SMOTE sonrası elde edilen örnekler.

Orijinal Dengesiz Veri			Rastgele Aşırı Örnekleme			SMOTE Aşırı Örnekleme		
Özellik 1	Özellik 2	Etiket	Özellik 1	Özellik 2	Etiket	Özellik 1	Özellik 2	Etiket
0.305	-1.040	0	0.305	-1.040	0	0.305	-1.040	0
0.750	0.941	0	0.750	0.941	0	0.750	0.941	0
-1.951	-1.302	0	-1.951	-1.302	0	-1.951	-1.302	0
0.128	-0.316	0	0.128	-0.316	0	0.128	-0.316	0
-0.017	-0.853	0	-0.017	-0.853	0	-0.017	-0.853	0
0.879	0.778	0	0.879	0.778	0	0.879	0.778	0
0.066	1.127	0	0.066	1.127	0	0.066	1.127	0
0.468	-0.859	0	0.468	-0.859	0	0.468	-0.859	0
3.369	2.041	1	3.369	2.041	1	3.369	2.041	1
3.878	2.950	1	3.878	2.950	1	3.878	2.950	1
			3.369	2.041	1	3.765	2.749	1
			3.878	2.950	1	3.641	2.526	1
			3.878	2.950	1	3.800	2.810	1
			3.369	2.041	1	3.748	2.718	1
			3.369	2.041	1	3.535	2.337	1
			3.878	2.950	1	3.639	2.523	1

Eksik örnekleme yaklaşımları da aşırı örneklemeye alternatif olarak görülebilir. Aşırı örneklemeden farklı olarak bu sefer çoğunluk sınıfındaki gözlemlerin bir kısmı çıkarılarak sınıf dengesini sağlanmaktadır. Bu eksiltme işleminin baskın sınıftaki bilgi ve örüntülerin kaybolmasına sebep olabileceği unutulmamalıdır. Ek olarak sınıf ağırlıklandırma veya topluluk yaklaşımları da sınıf dağılımındaki dengesizliğin etkisini azaltabilecek alternatif yaklaşımlardandır. Bu yöntemler uygulanmadan önce verinin tipi, dağılımı, gürültü varlığı gibi etkenler mutlaka göz önünde bulundurulmalıdır.

2.1.4 Özellik Uzayı ve Boyut

Boyut sayısı arttıkça veri kümelerinin görselleştirilmesi ve yorumlanması güçleşir. Bu nedenle yüksek boyutlu verilerle çalışırken özellik sayısının azaltılması gerekebilir. Bunun için uygulanan temel yaklaşımlardan birisi özellik seçimidir. Yani mevcut değişkenler arasından en tanımlayıcı olanlarının seçilmesi anlamına gelmektedir. Bu noktada ayırt edicilik kavramının bağlamı önem kazanmaktadır. Eğer gözlemlerin sınıfları biliniyorsa, bu sınıfları en yüksek derecede karakterize eden özellikler aranabilir. Sınıfların bilinmediği durumlarda ise denetimsiz kümeleme yaklaşımları ile veri gruplandırılarak, oluşan kümeler çeşitli metrikler ile nitelenebilir. Böylece en ideal kümeleri yaratan özellikler seçilerek özellik uzayı daraltılabilmektedir.

Bu noktada sistematik arama için farklı algoritmik yaklaşımlar önerilebilir. En temel yöntemlerden biri yorucu (exhaustive) aramadır. Bu yaklaşım, olası bütün özellik kombinasyonlarının denenerek skorlandırılmasına dayanmaktadır (Nersisyan vd., 2022). Ancak bu yaklaşımın çok büyük özellik uzayları için pratik olmayabileceği göz önünde bulundurulmalıdır. Bu yaklaşıma alternatif olarak rekürsif özellik eleme (Recursive Feature Elimination, RFE) gösterilebilir. RFE ise sistematik olarak özellik kombinasyonlarının değerlendirilmesine ve karakterizasyona en az katkı yapan özelliklerin adım adım elenerek, daha küçük ama daha ayırt edici bir özellik alt kümesi oluşturulmasına dayanmaktadır (Darst vd., 2018). Meta-sezgisel yaklaşımlar ile de en güçlü özelliklerin bulunabileceği unutulmamalıdır. Özellikle arama uzayının çok büyük ve sofistike

olduğu durumlarda genetik algoritma gibi alternatif meta-sezgisel yaklaşımlar özellik uzayını en iyi şekilde sadeleştirecek uygun özelliklerin seçilmesi için faydalı olabilir (Agrawal vd., 2021).

Özellik seçiminden farklı bir başka yaklaşım ise boyut indirgeme yani mevcut değişkenlerin daha az sayıda yeni değişkenlere dönüştürülmesidir. Bunun için birden fazla yöntem olmakla birlikte PCA (Principal Component Analysis) en yaygın kullanılagelmiş yöntemlerden birisidir. PCA tabanlı boyut azaltma veri varyansının mümkün olduğunca korunarak izdüşümünün bulunmasına dayanmaktadır (Berguin ve Mavris, 2015). Ortaya çıkan yeni matris üç boyutlu bir objenin iki boyutlu gölgesi gibi gerçek bilginin daha az boyutlu ancak varyansı korunmuş bir temsili olarak işlenebilmektedir. Bu sayede yüksek boyutlu veri kümelerinin kümeleme öncesinde daha anlaşılır hale getirilmesine yardımcı olmaktadır.

2.2 Veri Tipleri

2.2.1 Nitel (Kategorik) Veriler

Nitel veriler gözlemleri sayısal bir ölçüm ile karakterize etmek yerine onları gruplayacak şekilde bir değer veya etiket ile belirli kategorilere ayırır. Bu kategoriler sayısal bir değer ile ifade edilmiş bile olsa numerik değer olarak değil bir sınıf etiketi olarak değerlendirilmelidir. Bu nedenle doğrudan ortalama alma, toplama gibi aritmetik işlemlere tabi tutulmaları doğru olmayabilir. Nitel veriler, nominal ve ordinal olmak üzere iki alt tipe ayrılır.

2.2.1.1 Nominal Veri Tipi

Nominal özelliklerde kategori ayrımı yapan değerler birbirinden tamamen bağımsızdır. Kategoriler arasında herhangi bir sıralama hiyerarşi veya önem farkı yoktur. Nominal özellik sütunlarında değerler sayısal olarak verilmiş olsa bile (örneğin 1 = erkek, 2 = kadın) nicel anlam taşımadıklarından aritmetik işlemlere tabi tutmak anlamlı olmayacaktır. Mesela cinsiyet, sigara kullanma durumu, kan grubu, diyabet, kırışıklık tipi gibi veriler nominal veri

tipine örnek olabilir.

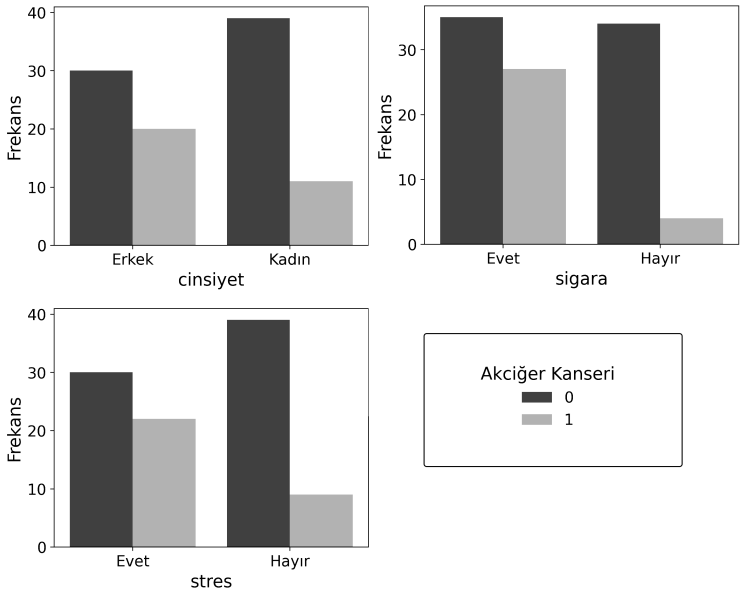
Bu tip değişkenler analiz edilirken aritmetik işlemler yerine frekans analizi, dağılım veya oran gibi değerlendirmeler kullanılmalıdır. İki kategorik değişken arasındaki ilişki ki-kare analizi ile incelenebilir (McHugh, 2013). Bu sayede kategorik değişkenlerin bağlılığına dair yorum yapmak mümkün olabilir. Nominal verinin analizini örneklendirmek için üç değişkenden oluşan 100 gözlemlik bir sentetik veri kümesi oluşturulmuştur. Verinin kısmi bir görünümü Tablo 2.4'te görülmektedir.

Tablo 2.4: Sentetik veri kümesinden örnek gözlemler

Cinsiyet	**Sigara**	**Stres**	**Akciğer Kanseri**
Erkek	Evet	Evet	1
Erkek	Hayır	Evet	1
Erkek	Hayır	Hayır	0
Kadın	Hayır	Hayır	0
...	...	...	...
Kadın	Evet	Evet	1
Erkek	Hayır	Evet	1

Tablo 2.4'te örneklenen veride Cinsiyet değişkeni %50 kadın ve %50 erkek olacak şekilde eşit olasılıkla dağıtılmıştır. Bu özellik sınıf değişkeni ile ilişkili olmayacak biçimde dağıtılmıştır. Sigara kullanımı ise %60 evet, %40 hayır olasılıkları ile dağıtılmış, stres ve sigara kullanımı değişkenleri bağımlı olacak şekilde veri kurgulanmıştır. Sigara içenlerde stresli işte çalışma olasılığı %70, içmeyenlerde ise %30'dur. Kümeleme değişkeninin akciğer kanserine rastlanma durumunu işaret ettiği farz edilmiştir. Buna göre sigara ve stres kombinasyonlarına göre farklı olasılıklarla dağıtılmıştır. Hem sigara kullanıp hem stresli işte çalışanlarda %50, sadece sigara içenlerde %30, sadece stresli işte çalışanlarda %20, hem sigara kullanmayan hem de stresli işte çalışmayanlarda ise %5 akciğer kanseri olasılığı görülecek şekilde gözlemler kümelenmiştir. Bu yaklaşım sayesinde yaratılan sentetik veride sigara ve stres değişkenleri hem birbirleriyle hem de akciğer kanseri ile ilişkili şekilde gözlenmelidir. Buna karşılık cinsiyet değişkeni akciğer kanserinden bağımsız bir değişken olarak bırakılmıştır. Şekil 2.5'te farklı

kategorilerin akciğer kanserine göre dağılımları görsel olarak temsil edilmiştir.



Şekil 2.5: Cinsiyet, sigara kullanımı ve stres değişkenlerinin akciğer kanseri durumuna göre frekans dağılımları

Şekil 2.5'te sigara kullanan ve stresli işte çalışan bireylerde akciğer kanseri görülme oranlarının belirgin şekilde arttığı, buna karşılık cinsiyet değişkeninin kanser görülme durumu ile anlamlı bir ilişki içinde olmadığı gözlenmektedir. Bu durum, sentetik verinin üretilirken belirlenen özellikleri ile uyumlu bir bulgudur. Bu noktada ki-kare testinin de bu gözlemle uyumlu olduğunu göstermek adına test uygulanarak sonuçları Tablo 2.5'te verilmiştir.

Tablo 2.5: Ki-kare test sonuçları

Değişken 1	**Değişken 2**	**Chi2**	**p-değeri**
Sigara	Akciğer Kanseri	10.517	0.00118
Stres	Akciğer Kanseri	5.421	0.01989
Sigara	Stres	14.581	0.00013
Cinsiyet	Akciğer Kanseri	2.992	0.08367

Sunulan ki-kare test sonuçları incelendiğinde, sigara kullanımı ile akciğer kanseri arasında istatistiksel olarak anlamlı bir ilişki ol-

duğu görülmektedir ($p < 0.01$). Benzer şekilde stresli işte çalışmanın da akciğer kanseri ile anlamlı düzeyde ilişkili olduğu gözlenmiştir ($p < 0.05$). Ayrıca sigara kullanımı ile stres değişkenleri arasında da güçlü bir ilişki bulunmuştur ($p < 0.001$). Bu durum veri setinin üretiminde kurgulanan bağımlılık ve oluşturulan grafiker iler uyumlu bir sonuçtur. Yine başlangıçta planlandığı gibi test sonuçlarında da cinsiyet ile akciğer kanseri arasında anlamlı bir ilişki saptanmamıştır ($p > 0.05$). Bu noktada ki kare testi yorumlanırken test sonucunda elde edilen ki kare değeri gözlenen ve beklenen dağılımlar arasındaki farkın büyüklüğünü ifade etmektedir. Ancak farkın anlamlılığı p değerine bakılarak yorumlanmaktadır. Tablo 2.5'te verilen sonuçlara göre sigara ve stres değişkenleri hem yüksek ki-kare değerleri hem de düşük p değerleri ile akciğer kanseriyle güçlü biçimde ilişkili bulunmuştur.

2.2.1.2 Ordinal Veri Tipi

Ordinal değişkenler nominal değişkenler gibi ayrık kategorik bilgi taşımaktadır. Bu veri tipinin en önemli farklılığı bir sıralama yapılarak gözlemlerin değerlerine göre bir biri ile mukayese edilebiliyor oluşudur. Ancak kategoriler arasındaki mesafelerin eşit kabul edilememesi nedeniyle analiz buna göre gerçekleştirilmelidir. Sağlık bilimlerinde sıkça kullanılan ağrı şiddeti skalası ordinal veriye örnek olarak gösterilebilir.

Ordinal verilerde kategoriler sıralı olmakla birlikte kategoriler arası mesafeler eşit olmayabilir. Bu nedenle ortalama hesaplamak istatistiksel olarak yanlış sonuçlar üretebilmektedir. Bu tür verilerde analiz yöntemi olarak medyan ve mod tercih edilebilir. Varyans analizi (ANOVA) veya Pearson korelasyonu gibi eşit aralıklarla değişim gerektiren analiz yöntemleri de bu nedenle dikkatle uygulanmaldır. Eğer eşit aralık varsayımı sağlanmıyorsa bu yöntemler uygulanmamalı bunun yerine Mann–Whitney U, Kruskal–Wallis veya Spearman korelasyonu gibi parametrik olmayan yöntemler tercih edilmelidir.

Küme analizi yapılırken özellikle gözlemlerin mesafesine dayalı algoritmalarda ordinal verinin nasıl değerlendirileceğine çok dikkat etmek gerekir. Eğer etiket kodlama yöntemleri tercih ediliyorsa

yöntemin değerler arasındaki mesafeyi eşit kabul edebileceği unutulmamalıdır. Bununla birlikte ikilik kodlama (one-hot encoding) uygulanırsa gözlemlerin sıralanması artık mümkün olmayacaktır. Bu nedenle ordinal verilerle çalışırken, kümeleme yaklaşımları Spearman veya Kendall sıra korelasyonu ya da Gower mesafesi gibi alternatif ölçüler kullanılarak uygulanmalıdır.

2.2.2 Nicel Veriler

2.2.2.1 Sürekli Veri Tipi

Sürekli veriler belli bir aralıkta herhangi değer alabilen nicel özelliklerdir. Ölçümle elde edilen veriler genellikle sürekli veri tipi şeklinde analiz edilmektedir. Buna örnek olarak hasta boyu ve kilosu gösterilebilir. Yine vücut sıcaklığı veya kan basıncı gibi parametreler de sürekli veri olarak karşımıza çıkabilir. Ayrıca biyopotansiyel sinyallerin özellikleri ve anlık değerleri de sürekli verilere örnek olarak gösterilebilir. Sürekli verinin istatistik analizinde genellikle aritmetik ortalama, standart sapma, varyans, korelasyon veya regresyon gibi hesaplamalar tercih edilmektedir. Histogramlar ve saçılım grafikleri gibi görselleştirme yöntemleri de bu tip verilerin dağılımını ortaya koymada yaygın olarak kullanılmaktadır.

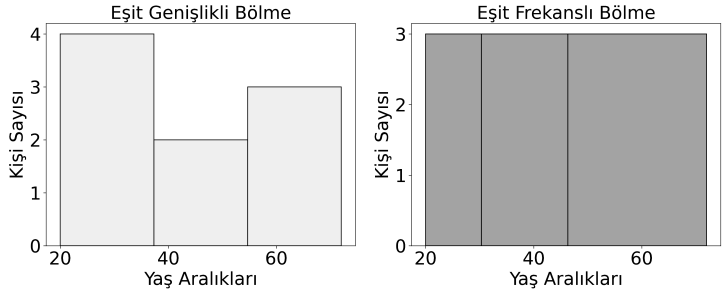
2.2.2.2 Ayrık Veri Tipi

Ayrık veriler nicel veri tipine örnek olmakla birlikte sonsuz sayıda değer alabilmek yerine yalnızca sayılabilir, tam sayı değerleri almaktadır. Dolayısı ile bu tip veriler genellikle tam sayılarla ifade edilir ve iki değer arasında ara değer görülmez. Belli bir zaman aralığında gözlenen hasta sayısı bu özelliğe örnek olabilir. Buna göre hasta sayısı 12, 13 veya 14 olabilir; ancak 13.5 hasta gibi bir değer görülmesi mümkün değildir. Nominal ve ordinal veriler kategorik (nitel) veri grubunda iken ayrık verilerin sayısal (nicel) veri grubunda olduğu unutulmamalıdır. Yani nominal veriler kategorilerden oluşmakta ve sayısal anlam ifade etmemektedir. Ordinal verilerde ise kategoriler arasında bir sıralama vardır ama aralıklar eşit olmayabilir. Ancak ayrık veriler nicel veri grubundadır ve tam sayılarla sayısal karakterizasyon yapmaktadır.

2.2.2.3 Veri Ayrıklaştırma

Sürekli özellikler bazı durumlarda analizi kolaylaştırmak amacı ile kodlanarak ayrık verilere dönüştürülebilirler. Bu sayede nominal/ordinal verilerle benzer şekilde işlenebilmesi mümkün olabilmektedir. Bu işleme ayrıklaştırma (discretization) ya da binleme (binning) adı verilir. Ayrıklaştırma işlemi için farklı yaklaşımlar bulunmaktadır. Bunlardan birisi eşit genişlikli bölme yöntemidir (Ferreira ve Figueiredo, 2012). Bu yöntemde verinin değişim gösterdiği sayısal aralık sabit uzunluklara bölünür ve her bir gözlem bir aralığa atanır. Örnek olarak bir veri kümesindeki yaş değişkenini ele alalım. Gözlemlerin yaşları 20 ile 80 arasında değişiyorsa ve bu aralık üçe ayrılmak istenirse, her bir bölüm aralığının genişliği 20 olacaktır. Böylece yaşlar 20–40, 40–60 ve 60–80 aralıklarına gruplandırılarak ayrık değerler atanabilir. Buna uygun olarak yaşı 33 olan bir kişi ilk aralığa, 57 olan ikinci aralığa, 72 olan ise üçüncü aralığa yerleştirilecektir. Basit ve hızlı bir yöntem olmakla birlikte bu yaklaşımda veri dağılımı dikkate alınmamaktadır. Verilerin belirli bir aralıkta yoğunlaştığı durumlarda bazı aralıklara görece olarak çok daha fazla gözlem düşerken bazıları neredeyse boş kalabilir. Buna ek olarak veri kümesindeki uç değerler (outlier) bölme genişliğine analizin doğruluğunu etkileyecek şekilde etki edebilir. Alternatif bir yaklaşım olan eşit frekanslı bölme yönteminde ise her sınıfta yaklaşık aynı sayıda gözlem bulunur (Ferreira ve Figueiredo, 2012). Bu yöntem, verileri küçükten büyüğe sıraladıktan sonra her bölmeye eşit sayıda gözlem düşecek şekilde kesme noktaları belirler. Böylece tüm bölmelerin dolu olması sağlanmış olur ve özellikle dengesiz dağılımlarda daha dengeli ve anlamlı gruplar elde edilir. Bu yaklaşımda bölme sayısı kritik bir parametredir ve dikkatle seçilmesi gerekir. Bölme sayısı gözlem sayısıyla doğrudan ilişkilidir ve çok az bölme seçilirse her bölmede çok sayıda gözlem toplanır ve analiz çözünürlüğü azalarak aşırı genellemeye sebep olabilir. Bölme sayısı gözlem sayısına yaklaştıkça bölmelere düşen gözlem sayısı bire yaklaşır ayrıklaştırma anlamsızlaşmaya başlar. Bu nedenle verinin büyüklüğü, dağılımı ve kullanım amacı dikkate alınarak orta düzeyde bir bölme sayısı belirlenmelidir. Eşit genişlikli ve eşit frekanslı bölme yöntemleri Şekil 2.6'da aynı veri seti

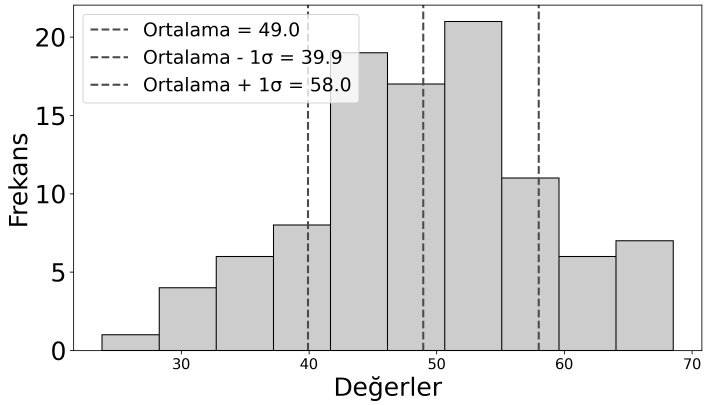
üzerinde örneklendirilmiştir.



Şekil 2.6: Eşit genişlikli ve eşit frekanslı bölme yöntemleri görsel olarak karşılaştırılmıştır

Şekil 2.6'da solda grafikte bölme aralıklarının eşit olmasına rağmen her bölüme düşen gözlem sayısının değişiklik gösterdiği görülmektedir. Sağdaki grafikte ise eşit frekanslı bölme yöntemi kullanılmış ve her sınıfta yaklaşık eşit sayıda gözlem bulunması sağlanmıştır. Bu durum, özellikle dengesiz dağılımlarda eşit frekanslı bölmenin daha dengeli ve karşılaştırılabilir gruplar elde etmeye olanak tanıdığını göstermektedir.

İstatistiğe dayalı bölme yönteminde ise verinin dağılımı gözönünde bulundurularak ortalama, medyan veya standart sapma gibi ölçekler aracılığı ile kesme noktaları belirlenir. Dolayısı ile bu yöntem verinin dağılım özelliklerine duyarlıdır. Mesela ortalama değer bir ölçüt olarak kullanılırsa gözlemler ortalamanın altı ve üstü olarak iki sınıfa ayrılabilir. Bu noktada unutulmamalıdır ki bu gibi bir yaklaşımın doğru çalışabilmesi için verinin dağılımının doğru tahmin edilmiş olması gerekir. Normale yakın dağılmış veride ortalama ölçütü etkinken dağılımın çarpık olduğu durumlarda sonuçlar beklenilenden farklı olabilir. İstatistiğe dayalı bölme yöntemine dair bir örnek Şekil 2.7'de verilmiştir.



Şekil 2.7: İstatistiksel bölme örneği üzerinde ortalama ve standart sapma sınırları işaretlenmiştir.

Son olarak, ayrıklaştırmanın en önemli alternatif yöntemlerinden birisi de uzman bilgisi, klinik eşikler veya kılavuzlar doğrultusunda veriyi kategorize etmektir. Bu bilgiye dayalı bir yaklaşım olduğundan pratik anlamda en güçlü yöntemlerden birisidir. Mesela vücut kitle endeksi (VKİ) gözönüne alınarak gerçekleştirilen obezite kategorizasyonu bu yönteme dair tipik bir örnek olabilir. Alternatif yaklaşımlar arasında seçim yapılırken, çalışmanın amacı ve veri setinin özellikleri mutlaka gözönünde bulundurulmalıdır. Tanımlayıcı analizlerde eşit genişlik veya eşit frekans tercih edilirken, keşif amaçlı çalışmalarda istatistiksel eşikler, klinik uygulamalarda ise uzman bilgisine dayalı kategorizasyon tercih edilebilir.

2.2.3 Karma Veriler

Gerçek hayat uygulamalarının pek çoğunde veri kümeleri tek tipte veriden oluşmamaktadır. Sayısal, kategorik, sıralı ve metin tabanlı değişkenler aynı veri kümesi içinde gözlemleri birlikte karakterize ediyor olabilirler. Bu tür veri kümeleri karma veri olarak adlandırılmaktadır. Analiz stratejisi belirlenirken her bir değişken tipinin kendine özgü yapısını dikkate almayı gerektirmektedir. Karma yapılı veri kümeleri analiz edilirken farklı veri tiplerinin aynı kümeleme veya sınıflandırma algoritması ile uyumlu olacak biçimde

dönüştürülmesi gerekebilir. Ayrıca veri analize hazırlanırken sayısal verilerin ölçeklerine, kategorik verilerin uygun kodlanmasına ve ordinal verilerin doğru biçimde modellenmesine dikkat edilmesi gerekir. Karma verinin analize hazırlanmasını örneklendirmek üzere sentetik bir karma veri kümesi hazırlanarak Tablo 2.6'te gösterilmiştir.

Tablo 2.6: Karma verinin orijinal hali

Yaş	**Cinsiyet**	**Eğitim**
25	Kadın	İlkokul
32	Erkek	Lise
47	Erkek	Üniversite
51	Kadın	Lise
62	Erkek	Üniversite

Karma veri setlerinin içerdiği farklı tiplerdeki değişkenlerin aynı analizde kullanılabilmeleri için uygun dönüşümlere ihtiyacı vardır (Liu vd., 2024). Buna uygun olarak Tablo 2.6'te görülen örnek veri setinde yer alan yaş, cinsiyet ve eğitim düzeyi değişkenleri ön işleme adımlarından geçirilmiştir. Cinsiyet, ikili biçimde kodlanmış, eğitim düzeyi ordinal yapısına uygun olarak sayısal değerlere dönüştürülmüş, yaş değişkeni ise belirli aralıklara ayrılarak kategorilere (genç, orta, yaşlı) ayrılmıştır. İşlemler sonrası elde edilen yeni veri Tablo 2.7'da gösterilmiştir.

Tablo 2.7: Karma verinin ön işleme sonrası dönüştürülmüş hali

Yaş Grup	**Cinsiyet**	**Eğitim**
1	1	1
1	0	2
2	0	3
2	1	2
3	0	3

Tablo 2.7'da bütün değişkenler sayısal biçimde temsil edilmiştir ancak unutulmamalıdır ki hala farklı veri tiplerini yansıtmaktadırlar. Buna göre yeni tabloda cinsiyet nominal, eğitim ordinal, yaş gruplamaları ise kategorik veri tiplerine örnektir (0=Erkek, 1=Kadın; 1=İlkokul, 2=Lise, 3=Üniversite; 1=Genç, 2=Orta, 3=Yaşlı). Bu noktada kümeleme yapılırken doğrudan standart benzerlik öl-

çütleri kullanılarak gözlemlerin değerlendirilmesi değişkenler arasındaki anlamlı ilişkilerin kaybedilmesine ve hatalı yorumlamalara sebep olabilmektedir. Bu nedenle karma veri setleri için veri tipi farklılıklarını dikkate alan özel yöntemler (Gower mesafesi gibi) tercih edilmelidir.

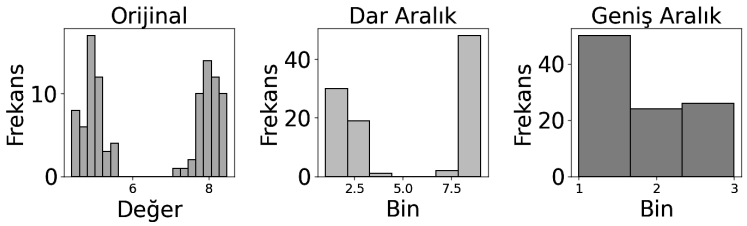
2.3 Veri ön işleme

Pratik saha uygulamalarında veri toplandığı ve doğrudan depolandığı hali ile analize hazır olmayabilir. Bu nedenle genellikle ham (raw) veri olarak adlandırılmaktadır. Verinin herhangi analizden önce bazı işlemlere tabi tutularak hazırlanması işlemine ön işleme (pre-processing) adı verilir. Araştırmacının bu safhadan gürültünün azaltılmasını, aykırı değerlerin belirlenmesini ve verilerin uygun ölçeklere getirilmesini beklemelidir. Bu sayede analizin doğruluğunun ve güvenilirliğinin artırılması hedeflenmektedir.

2.3.1 Gürültü

Ham veri kümelerinde hatalı girişlerden, ölçüm hatalarından, ekipman/enstrüman hata aralıklarından veya kayıttaki teknik sorunlardan kaynaklanan beklenmedik değerler bulunuyor olabilir. Verideki rastlantısal ve sistematik olmayan sapmalar gürültü olarak tanımlanmaktadır. Gürültülü veriler özelliklerin sınıfları karakterize etme yeteneklerini ve gözlemlerin sınıflara göre dağılımlarını bozarak analiz sonuçlarında hatalara sebep olabilmektedir. Bu nedenle analiz öncesi mümkün olduğunca temizlenmeleri hedeflenmelidir. Gürültünün kaynağına ve doğasına bağlı olarak farklı temizleme stratejileri belirlemek mümkündür. örneğin sensör okumaları gibi sürekli ve ardışık alınan veriler için hareketli ortalama veya medyan filtreleme uygulanabilirken, anket çalışmaları ile oluşturulmuş veri kümelerinden tutarsız cevapların temizlenmesi faydalı olabilir. Metin tabanlı verilerde dil ve anlam denetimi kaba bir gürültü ayıklama olarak görülebilir. Bu noktada gürültüden arındırma işleminin otomatikleştikçe bilgi kaybına da neden olabileceği unutulmamalıdır. Bu nedenle uygulanan metodoloji dengeli bir yaklaşımla kurgulanmalıdır. Mesela sürekli veri belli binlere bölüne-

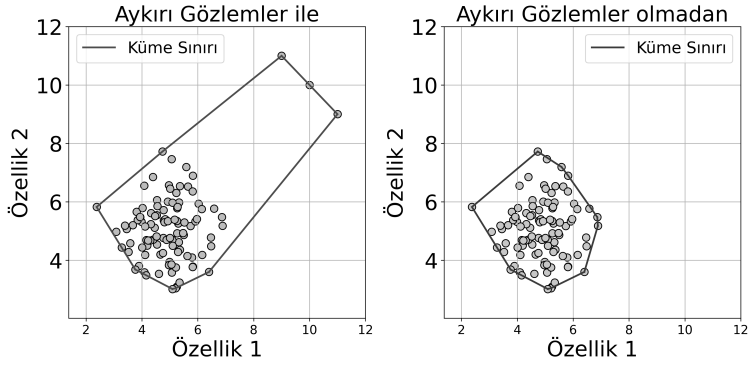
rek ayrıklaştırılırsa seçilen aralıktaki gürültü sayılabilecek önemsiz varyans yok edilmiş olur ancak aralıklar büyüdükçe anlamlı bilginin kaybolma ihtimali de artacaktır. Şekil 2.8'de aynı veri kümesi farklı bin genişlikleriyle ayrıklaştırılmış ve gürültü–bilgi kaybı ilişkisi görselleştirilmiştir.



Şekil 2.8: Sürekli bir veri kümesinin farklı bin genişlikleriyle ayrıklaştırılması görselleştirilmiştir.

2.3.2 Aykırı değerler

Aykırı (outlier) değerler veri kümesinin genel dağılımından belirgin şekilde farklılaşmış gözlemlere verilen isimdir (Kwak ve Kim, 2017). Bu gibi gözlemler ölçme hatası veya gürültü olabileceği gibi kendi sınıfının ender bir uç örneği de olması muhtemeldir. Sınıfının genel özelliklerinden oldukça ayrıştığı için ortalama, varyans gibi istatistiksel ölçütlerin sınıfı doğru nitelemesine engel olabilmektedirler. Aynı şekilde kümeleme algoritmalarını da yanıltarak sınırları kaydırmaları veya küçük kümelere sebep olmaları muhtemeldir. Şekil 2.9'da aynı veri kümesi aykırı gözlemler ile ve aykırı gözlemler olmadan gösterilmiştir. Şekilde aykırı değerler kümenin kapalı sınırını genişletmekte ve veri kümesinin gerçek yapısının olduğundan farklı görünmesine neden olmaktadır.



Şekil 2.9: Aykırı değerlerin küme sınırına etkisi saçılım grafiği üzerinde gösterilmiştir.

Sistematik yöntemler ile bu gibi gözlemleri veri setinde tespit edip çıkarmak veya işaretlemek mümkündür. Ancak nadir ama anlamlı bir gözlemin filtrelenmeden ayrı analiz edilmesi araştırma sonuçları için önemli olabilir. Buna örnek olarak nadir bir hastalık bulgusunun veya semptomun tespiti gösterilebilir.

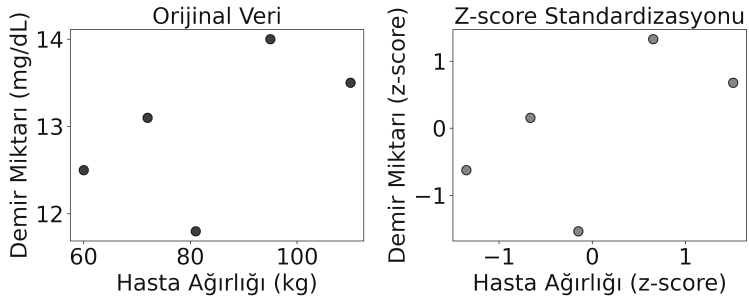
2.3.3 Standardizasyon ve Normalizasyon

Veri kümelerinin işlenmesini zorlaştıran etkenlerden birisi de değişkenlerin farklı aralıklarda ve büyüklüklerde değer alabilmesidir. Örneğin aynı veri kümesinde hasta ağırlığı ve kandaki demir miktarı gibi iki değişken birlikte analiz edilmek istendiğinde, değişkenler sınıfları farklı ölçeklerde karakterize edeceklerdir. Hasta ağırlığı onlarca kilogramla ifade edilirken, demir miktarı çoğunlukla ondalıklı değerler şeklinde ölçülmektedir. Bu görece büyüklük farkı nedeniyle sınıflar arasındaki anlamlı ayrımlar otomatik kümeleme algoritmalarında da doğrudan yürütülen araştırmalarda da ortaya çıkarılamayabilmektedir. Bu nedenle özelliklerin aynı ölçekte karşılaştırılabilmesi için dönüştürülmeleri pratik uygulamalarda önemli bir adımdır. Yaygın kullanılan yöntemlerden birisi Z-score standardizasyonu veya başka bir deyişle standardizasyondur. Bu yöntem verilerin ortalaması sıfır ve standart sapması bir olacak şekilde dönüştürülmesine dayanmaktadır (Milligan ve Cooper, 1988). Bunun sağlanması için her bir gözlem değeri, ait olduğu de-

ğişkenin ortalamasından çıkarılıp standart sapmasına bölünmektedir. Böylece farklı ölçeklerdeki değişkenler aynı standart ölçeğe indirgenmiş olur. Burada dikkat edilmesi gereken husus dönüşüm sonrası gözlemlerin değişim ölçekleri standart hale gelmiş olsa da sınıflar arası ilişkilerin ve örüntülerin birebir korunuyor olmasıdır. Z-score standardizasyonunun formülü aşağıdaki gibidir:

$$z = \frac{x - \mu}{\sigma} \tag{2.1}$$

Burada x gözlemi, μ değişkenin ortalamasını ifade ederken, σ standart sapmayı ifade etmektedir. Özellikle dağılımların normal kabul edildiği analizlerde ve uzaklık temelli algoritmalarda uygulanması analizin doğruluğu açısından kritik olabilir. Standardizasyon işleminin etkisi Şekil 2.10'da gösterilmiştir. Şekilde gözlemlerin numerik temsilinin değişmesine rağmen aralarındaki mesafe ve konum ilişkisinin korunduğu açıkça görülmektedir.



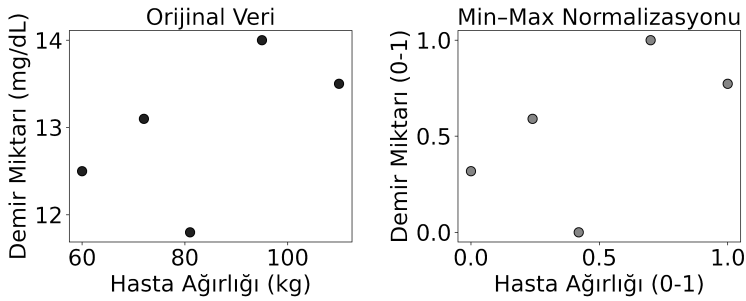
Şekil 2.10: Örnek bir veri kümesi üzerinde Z-score standardizasyonunun etkisi gösterilmiştir.

Bir diğer ölçeklendirme yaklaşımı normalizasyon işlemidir. Bu işlemde değişkenler belirli bir aralığa yeniden ölçeklendirilmektedirler (Al Shalabi vd., 2006). Farklı tipleri bulunmakla birlikte yaygın kullanılan tipi Min/Max normalizasyonudur. Veriler işlemden sonra genellikle 0 ve 1 aralığına sıkıştırılmaktadır. Bunun sağlanması için her bir gözlem değeri, değişkenin minimum değeri çıkarıldıktan sonra yine aynı gözlemin maksimum ve minimum değerler farkına bölünür. Böylece en küçük değer 0, en büyük değer 1 olacak şekilde yeniden ölçeklendirme yapılmış olur. Aşağıdaki gibi

formülize etmek mümkündür:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2.2}$$

Burada x gözlemi, $x_{\min}$ ilgili değişkenin minimum değerini, $x_{\max}$ ise maksimum değerini ifade etmektedir. Normalizasyon sonucunda gözlemler $[0, 1]$ aralığında kalırken, gözlemler arasındaki sınıfsal farklılıklar ve ilişkiler korunmuş olur. Normalizasyon işleminin etkisi Şekil 2.11'de gösterilmiştir. Şekilde standardizasyon işlemi ile benzer şekilde gözlemlerin aralarındaki mesafe ve konum ilişkisinin korunduğu görülmektedir. Ancak bu sefer bütün gözlem değerlerinin 0 ve 1 arasında değişmekte olduğuna dikkat edilmelidir.



Şekil 2.11: Örnek bir veri kümesi üzerinde Min/Max normalizasyonun etkisi gösterilmiştir.

Bu noktada her iki yaklaşımın da değişkenleri ortak bir ölçeğe taşımayı amaçladığı görülmektedir. Ancak yöntemlerin uygulanma biçimleri ve elde edilen sonuçlar farklıdır. Z-score standardizasyonunda veriler ortalama etrafında standart sapmaları bir olacak şekilde yeniden ölçeklenmekteyken, Min/Max normalizasyonunda tüm değerler belirli bir aralığa sıkıştırılmaktadır. Bu nedenle hangi yöntemin kullanılacağına, veri setinin yapısına ve seçilen algoritmanın özelliklerine göre karar verilmelidir. Uzaklık temelli yöntemlerde (KNN, k-means, PCA gibi) değişkenlerin dağılım özelliklerinin korunması gerektiğinden Z-score standardizasyonu uygunken sinir ağları veya gradyan tabanlı yöntemlerde, özellikle farklı ölçeklerde değişkenlerin bulunduğu durumlarda normalizasyon daha

uygun bir seçim olabilir.

2.4 İstatistiksel kavramlar

2.4.1 Ortalama

Ortalama kavramı yaklaşılan açıya göre farklı perspektiflerden işlevsellik kazanabilecek bir matematiksel ölçüttür. Küme yapısı açısından bakıldığında bir sınıf gözlemin eğilimlerini anlamak ve karakterini özetleyerek sayısal olarak temsil etmek için faydalı olabilir ve bir kümedeki gözlemlerin merkezi olarak kabul edilmektedir. Buna ek olarak ardışık gözlemler söz konusu olduğunda ortalama işlemi belli bir pencere aralığındaki gözlemlerin eğilimini göstererek gürültü veya ayrık değerlerden kaynaklı sapmaları yumuşatmak için etkin olabilmektedir. Ayrıca bir çok ölçümün özetlenerek bir gözlemi temsil edecek tek bir özellik değerine dönüştürülmesinde de etkin şekilde faydalanılabilir. En yaygın biçimi aritmetik ortalamadır ve tüm gözlemlerin toplamının gözlem sayısına bölünmesiyle elde edilir:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.3}$$

Burada x_i gözlemleri, n ise toplam gözlem sayısını ifade etmektedir. Aritmetik ortalama merkezi bir karakterizasyon için faydalı olmakla birlikte işleme sokulan gözlemler arasındaki aykırı değerlere karşı oldukça hassastır bu nedenle, özellikle uç değerlerin bulunduğu dağılımlarda baskın eğilimleri tam olarak temsil etmeyebilir. Bu noktada geometrik ortalama verilerin çarpımsal bir yapıya sahip olduğu (büyüme oranları, katlanarak artan değerler gibi) durumlar için alternatif olarak gösterilebilir. Şu şekilde tanımlanmaktadır:

$$x_g = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} \tag{2.4}$$

Burada x_i gözlem değerlerini, n ise toplam gözlem sayısını ifade etmektedir. Bu formülasyona göre örneğin bir tümörün boyunun

üç gün boyunca sırasıyla 2, 3 ve 4 kat arttığını farz edersek, artış oranlarını toplamak yerine çarparak özetlemek ortalama büyüme katsayısını verecektir. Bu örneği matematiksel olarak şöyle gösterebiliriz:

$$x_g = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n} \tag{2.5}$$

Yani geometrik ortalama tüm değerler çarpımının n'inci kökü ile hesaplanmaktadır.

Bir başka alternatif yaklaşım olan harmonik ortalama ise oranlar veya hızlar gibi pozitif büyüklüklerin değerlendirilmesinde kullanılmaktadır ve şu şekilde tanımlanmaktadır:

$$x_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{2.6}$$

Burada x_i her bir gözlemi, n toplam gözlem sayısını ifade eder.

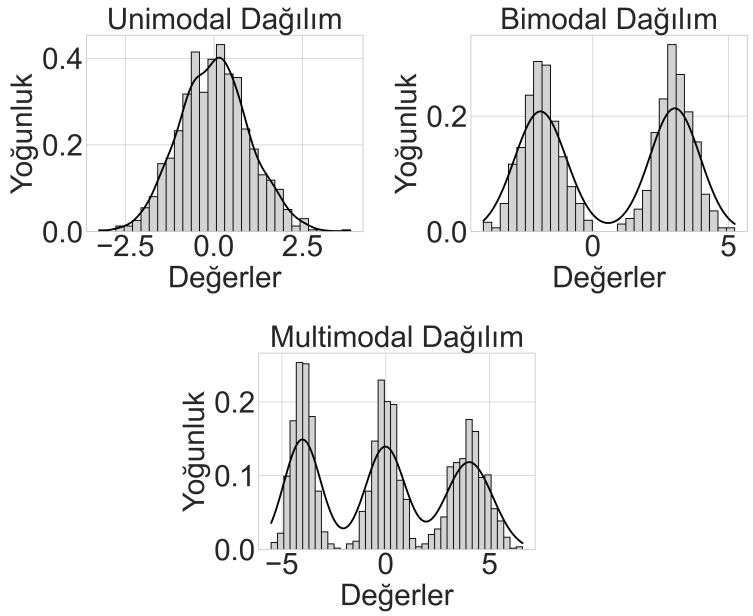
Bu alternatif ortalama bulma yaklaşımları verilerin doğasına göre farklı miktarlarda etkinlik gösterebilirler. Bu noktada gözlemlerin dağılımının incelenmesi doğru yaklaşımın seçimi için kritik olabilir. Unutulmamaldır ki dağılımın simetrik olmadığı durumlarda alternatif yaklaşımlar arasında belirgin farklar gözlenebilir.

2.4.2 Medyan ve Mod

Ortalama kavramı kümelerin genel karakterini özetlemek için faydalı bir araç olmakla birlikte gürültü, ayrık değerler ve veride bozunmaların varlığı başka yaklaşımlara olan gereksinimi artırmaktadır. Medyan kavramı da perspektif sunan alternatif yaklaşımlardan birisidir. Bu yaklaşım kabaca gözlemlerin değerlerine göre sıralanarak ortaya denk düşen gözlem değerinin seçilmesine dayanmaktadır. Gözlem miktarı tek sayı ise kümeyi ortadan ikiye bölen gözlemi seçmek kolaydır. Eğer çift sayı ise bu sefer ortadaki iki gözlem seçilerek ortalaması alınmalıdır. Bu sayede gözlemlerin sıralanmalarına dayanılarak kümelerin merkez eğilimi gözlenmektedir.

Mod kavramı ise özellikle kategorik veri tiplerinde gözlemlerin yoğunlaştığı kategorinin tespiti için tercih edilen yaklaşımlardan

birisidir. Buna göre en fazla miktarda gözlemin aldığı kategorik değer mod değeri olarak kabul edilmektedir. Yani kategorik veriler için frekans tablosu veya çubuk, sürekli veriler için histogram grafiklerinde görülen bir veya daha fazla tepe nokta mod değerini gösteriyor diyebiliriz. Eğer histogramda tek tepe nokta var ise dağılım unimodal, iki tepe var ise bimodal, ikiden fazla tepe var ise multimodal olarak adlandırılmaktadır. Mod yorumu histogram ile birlikte çizilen yoğunluk eğrisi üzerinden de gerçekleştirilebilir. Unimodal, bimodal veya multimodal karakterde dağılım örnekleri Şekil 2.12'de verilmiştir.



Şekil 2.12: Unimodal, bimodal ve multimodal dağılımlar görsel olarak örneklendirilmiştir.

2.4.3 Varyans ve Standart Sapma

Veri kümesindeki gözlemleri merkezleyerek özetlemeye yarayan ortalama kavramı tek başına veri kümesinin karakterini nitelemeye yetmeyebilmektedir. Bu noktada gözlemlerin bu merkez etrafında ne derece dağıldığını ölçerek özete katkıda bulunan temel istatistik ölçütleri varyans ve standart sapmadır. Bu tanıma göre varyans

ve standart sapma merkez etrafında gözlemlerin birbirinden ne kadar farklılaştığını karakterize edebilmektedir. Tıbbi veri analizinde kümeler arasındaki değerlendirilmesi ve kümelerin nitelenebilmesi için bu ölçütler etkin şekilde kullanılmaktadır. Varyans matematiksel olarak gözlemlerin ortalamaya olan farklarının karesinin ortalaması olarak tanımlanmıştır. Şu şekilde formülize edilir:

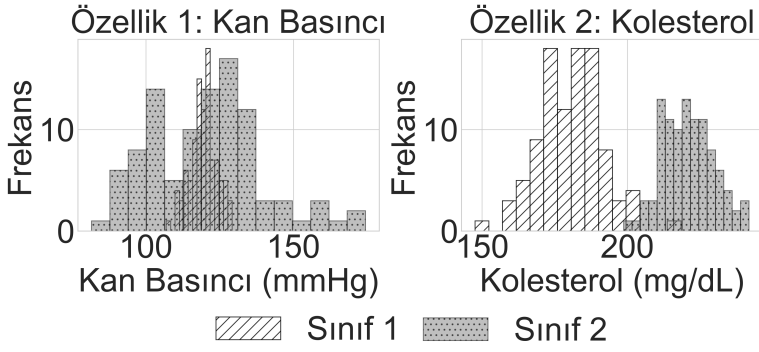
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.7}$$

Burada x_i gözlemleri, $\bar{x}$ ortalamayı, n gözlem sayısını ve σ^2 varyansı ifade etmektedir.

Standart sapma ise matematiksel olarak varyansın karekökü olarak tanımlanmıştır:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.8}$$

Bu formulasyona göre standart sapma varyansın bir fonksiyonudur ancak verilerle aynı birime sahiptir. Bu nedenle genellikle saçılımın ve farklılaşmanın ölçütü olarak standart sapma tercih edilmektedir. Standart sapmanın sınıf dağılımları ve küme ayrılabilirliği üzerindeki etkisi Şekil 2.13'de gösterilmiştir.



Şekil 2.13: Standart sapmanın sınıf dağılımlarına ve ayrılabilirliğe etkisi örnek bir veri kümesi üzerinde görülmektedir.

Şekil 2.13'de kan Basıncı (Özellik 1) panelinde sınıfların ortalamaları benzerdir ve farklı standart sapmalara sahip olmalarına

rağmen sınıflar üst üste binmiş olarak görülmektedir. Kolesterol (Özellik 2) panelinde ise sınıfların ortalamaları birinden farklı, standart sapmaları ise düşüktür bu sayede sınıfların ayrılabilirliği yükselmiştir.

2.4.4 Normallik ve Normallik Testleri

Veri kümesini oluşturan gözlemlerin dağılımlarının istatistiksel olarak karakterize edilebiliyor olması veri analizi için önemli bir perspektif sunmaktadır (Habibzadeh, 2024). Ayrıca kestirim ve sınıflandırma problemlerinin çözümünü kolaylaştırmaktadır. Çünkü gözlemlerin hangi oranda hangi değerlerde toplanacağını kestirmek mümkün olabilmektedir. En çok bilinen ve kullanılan istatistiki dağılım modellerinden birisi de normal dağılımdır. Bu dağılım karakterinde gözlemler ortalama değer etrafında simetrik bir çan eğrisi (veya Gauss eğrisi) şeklinde kümelenmektedir. Pek çok parametrik yöntem gözlemlerin normal karakterde dağıldığı varsayımı gerçek ise etkin şekilde kullanılabilmektedir. Örneğin ANOVA, t-test, Pearson korelasyon katsayısı veya Gaussian Naive Bayes gibi yaklaşımların uygulanabilmesi için gözlemlerin normal dağılım sergiliyor olmaları gerekebilmektedir. Normal dağılım varsayımı daha fazla gözlemin ortalamaya yakın değerlerde toplanırken, ekstrem değerlere doğru gittikçe gözlem sayısının düşeceği öngörüsü üzerine kurulmuştur. Pratik uygulamalarda normal dağılımı mükemmel hali ile nadir görülse de normale yakın benzer dağılımları gözlemek mümkün olabilmektedir.

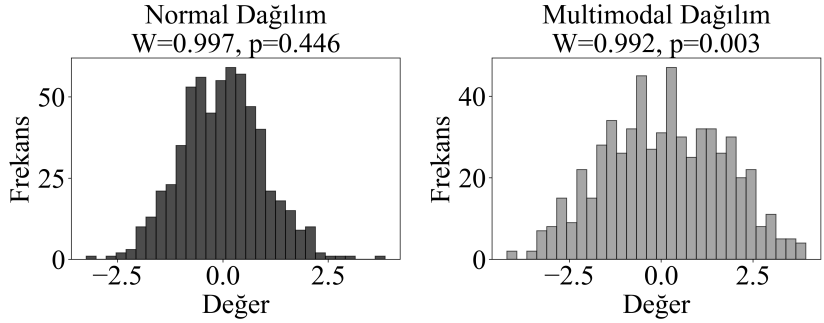
Normal dağılım varsayımının test edilmes için histogram veya Q-Q grafikleri gibi görselleştirme yaklaşımları etkin şekilde kullanılabilmektedir. Görsel incelemeye ek olarak matematiksel olarak derecelendirebilmek adına pek çok yaklaşım bulunmaktadır. Bunlardan en yaygın kullanılanlardan birisi de Shapiro–Wilk testidir (González-Estrada vd., 2022). Bu test özellikle görece küçük veri kümelerinin normal dağılım varsayımını test etmek için tercih edilmektedir (Shapiro ve Wilk, 1965). Test çıktı olarak bir w istatistiği ve bir p değeri verir ve w değeri genellikle 0 ile 1 arasında değişmektedir. Bu değer 1'e yaklaştıkça normal dağılım uyumluluğu artmaktadır. Bununla birlikte p değeri büyüdükçe normal dağılım

varsayımı güçlenmektedir. Test şu hipotezleri test etmektedir. H_0: Veri normal dağılım karakterindedir. H_1: Veri normal dağılım karakterinde değildir. Testi şu şekilde formülize etmek mümkündür:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.9}$$

Verilen formülasyonda $x_{(i)}$ küçükten büyüğe sıralanmış gözlemleri, $\bar{x}$ gözlemlerin ortalamasını, a_i katsayıları ise normal dağılımdaki sıralanmış değerlerin ortalamalarından çıkarılan sabitleri ifade etmektedir. Örneklem büyüklüğü arttıkça test hassasiyetinin artarak küçük sapmaların bile anormal sayılabileceği unutulmamalıdır.

Normallik testlerinin bir gözlem kümesinde tek bir normal dağılım varsayımını değerlendirdiği unutulmamalıdır. Yani multi modal bir dağılımda, klasik normallik testi yaklaşımları genellikle normal dağılım varsayımını reddedecektir. Çünkü tüm verinin tek bir çan eğrisine oturup oturamadığı analiz edilmeye çalışılmaktadır. Şekil 2.14'te normallik testinin unimodal ve multimodal yapılarda nasıl sonuç vermekte olduğu görselleştirilmiştir.



Şekil 2.14: Normal dağılım ve multimodal dağılım için Shapiro–Wilk test sonuçları histogram üzerinde gösterilmiştir.

Bölüm 3

Kümelerin Özellikleri

3.1 Uzaklık

Veri kümelerinin karakterize edilmesi, temsil ettikleri gruplar arasındaki ilişkilerin ve gözlemlerin oluşturduğu arama uzayının incelenmesi açısından kritik bir adımdır. Bu inceleme sayesinde, veri uzayında yer alan örüntüler ve ilişkiler ortaya çıkarılabilir.

Karakterizasyonun dayandırılabileceği pek çok matematiksel ölçüt olmakla birlikte gözlem uzayındaki görece konum, benzerliğin ve ayrılabilirliğin ölçülebilmesi adına temel parametrelerden birisidir. Bu noktada iki gözlem arasındaki uzaklığın ölçülerek kıyaslanabilmesi büyük önem kazanmaktadır.

iki boyutlu bir düzlemde uzaklık kavramını ölçmek ve değerlendirmek görece daha basit olmakla birlikte gözlemlerin çok miktarda özellikle temsil edildiği çok boyutlu uzaylarda uzaklığın hesaplanması boyutlar arttıkça giderek daha karmaşık hale gelmektedir. Bu nedenle, klasik Öklidyen uzaklık ölçütüne ek olarak farklı varsayımlara dayanan pek çok alternatif uzaklık metriği geliştirilmiştir.

3.1.1 Öklidyen Uzaklığı

En yaygın uzaklık metriklerinden birisi olan öklid uzaklığı (L2 uzaklığı olarak da adlandırılır) iki nokta arasındaki en kısa (doğrusal) mesafeyi ölçmektedir. Sürekli ve nicel veriler için uygun bir ölçüt olan bu metrik aşağıdaki formül ile ifade edilir:

$$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3.1}$$

Bu formülde x_1, y_1,x_2,y_2 iki boyutlu bir uzayda iki noktanın koordinatlarını ifade etmektedir. Örneğin yaş ve prostat hacmi gibi her hastanın iki özellikle karakterize edildiği bir uzayda iki gözlem arasındaki öklid uzaklığı yaş ve prostat hacmi farklarının karelerinin toplamının karekökü alınarak hesaplanır.

Gerçek hayat uygulamalarında ise veri kümesinin çok daha fazla özellik içermesi muhtemeldir. Böyle bir durumda Bu durumda formül, p boyutlu bir uzay için aşağıdaki şekilde genelleştirilebilir:

$$d(X,Y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \tag{3.2}$$

Burada X ve Y iki gözlemi (vektörü), p ise veri kümesindeki özellik sayısını temsil etmektedir (Malkauthekar, 2013). Bu genelleştirmenin uygulanışını örneklendirmek amacı ile 3 hastadan oluşan bir veri seti yaratılmıştır. Tablo 3.1'de örnek veri seti görülmektedir.

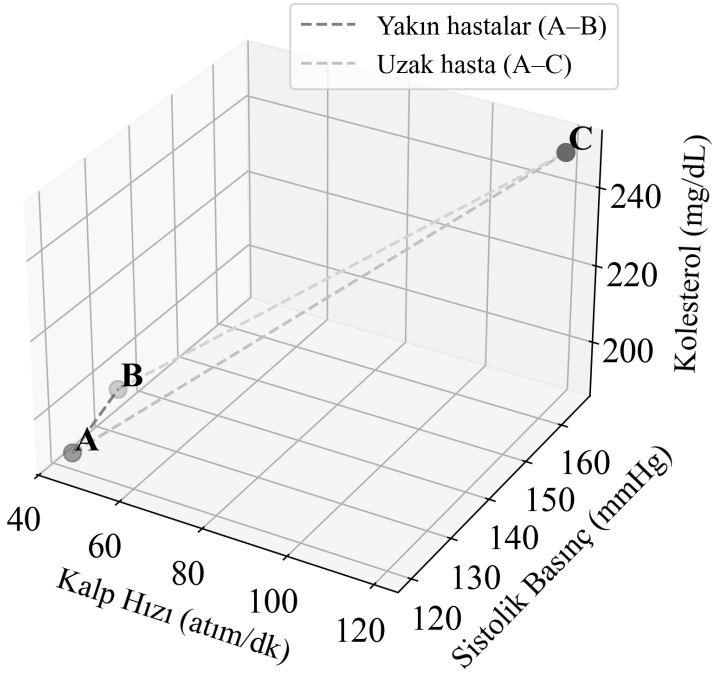
Tablo 3.1: Örnek tıbbi veri kümesi (3 gözlem, 3 özellik)

Hasta	**Kalp Hızı (atım/dk)**	**Sistolik Basınç (mmHg)**	**Kolesterol (mg/dL)**
A	45	120	190
B	48	128	200
C	120	165	250

3.1'de verilen örnek hasta grubunda her gözlem üç özellik ile temsil edilmektedir. Buna göre gözlemler arası öklidyen uzaklığı şu açık formül ile hesaplanmıştır:

$$d(X,Y) = \sqrt{(\mathrm{KH} - \mathrm{KH}')^2 + (\mathrm{SB} - \mathrm{SB}')^2 + (\mathrm{KOL} - \mathrm{KOL}')^2} \tag{3.3}$$

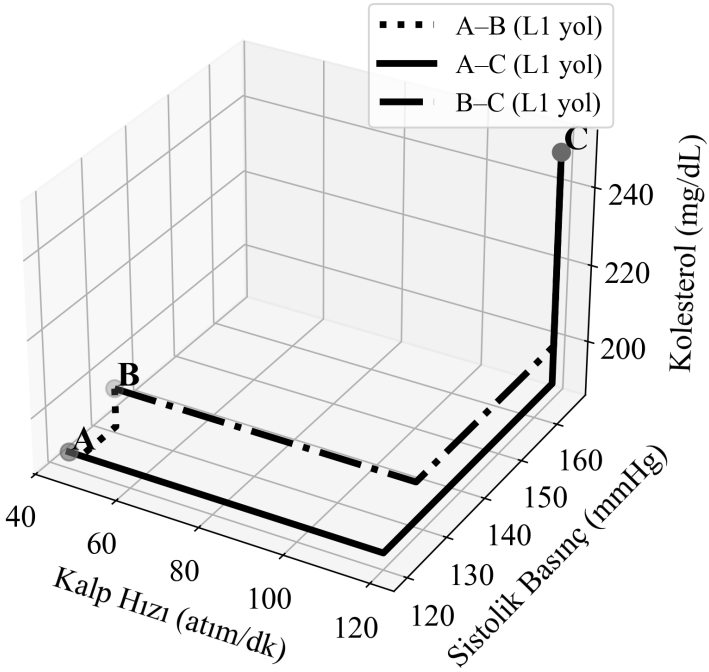
Yani iki gözlemin ilgili özelliklerinin farklarının karelerinin toplamının karekökü hesaplanmaktadır. Buna göre özellik uzayına yerleştirilmiş gözlemler Şekil 3.1'de gösterilmiştir:



Şekil 3.1: Üç hastanın üç özellikli özellik uzayındaki konumları gösterilmiştir.

3.1.2 Manhattan Uzaklıklığı

L1 uzaklığı olarak da bilinen bu metrik gözlem koordinatlarının mutlak farklarının toplanması ile hesaplanmaktadır (Suwanda vd., 2020). En kısa mesafeyi doğrusal bir çizgi olarak hesaplayan öklidyen uzaklığından farklı olarak birbirini dik kesen doğrusal yolların toplam uzunluğu olarak mesafeyi ölçmektedir. Tablo 3.1'de verilen örnek veri seti kullanılarak Manhattan mesafeleri hesaplanmış ve Şekil 3.2'de görselleştirilmiştir.



Şekil 3.2: Üç hastanın üç özellikli özellik uzayında aralarındaki Manhattan uzaklıkları gösterilmiştir.

İki boyutlu bir uzaydaki gözlemler için şu şekilde formülize edilebilir:

$$d_1(X, Y) = \sum_{i=1}^{p} |x_i - y_i|. \tag{3.4}$$

Bu formülü Tablo 3.1'de verilen örnek veri seti için genelleştirdiğimizde örnekler arasındaki Manhattan uzaklıklarını şu şekilde hesaplamak mümkün olacaktır:

$$d_1(X, Y) = |\mathrm{KH}_X - \mathrm{KH}_Y| + |\mathrm{SIS}_X - \mathrm{SIS}_Y| + |\mathrm{KOL}_X - \mathrm{KOL}_Y|. \tag{3.5}$$

Manhattan uzaklığı bulunurken kare hesaplaması yapılmadığından analiz ayrık gözlemlerden daha az etkilenebilmektedir. Bu gerçek hayat uygulamalarında önemli bir avantaj olabilir.

Bu metriğin de benzer uzaklık ölçütleri gibi ölçeğe duyarlı

olduğu unutulmamalıdır. Bu nedenle farklı birimlerde (mmHg, mg/dL gibi) özelliklerin bir arada olduğu veri setlerinde analiz öncesi normalizasyon veya standardizasyon uygulanması analizin doğruluğu açısından önemlidir.

3.1.3 Chebyshev Uzaklıklığı

Satranç tahtası ölçütü olarak da bilinen bu metrik Manhattan ve Öklidyen uzaklıklarından farklı olarak eksenler farklarını birbiri ile toplamaz. Bunun yerine bütün eksen farkları içinde en büyük olan çıktı olarak kabul edilir (Rodrigues, 2018). Yani iki gözlem arasındaki mesafeyi en büyük fark gösteren özelliğin üzerinden ifade etmektedir. Matematiksel olarak şu şekilde ifade edilir:

$$d_\infty(X_1, X_2) = \max_{j=1,\ldots,p} |x_{1j} - x_{2j}|. \tag{3.6}$$

bu formülasyonda X_1 ve X_2 iki farklı gözlemi (örneğin hasta A, hasta B gibi), x_{1j} ve x_{2j} ise her iki hastanın j özelliğini (örneğin kalp hızı, sistolik basınç gibi) ifade etmektedir. p ise toplam özellik sayısıdır. Tablo 3.1'de gösterilen veri setindeki A, B ve C hastalarının Chebyshev uzaklıkları verilen formülasyona uygun olarak aşağıda hesaplanmıştır.

$$\begin{aligned}
d_\infty(A,B) &= \max\left(|45-48|,\ |120-128|,\ |190-200|\right) \\
&= \max(3,\ 8,\ 10) = 10 \\
d_\infty(A,C) &= \max\left(|45-120|,\ |120-165|,\ |190-250|\right) \\
&= \max(75,\ 45,\ 60) = 75 \\
d_\infty(B,C) &= \max\left(|48-120|,\ |128-165|,\ |200-250|\right) \\
&= \max(72,\ 37,\ 50) = 72
\end{aligned} \tag{3.7}$$

Chebyshev uzaklığı hesabı toplama veya kare alma işlemi içermediği için görece işlem yükü daha azdır ancak yalnızca en büyük fark gösteren özellik kullanıldığından bu tek özelliğin baskın farkı tüm mesafeyi karakterize etmektedir. Bu nedenle ölçek ve normalizasyon/standardizasyon yine büyük önem kazanmaktadır. Veri

setinde farklı ölçeklerde özellikler söz konusu ise normalizasyon veya standardizasyon yapılması önemlidir.

3.1.4 Minkowski Uzaklıklığı

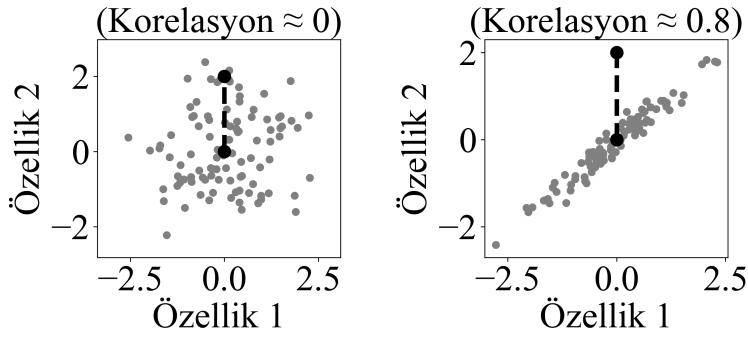
Bu ölçüt birden fazla metriği genelleştirerek mesafe ölçümünün esnekleştirilmesini sağlamaktadır. Buna göre uzaklık değeri iki gözlem arasındaki farkların belirli bir üstel parametreye (p) göre değerlendirilmesi ile hesaplanmaktadır (Rodrigues, 2018). n boyutlu bir gözlem uzayı için şu şekilde formulize etmek mümkündür:

$$d_p(X_1, X_2) = \left(\sum_{j=1}^{n} |x_{1j} - x_{2j}|^p \right)^{1/p} \tag{3.8}$$

Bu ifadede X_1 ve X_2 iki farklı gözlemi (örneğin hasta A, hasta B gibi), x_{1j} ve x_{2j} ise her iki hastanın j özelliğini (örneğin kalp hızı, sistolik basınç gibi) ifade etmektedir. p değeri, uzaklık hesabında farkların hangi ölçüde baskın olacağını belirleyen parametredir. Buna göre $p = 1$ olduğunda Manhattan, $p = 2$ olduğunda Öklidyen, $p \to \infty$ olduğunda ise Chebyshev uzaklıkları hesaplanmaktadır. Bu şu anlama gelir: p değeri arttıkça gözlemler arasındaki büyük farklar daha belirleyici hale gelirken, küçük farkların daha az etkili olduğu görülmektedir. Bu sayede metrik veri yapısına göre esnek mesafe ölçümü sağlayabilmektedir.

3.1.5 Mahalanobis uzaklığı

Öklidyen uzaklığı hesaplanırken değişkenler arası bağımlılık ilişkileri gözönüne alınmamaktadır. Ancak değişkenlerin birbiri ile korelasyon içinde olduğu durumlarda aksların birbirine dik olduğu varsayımı geçerliliğini yitirebilir. Bu gibi bir durumda ölçüt olarak Mahalanobis uzaklığı önerilmektedir. Buna göre eğer özellikler gerçekten birbirinden bağımsız ise öklidyen ve Mahalanobis mesafeleri birbirine oransal olarak eşit olmalıdır. Şekil 3.3'de akslar arasında korelasyon görülen ve görülmeyen iki farklı uzay içinde aynı gözlemin konumu gösterilmiştir.



Şekil 3.3: Korelasyonun varlığı, Öklidyen uzaklığın gerçek geometrik farklılığı yansıtamamasına neden olmaktadır.

Matematiksel olarak Mahalanobis uzaklığı aşağıdaki şekilde ifade edilir:

$$d_M(X_1, X_2) = \sqrt{(X_1 - X_2)^T S^{-1} (X_1 - X_2)} \tag{3.9}$$

Bu formülde X_1 ve X_2, uzaklığı hesaplanacak iki farklı gözlemi temsil etmektedir. Buna göre gözlem vektörleri, bir hastaya ait özellik değerlerini içeren veri matrisleridir. $(X_1 - X_2)$ gözlemlerin fark vektörünü, $(X_1 - X_2)^T$ ise bu farkın transpozunu (satır vektörü biçiminde) göstermektedir. S veri kümesindeki tüm değişkenlerin birlikte nasıl değiştiğini ifade eden kovaryans matrisidir (Brereton, 2015).

Kovaryans kavramı, iki değişkenin birlikte nasıl değiştiğini ölçmektedir. İki değişken aynı yönde artıp azalma gösteriyorsa pozitif, biri artıp diğeri azalıyorsa negatif kovaryans söz konusu olacaktır. Matematiksel olarak X_1 ve X_2 gibi iki değişken arasındaki kovaryans şu şekilde hesaplanmaktadır:

$$\mathrm{Cov}(X_1, X_2) = \frac{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{n - 1} \tag{3.10}$$

Burada X_{1i} ve X_{2i} her bir gözleme ait değerleri, $\bar{X}_1$ ve $\bar{X}_2$ ise bu değişkenlerin ortalamalarını göstermektedir. Yani her gözlemin gösterdiği ortalamadan sapma miktarı çarpılıp ortalaması alınmaktadır. Örnek veri kümesindeki üç değişken (örneğin kalp hızı, sistolik basınç, kolesterol gibi)için kovaryans matrisinin oluşturul-

ması istenirse bütün özellik çifti kombinasyonları için kovaryans hesaplanmalıdır. Buna göre aşağıdaki gibi bir matris biçiminde oluşturulacaktır:

$$S = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \mathrm{Cov}(X_1, X_3) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \mathrm{Cov}(X_2, X_3) \\ \mathrm{Cov}(X_3, X_1) & \mathrm{Cov}(X_3, X_2) & \mathrm{Cov}(X_3, X_3) \end{bmatrix} \tag{3.11}$$

Bu ifade her bir değişkenin kendi varyansı ile birlikte diğer değişkenlerle olan ilişkisini de göstermektedir. Mahalanobis uzaklığı hesabında bu matrisin tersi kullanılmaktadır. Matrisin tersini almak uzaklık hesabının değişkenler arası korelasyon ilişkilerinden bağımsız yapılabilmesini sağlamaktadır.

3.1.6 Kosinüs benzerliği

Küme aidiyetinin sorgulanması, gözlemlerin görece uzaklıklarının ölçülebilmesi ve küme iç yapılarının değerlendirilebilmeleri için önerilebilecek bir başka alternatif metrik de kosinüs benzerliğidir. Bu metrik gözlemleri temsil eden özellikleri vektörler olarak ele alıp aralarındaki açısal benzerliği ölçmektedir (Huang vd., 2008).

Özellik uzayının boyutundan bağımsız olarak her gözlemin bir nokta olarak kabul edilmesi mümkündür. Alternatif olarak gözlemleri orijin noktasından gözleme doğru çizilen bir vektör olarak da kabul etmek mümkün olabilir. Buna göre özelliklerin değer değişimine göre bu vektörün yönü ve büyüklüğü değişecektir. İki farklı gözlemi değerlendirmek istediğimizde bu sefer orijinden başlayıp iki farklı noktaya yönelen iki vektör söz konusu olacaktır. Dolayısıyla iki farklı gözlemi temsil eden iki vektör arasında bir açı ölçmek mümkündür. Bu açı, gözlemlerin benzerliğini, yani özellikler arasındaki göreli değişimi karakterize eden parametrelerden birisidir. Buna göre iki gözlemin benzerliği onları temsil eden özellikler benzer oranlarda artıp azalmasına göre artmakta veya azalmaktadır. Bu açı büyüdükçe gözlemler arasındaki yönsel benzerliğin azaldığı düşünülebilir. Kosinüs benzerliği sayısal olarak ifade edilmek üzere şu şekilde tanımlanmıştır:

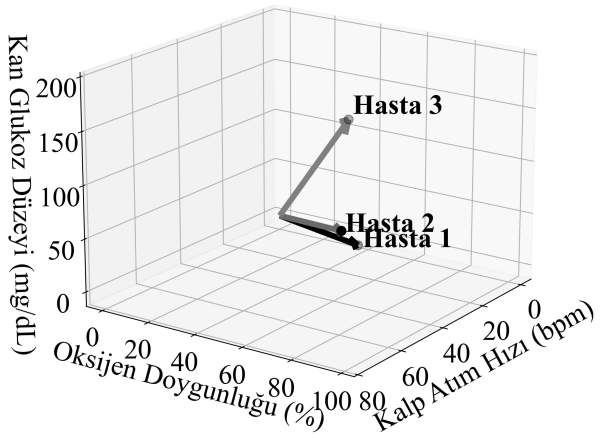
$$\text{Kosinüs Benzerliği}(X_1, X_2) = \frac{X_1 \cdot X_2}{\|X_1\|\|X_2\|} \tag{3.12}$$

Burada X_1 ve X_2 iki farklı gözlemi ifade etmektedir. Buna göre kosinüs benzerliği hesabı gözlemlerin noktasal çarpımları, normlarının çarpımına bölünerek gerçekleştirilmektedir. Formülasyonun sayısal sonucu, yönsel benzerliğin derecesini ifade etmektedir. Bu değerin 1'den çıkarılarak kosinus mesafesi olarak da kullanılması mümkündür. Buna göre Kosinüs uzaklığı değerinin 0'a yaklaşması gözlem benzerliğinin, 1'e yaklaşması ise farklılaşmanın arttığını gösterecektir. Bu sayede gözlemlerin yönsel benzerliği büyüklük farklarından etkilenmeden yönsel ilişkiler ele alınarak değerlendirilebilmektedir. Kosinüs benzerliğini örneklendirmek üzere üç hastadan ve üç özellikten oluşan bir veri kümesi yaratılarak Tablo 3.2'de verilmiştir.

Tablo 3.2: Üç hasta için yaratılan sentetik özellik değerleri ve diğer hastalar ile aralarındaki kosinüs benzerlikleri verilmiştir.

H	**Kalp(bpm)**	**O2(%)**	**Glukoz (mg/dL)**	**H2**	**H3**	**H1**
H1	72	98	90	0.997	0.917	-
H2	76	95	105	-	0.944	0.997
H3	65	88	190	0.944	-	0.917

Tablo 3.2'e göre hasta 3'ün artan kan glukozu diğer iki hastaya göre görece daha küçük kosinüs benzerliği ile karakterize edilmesine neden olmuştur. Hastaları temsil eden vektörler Şekil 3.4'de özellik uzayında görülmektedir.



Şekil 3.4: Üç gözlemin özellik uzayındaki vektörel temsili gösterilmiştir. Hastalardan birisi yüksek kan glukoz düzeyi nedeniyle diğer iki hastadan ayrıklaşmaktadır.

3.2 Kümelerin Yapısal Özellikleri

Veri kümeleri küme analizine tabi tutulduğunda yani gözlemler farklı özelliklerine veya karakteristiklerine göre bir araya gruplandığında ortaya çıkan yapı mercek altına alınan veriye dair içgörü sağlamaya adaydır. Veri kümeleri çeşitli dağılım özelliklerine göre gözlemlerin doğasına dair bilgi sağlayabilmektedir. Küme merkezleri bu açıdan önemli bir parametredir. Bir kümenin merkezi o kümeye ait gözlemlerin genel karakterini ortaya koyan bir genelleştirilme noktasıdır. Matematiksel olarak n gözlemden oluşan bir küme için merkez noktası şu şekilde ifade edilebilir:

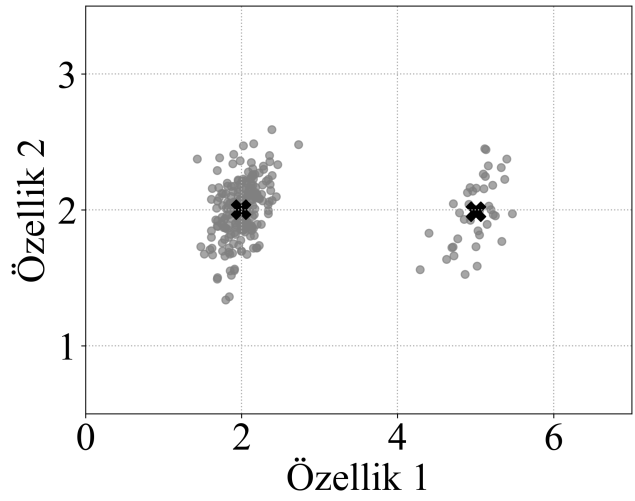
$$C_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \tag{3.13}$$

Burada merkez nokta her bir özellik için ayrı ayrı hesaplanmalıdır yani iki özellikten oluşan bir örnek uzayında merkez noktayı bulmak için iki ortalama nokta bulunmalıdır. Buna göre burada C_j

j'inci özelliğe ait merkez değerini, x_{ij} ise her bir gözlemin j'inci özelliğinin değerini temsil etmektedir. Küme merkezi gözlemlerin tekil farklılıklarından arınmış bir bir toplu eğilimi yansıtmaktadır.

3.2.1 Yoğunluk

Bir veri kümesinin yoğunluğu, küme merkezi çevresinde belirli bir alan veya hacim içerisinde yer alan gözlem sayısını ifade eder. Buna göre bu ölçüt gözlemlerin merkeze ne kadar yakın toplandığını göstermektedir. Bu noktada kaplanan toplam küme hacmi veya alanı önem kazanmaktadır. Örneğin iki küme de aynı hacimde yer kaplıyorsa gözlem sayısı fazla olanın yoğunluğu daha yüksek olacaktır. Ancak kümelerin kapladığı alan ya da hacim farklı ise yoğunluğu karşılaştırmak için toplam gözlem sayısı toplam alan ya da hacme bölünmelidir. Şekil 3.5'de farklı yoğunluklarda kümeler aynı özellik uzayı üzerinde gösterilmiştir.

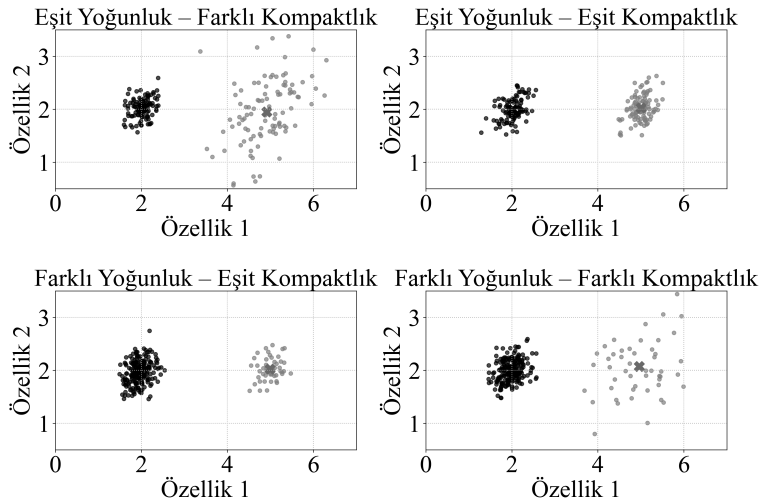


Şekil 3.5: Aynı gözlem uzayında farklı yoğunluklara sahip iki küme görülmektedir. Küme merkezleri çarpı ile gösterilmiştir.

3.2.2 Kompaktlık

Bu ölçüt bir veri kümesinin merkez çevresinde ne kadar saçıldığını nitelemektedir. Buna göre eğer iki veri kümesi eşit yoğunlukta ise

yüksek varyans kompaktlığın azalmasına neden olacaktır. Burada şuna dikkat edilmelidir: Yoğunluk yüksek olsa bile gözlemler birbirinden uzaksa kompaktlık düşük olabilmektedir. Kompaktlık ve yoğunluk özellikleri Şekil 3.6'da dört farklı senaryo üzerinden görselleştirilmiştir.

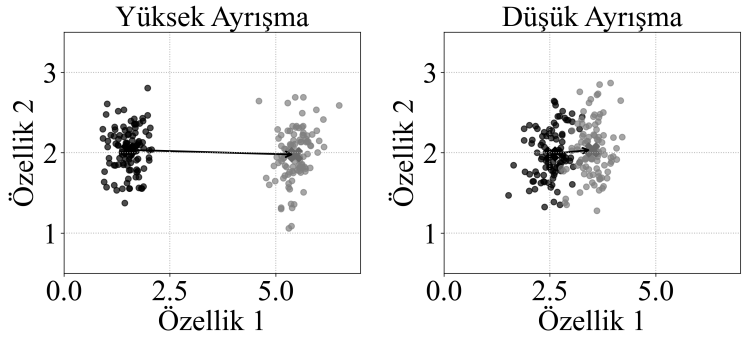


Şekil 3.6: Aynı gözlem uzayında farklı yoğunluk ve kompaktlık kombinasyonlarına sahip küme senaryoları görselleştirilmiştir. Küme merkezleri çarpı ile gösterilmiştir.

Şekil3.6'da küme yapısının genel görünümü bu iki ölçütün etkileşimiyle şekillendiği görülmektedir. Buna göre benzer sayıda gözlem içeren eşit yoğunluktan kümelerden birisinde gözlemler merkeze daha yakın toplanmışsa daha kompakt kümenin sıkı ve belirgin, diğerinin ise görece daha gevşek bir dağılım sergilemesi beklenebilir. Yoğunluk da kompaktlık da benzer ise gözlemler farklı konumlarda olsalar da kümelerin birbirine benzer bir bulut görüntüsünde olması muhtemeldir. Sadece kompaktlığın benzediği durum senaryolarında ise her iki küme benzer sıkılıkta olacaktır ancak gözlem sayısı fazla olan küme daha yoğun görünebilir. Eğer yoğunluk da kompaktlık da farklı ise birbirinden farklı karakterde gözüken kümeler oluşacaktır.

3.2.3 Ayrışma

Ayrışma Kümelerin gözlem uzayında birbirinden ne kadar ayrılabildiğini ya da ne kadar iç içe geçtiklerini tanımlayan özelliktir. Bu tanıma göre farklı kümelere ait gözlemler farklı küme merkezlerine yaklaşarak diğer küme gözlemleri ile kaynaştıkça ayrışma azalmaktadır. Bu azalma aslında başka bir açıdan sınıflama veya kümeleme probleminin de zorlaşarak doğrusal sınıflayıcılarla çözülebilme ihtimalini azaltmaktadır. Ayrışma genellikle kümelerin arası merkez mesafesiyle değerlendirilmektedir. Buna göre merkezler arası uzaklık arttıkça kümeler arasındaki ayrışma da artacaktır. Şekil 3.7'de yüksek ve düşük küme ayrışması görselleştirilmiştir.



Şekil 3.7: İki kümeli iki farklı gözlem uzayı ile yüksek ve düşük ayrışma görselleştirilmiştir. Küme merkezleri arası mesafe okla belirtilmiştir.

3.3 İç metrikler

3.3.1 Silhouette

Gözlemlerin ait oldukları sınıfların bilinmediği durumlarda veri kümesinin optimum sayıda kümeye bölünmesi önemli bir problemdir. Bu noktada küme geçerliliğinin ölçülmesi veya oluşturulan kümelerin nesnel ölçütlerle ne kadar başarılı bir şekilde veriyi sınıflara böldüğünün anlaşılabilmesi için nesnel metrikler geliştiril-

miştir. Bu metrikler eğer kümelerin saçılım, yoğunluk ve ayrılabilirlik gibi öz niteliklerine dayanıyorsa ve değerlendirme herhangi dış referans veya etiket bilgisi olmadan tamamlanıyorsa iç metrikler olarak adlandırılmaktadır. Silhouette ölçütü literatürde yaygın kullanım bulmuş pek çok iç metrikten birisidir. Bu metrik, her bir gözlemin ait olduğu kümenin diğer gözlemlerine olan ortalama uzaklığı ile en yakın diğer kümeye ait gözlemlere olan ortalama uzaklığın karşılaştırılmasına dayanmaktadır (Shahapure ve Nicholas, 2020). Böylece gözlemlerin ait oldukları kümeye veya diğer kümelere olan uygunlukları değerlendirilmektedir. Silhouette değeri $S(i)$ her bir gözlem i için şu şekilde tanımlanır:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i),\, b(i))} \tag{3.14}$$

Burada $a(i)$ gözlemin kendi kümesi içindeki diğer gözlemlere olan ortalama mesafesi, $b(i)$ ise en yakın kümeye ait gözlemlerle olan ortalama mesafesidir. Bu formülasyona göre Silhouette değeri -1 ile $+1$ arasında değişmektedir. $+1$'e yakın değerler gözlemin kendi kümesine yüksek derece ile uyum içinde olduğunu göstermekte iken 0 civarında olması kümeler arası sınırda kaldığını veya kümelerin üst üste bindiğini, -1'e yakın olması ise gözlemin yanlış kümeye atanmış olabileceğini ifade etmektedir. Kümeye dair genel bir değerlendirme yapılmak istendiğinde genellikle tüm gözlemler için ortalaması alınmaktadır. Ortalama Silhouette değerinin yüksek olması, kümelerin hem iyi ayrıştığını hem de kompakt olduğunu ifade etmektedir.

3.3.2 Davies–Bouldin

Davies–Bouldin (DB) indeksi herhangi dış referans veya etiket bilgisine ihtiyaç duymadan, verinin kendi yapısına dayanarak değerlendirme yapmaya yarayan iç metriklerden birisidir. Silhouette ölçütüne benzer şekilde kümelerin ne kadar kompakt ve ayrık olduğuna bağlı olarak değişmektedir. Ancak Silhouette ölçütünün tam tersi şekilde DB indeksinin düşük bir değer alması daha homojen bir kümelenmenin gerçekleştiğini ve kümelerin daha ayrık olduğunu işaret etmektedir. DB indeksi küme merkezleri arasındaki

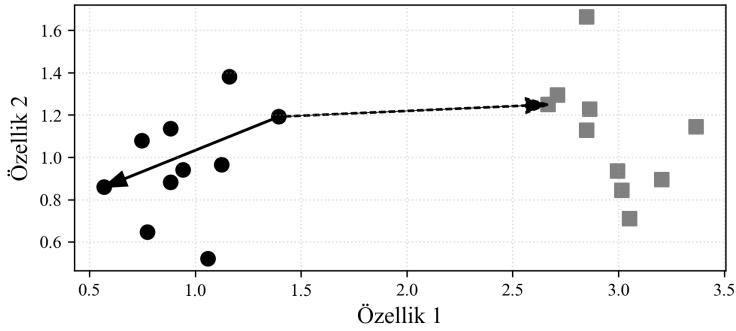
uzaklığın ve kümelerin ortalama saçılımlarının (kompaktlık) oranı olarak tanımlanmaktadır (Xiao vd., 2017). K adet küme için DB indeksi şu şekilde formülize edilir:

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \tag{3.15}$$

Burada: S_i, i'inci kümedeki gözlemlerin küme merkezine olan ortalama uzaklığını göstermektedir. M_{ij}, i ve j kümelerinin merkezleri arasındaki uzaklıktır. Böylece DB indeksi, hem kümelerin ne kadar ayrık olduğunu hem de küme kompaktlığını tek bir sayısal değerle ifade eder. Değerinin düşük olması iyi ayrılmış ve sıkı kümelenmiş bir yapıyı işaret etmektedir.

3.3.3 Dunn

Silhouette ölçütüne benzer şekilde kümeler arası ayrışmaya ve küme içi sıkılığa bağlı olarak değişmektedir. Benzer şekilde Yüksek Dunn indeksi değeri, kümeler arası uzaklığın yüksek, kümeler içi saçılımın ise düşük olduğunu, dolayısıyla kümelenmenin başarılı olduğunu ifade etmektedir. Matematiksel olarak en küçük kümeler arası mesafe, en büyük küme içi mesafe oranı ile hesaplanmaktadır. Not edilmelidir ki bu iki paremetre de aykırı gözlemlere karşı oldukça hassastır (Bezdek ve Pal, 1995). Dunn ölçütünün dayandığı bu iki parametre Şekil 3.8'de iki kümeli bir gözlem uzayı üzerinde gösterilmiştir.



Şekil 3.8: İki kümeli bir gözlem uzayı üzerinde küme içi maksimum uzaklık (düz ok) ve kümeler arası minimum uzaklık oku (kesik çizgi)gösterilmiştir.

3.3.4 Calinski–Harabasz

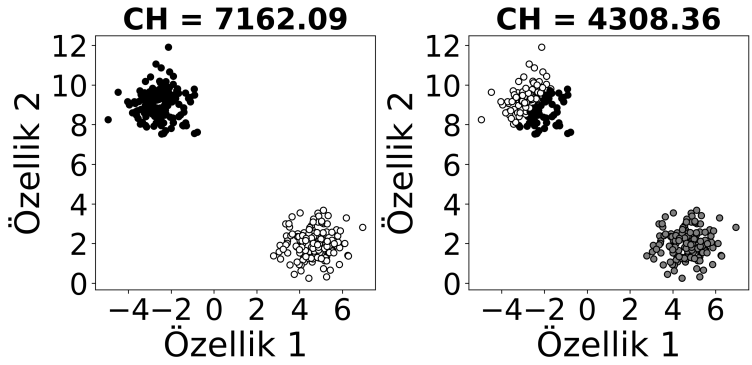
Calinski–Harabasz (CH) indeksi, kümeler arası ayrışma ve küme içi yoğunluk (kompaktlık) oranına dayanan bir ölçüttür (Cengizler vd., 2023). Bu sayede gözlemlerin yapısal olarak ne kadar başarılı kümelendiklerini derecelendirebilmektedir. Buna göre kümeler arası uzaklık artarken küme içi saçılım azaldıkça kümeleme başarısının artacağı varsayılmaktadır. Matematiksel formülasyonu şu şekildedir:

$$CH = \frac{B}{W} \times \frac{n-k}{k-1} \tag{3.16}$$

Bu formülde B, küme merkezlerinin genel ortalamadan uzaklıklarının karelerinin, ilgili kümedeki gözlem sayısı ile ağırlıklandırılarak toplanmasıyla elde edilen kümeler arası saçılımı; W ise her bir kümedeki gözlemlerin kendi merkezlerine olan uzaklıklarının karelerinin toplamıyla elde edilen küme içi saçılımı, n toplam gözlem sayısını, k ise toplam küme sayısını ifade etmektedir.

CH indeksi küme sayısının belirlenmesi amacıyla kullanılan en yaygın metriklerden birisidir. Ayrıca denetimsiz kümeleme sonuçlarının değerlendirilmesi için de kullanımı mümkündür. Genel olarak CH indeksinin yüksek değer alması, kümeler arası ayrışmanın güçlü, küme içi yayılımın ise zayıf olduğu, dolayısıyla iyi bir kümelenmenin gerçekleştiği anlamına gelmektedir. Küme kalitesi-

nin ölçümünde tercih edilirken bu yaklaşımın küme yoğunluğuna ve dağılımın şekline hassas olduğu mutlaka göz önünde bulundurulmalıdır. Şekil 3.9'da küme sayısının ve formlarının CH puanı üzerindeki etkisi görselleştirilmiştir.



Şekil 3.9: İki kümeli bir gözlem uzayı üzerinde optimum küme sayısı seçiminin CH puanı üzerindeki etkisi görülmektedir.

3.3.5 Yoğunluk Tabanlı Küme Doğrulama

Yoğunluk tabanlı küme doğrulama (Density-Based Cluster Validation, DBCV) hesabında kümeler arası yoğunluk farkları dikkate alınarak kümeleme başarı ölçülmektedir. Bu özelliği yoğunluk tabanlı küme doğrulama ile merkez odaklı ve küre şeklinde küme dağılımı varsayımlı diğer yaygın iç metriklerden ayrılmaktadır. Bu sayede farklı yoğunluklara sahip ve düzensiz şekilli kümeler nesnel bir biçimde değerlendirebilmektedir (Moulavi vd., 2014). Şu şekilde formülize edilmektedir:

$$DBCV = \frac{1}{K} \sum_{i=1}^{K} V_i \tag{3.17}$$

Burada V_i her bir küme için tanımlanmış bir yerel puandır, kümenin yoğunluk ve ayrışma dengesini ifade etmektedir ve şu şekilde hesaplanır:

$$V_i = \frac{sep_i - dens_i}{\max(sep_i, dens_i)} \tag{3.18}$$

Burada K, toplam küme sayısını göstermek için kullanılmıştır. $dens_i$, i'inci kümenin ortalama çekirdek mesafesine dayalı yoğunluk ölçüsüdür ve bir noktanın kendi kümesindeki diğer gözlemlerle ne kadar ilişkili olduğunu ifade etmektedir. sep_i, i'inci kümenin en yakın diğer kümeye olan yoğunluk tabanlı etkin uzaklığıdır (mutual reachability) ve bir noktanın en yakın dış kümeye olan yoğunluk tabanlı mesafesini ifade etmektedir.

DBCV değeri -1 ile +1 arasında değişmektedir. Pozitif değerlere sahip kümeler iyi tanımlanmış, negatif değerler ise düşük kaliteli kümeleri ifade eder. Yani skor +1'e yaklaştıkça küme daha kompakt ve iyi ayrışmış kabul edilirken, -1'e yaklaştıkça kümeler arası karışma veya düşük yoğunluk gözlenmesi beklenmektedir.

3.4 Dış metrikler

3.4.1 Rand Ölçütü

İç metrikler kümelerin oluşturulduğu gözlem uzayına dair herhangi sınıf bilgisinin bilinmediği durumlarda kümelerin geçerliliğinin değerlendirilmesi için işlevseldirler. Alternatif gerçek hayat senaryolarında ise gözlemlerin ait oldukları gerçek sınıflar biliniyor olabilir. Bu gibi bir durumda denetimsiz kümeleme yaklaşımlarının yarattığı kümeler ile gerçek sınıflara göre gözlem dağılımının karşılaştırılması hem verinin değerlendirilmesi hem de kümeleme işleminin performansının ölçülmesi adına kritik olabilir. Dış metrikler gözlemlere dair sınıf bilgisinin bilindiği durumlarda bu karşılaştırmanın yapılabilmesi için nesnel değerlendirme ölçütleri olarak kullanılabilmektedir. Rand ölçütü bu dış metriklerden birisidir. Temelde iki kümenin birbirine ne kadar benzediğini ölçen bu yaklaşım gerçek sınıfların oluşturduğu kümeler ile denetimsiz kümeleme algoritmalarının oluşturduğu kümeleri karşılaştırmak için de kullanılabilir. Matematiksel formülasyonu şu şekildedir:

$$R = \frac{a+b}{a+b+c+d} \tag{3.19}$$

Bu formülasyonda a her iki yöntemde de aynı kümede bulunan, b her iki yöntemde de farklı kümelerde bulunan, c gerçek sınıflarda

aynı kümede olup kümeleme sonucunda farklı kümelerde bulunan, d gerçek sınıflarda farklı kümelerde olup kümeleme sonucunda aynı kümede bulunan gözlem çiftlerinin sayısını ifade etmektedir (Campello, 2007). Rand indeksi puanı 0 ile 1 arasında değişmektedir. Sonucun 1'e yaklaşması gerçek kümeleme ile kestirim kümeleme arasındaki benzerliğin artması anlamına gelmektedir.

Formülasyon doğruluk (accuracy) hesabına benzemekle birlikte Rand indeksi ile doğruluk hesabının aynı şey olmadığına dikkat etmek gerekir. Doğruluk değeri tekil gözlemler ile hesaplanırken, Rand indeksinin gözlem çiftleri arasındaki ilişki tutarlılığı üzerinden hesaplandığı unutulmamalıdır.

3.4.2 Düzeltilmiş Rand Ölçütü

Rand indeksinin gerçek dünya uygulamalarında görülen zayıflıklarından birisi de tamamen rastgele gerçekleştirilen bir kümelemede dahi pozitif değerler alabilmesidir. Özellikle küme sayısı veya dağılımı dengesiz olduğunda gerçekte olması gerekenden daha yüksek puanlarla değerlendirme yapabilmektedir. Bu problemi çözebilmek adına rastlantısal olarak düzeltilmiş bir versiyonu önerilmiştir. Düzeltilmiş Rand Ölçütü (Adjusted Rand Index, ARI) olarak adlandırılan bu versiyon rastgelelikten kaynaklanan benzerlik payını toplam puandan çıkararak daha güvenilir bir değerlendirme sunmaktadır (Campello, 2007). Matematiksel olarak şu şekilde ifade edilir:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \tag{3.20}$$

Bu formülasyonda RI Rand indeksi değerini, $E[RI]$ ise rastgele bir kümeleme durumunda alınacak muhtemel Rand indeksi değerini ifade etmektedir. Bu yaklaşımda ARI değeri -1 ile 1 arasında değişmektedir. Puanlamanın 1'e yaklaşması küme benzerliğinin yükseldiğini, 0 civarında olması olası rastgeleliği, negatif değerlere inmesi ise beklenenden benzerliğin azalmakta olduğu anlamına gelmektedir.

3.4.3 F-Ölçütü

F-ölçütü genellikle otomatik sınıflayıcı başarısının değerlendirilmesinde kullanılan performans ölçütlerinden birisidir. Özellikle sınıf dağılımının görece olarak daha dengesiz olduğu durumlarda nesnel bir ölçüt olarak tercih edilmektedir. Buna göre kesinlik (precision) ve duyarlılık (recall) değerlerini tek ve dengeli bir metrik ile ifade etmeyi amaçlamaktadır. Unutulmamalıdır ki kesinlik ve duyarlılık ölçütleri arasındaki denge modelin yalnızca birini maksimize etmeye çalıştığında diğerinin genellikle azalması ile ilgilidir. F-ölçütü, bu dengenin bir harmonik ortalaması olarak şu şekilde tanımlanmıştır:

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.21}$$

Bu formülde geçen precision ve duyarlılığı şu şekilde tanımlamak mümkündür:

$$Precision = \frac{TP}{TP + FP} \tag{3.22}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.23}$$

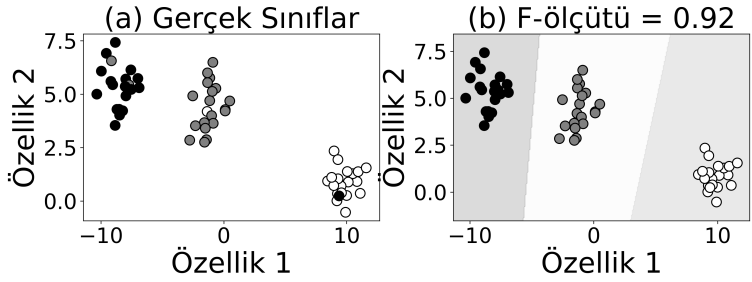
Burada TP (True Positive) doğru pozitifleri, FP (False Positive) yanlış pozitifleri ve FN (False Negative) yanlış negatifleri ifade etmektedir (Rendón vd., 2011). Genellikle sadece iki kümeli bir sınıflandırma probleminin değerlendirilmesine yönelik geliştirilmiş gibi dursa da F-ölçütü çoklu sınıf veya çoklu küme durumlarına da uyarlanabilmektedir. Bu durumda her bir sınıf için ayrı ayrı F-değerleri hesaplanır (birine karşı hepsi şeklinde) ve bunların ortalaması son puan olarak kabul edilir. Tüm sınıflar eşit ağırlıklı ise aritmetik ortalama, gözlem sayısına göre baskınlık tanımlanması isteniyorsa ağırlıklı ortalama tercih edilebilmektedir. Bu sayede dengesiz sınıf dağılımlarının gözlendiği veri kümelerinde çoklu sınıflandırma problemlerinde de model performansını F-Ölçütü ile değerlendirilebilmektedir.

Unutulmamalıdır ki verilen formülasyonda doğruluk ve duyarlılığın eşit ağırlıkta oldukları varsayılmıştır. Ancak bazı senaryo-

larda birinin baskın olduğu uygulamalar için F_β versiyonu kullanılabilir. Burada β parametresi duyarlılığa verilen ağırlığı kontrol eder:

$$F_\beta = (1 + \beta^2) \times \frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall} \tag{3.24}$$

Burada $\beta > 1$ seçilirse duyarlılığa, $\beta < 1$ seçilirse kesinliğe daha fazla önem verilmiş olur. Şekil 3.10'da örnek bir gözlem uzayı üzerinde otomatik kümeleme uygulanarak F-Ölçütü cinsinden değerlendirilmiştir.



Şekil 3.10: Örnek bir gözlem uzayı üzerinde otomatik kümeleme uygulanarak F-Ölçütü cinsinden performans değerlendirmesi yapılmıştır. (a) panelinde gözlemlerin ait oldukları gerçek sınıflar; (b) panelinde ise kümeleme sonucunda atandıkları sınıflar ve bu kümeleme sonucunun F-ölçütü değeri gösterilmiştir.

3.5 Metrik seçimi

Özellik uzayı içinde gözlemlerin kümelere atanarak bu kümelerin değerlendirilmesi pek çok farklı açıdan bilgi sağlamaktadır. Özelliklerin öneminin araştırılması, kümeleme algoritmalarının performansının ölçülmesi, gözlemlerin birbirleri ile olan ilişkilerinin araştırılması veya en uygun sayıda kümenin tespit edilmesi gibi pek çok farklı görev için küme doğrulama yaklaşımları tercih edilmektedir. Ancak bu noktada veriye ve uygulanan yönteme en uygun metriğin tercih edilmesi analizin etkinliği açısından büyük önem taşımaktadır.

Bu noktada önemli hususlardan birisi gerçek etiketlerin bilinip bilinmemesidir. Eğer gözlemlerin gerçek sınıfları biliniyorsa uygun bir dış metrik seçilebilir.

Ayrıca, seçilen ölçütün hangi düzeyde analiz yaptığı da seçimi doğrudan etkileyebilir. Nokta düzeyinde veya küme düzeyinde yapılan değerlendirmeler elde edilen sonuçların yorumlanma biçimini değiştirmektedir. Mesela Silhouette metriği her bir gözlemin kendi kümesine olan uyumunu ve en yakın diğer kümeye uzaklığını dikkate alarak nokta düzeyinde analiz yaparken, DB ve Dunn metrikleri kümeler arası ayrışma ile küme içi kompaktlık oranlarını karşılaştıran küme düzeyinde ölçütlerdir. Kümeler görece dengeli dağılım göstermekte ve küre biçiminde kompakt yapılar oluşturmaktaysa Silhouette, DB veya Dunn gibi indekslerin seçilmesi uygun olabilir. Ancak Dunn indeksinin gürültü hassasiyeti mutlaka göz önünde bulundurulmalıdır.

Bazı senaryolarda ise değerlendirmenin amacı önem kazanmaktadır. Eğer hedef optimum küme sayısını tespit etmek veya genel ayrışma düzeyini istatistiksel olarak derecelendirmek ise, varyans temelli ölçütler daha uygun bir seçim olabilir. Bu noktada CH, veri seti genelinde kümeler arasındaki farkların ne ölçüde anlamlı olduğunu varyans kullanarak ölçen global bir ölçüt olarak öne çıkmaktadır. Hesaplama açısından yalnızca küme merkezleri ve varyans değerlerini gerektirdiği için diğer metriklere göre görece hızı daha yüksektir ve büyük veri kümelerinde etkinliği yüksektir. Ancak bu indeks, kümelerin küresel ve benzer yoğunlukta olduğu varsayımına dayanır; farklı şekil veya yoğunluktaki kümelerde yanıltıcı sonuçlar verebilir.

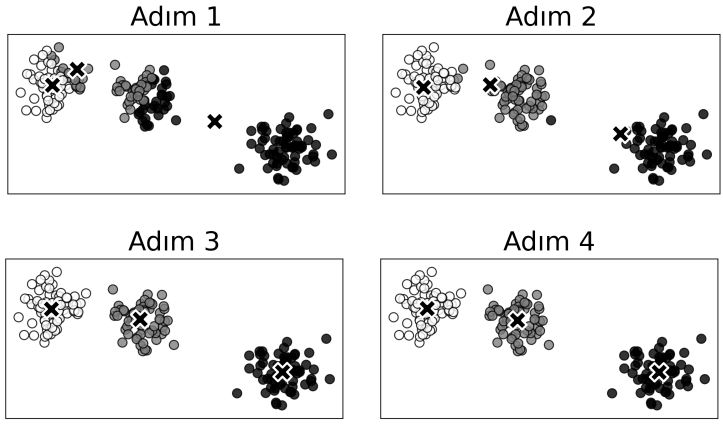
Metrik seçerken dikkat edilecek bir başka önemli parametre de şekil ve yoğunluk dağılımıdır. Eğer yoğunluğa dayalı bir kümeleme yapılacaksa veya verinin yoğunluk dağılımı küresel değilse yoğunluk tabanlı küme doğrulama uygun bir seçim olabilir.

Bölüm 4

Kümeleme Algoritmaları

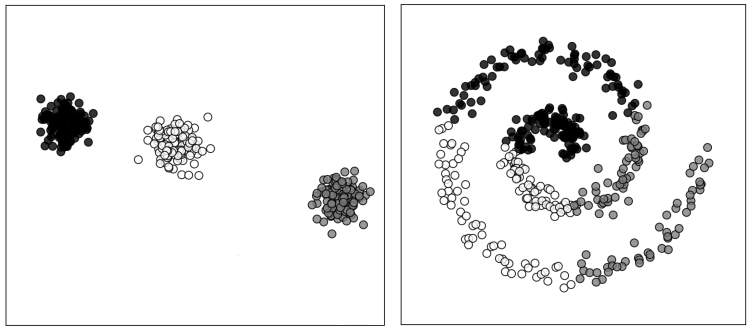
4.1 K-means Kümeleme

Otomatik kümeleme yaklaşımları birbiri ile ilişkisi en yüksek gözlemleri bir araya getirerek veriyi kümelere bölmeyi hedefleyen algoritmalardır. Bu algoritmalar özellik uzayında gözlemleri çeşitli kriterlere göre ilişkilendirmeye çalışırlar. K-means en yaygın kullanım bulmuş denetimsiz otomatik kümeleme yaklaşımlarından birisidir. Bu yaklaşımda bütün gözlemler en yakın küme merkezine atanarak küme merkezleri iteratif şekilde güncellenmektedir. Bu sayede küme içi toplam varyansın minimize edilmesi hedeflenmektedir (Rendón vd., 2011). Buna göre algoritma aşağıda verilen amaç fonksiyonunu (objective function) minimize etmeye çalışmaktadır. Şekil 4.1'de K-means algoritmasının rastgele merkezler ile başlayıp küme aidiyetlerini ve merkezleri adım adım güncelleyerek optimum çözüme doğru gidişi görselleştirilmiştir.



Şekil 4.1: K-means algoritmasının işleyişi 4 adımda gösterilmiştir. Çarpı işaretleri küme merkezlerini göstermektedir. Gözlemler atandıkları kümelere göre farklı tonlarda görselleştirilmişlerdir.

Otomatik kılavuzsuz kümeleme için K-means seçiminde önemli faktörlerden birisi sınıfların şekli ve gözlemlerin dağılımıdır. kümeler Öklidyen mesafesine göre ayırıldığı için kümeler dairesel veya küresel şekle benzedikçe algoritmanın etkinliği artacaktır. Yani Küme yayılım ve yoğunlukları birbirine yaklaştıkça, merkez tabanlı algoritmanın verimliliği artmaktadır. Bir başka önemli parametre de özellik ölçeğidir. Gözlemlerin görece uzaklıkları Öklid mesafesi ile ölçüldüğünden verinin normalizasyonu veya standardizasyonu K-means kümeleme başarısı için önemli bir önişlemdir. Bunlara ek olarak özelliklerin ayırt ediciliği ve ayrık gözlemlerin varlığı ve miktarı da genel performansı etkileyen etkenler arasındadır. Şekil 4.2'de K-means algoritması için farklı derecelerde uygun olan iki farklı veri kümesine örnek gösterilmiştir.



Şekil 4.2: K-means için uygun olabilecek veri kümesi solda gösterilmiştir. Sağda ise algoritmanın zorlanabileceği bir veri dağılımı görülmektedir.

4.1.1 K-means Varyantları

K-means yaklaşımı farklı veri işleme problemlerini çözebilmek adına zaman içinde adapte edilerek çeşitli farklı varyantlar üretilmiştir. Bunlardan en bilinenlerden birisi C-means algoritmasıdır. Bu yaklaşımda gözlemler bir seferde tek bir kümeye atanmak yerine farklı kümelere olan aidiyet dereceleri ayrı ayrı hesaplanmaktadır. Bu nedenle Fuzzy (bulanık) C-means olarak da adlandırılan bu varyant geçiş bölgelerini daha iyi analiz edebilmektedir.

Bir başka varyant olan K-medoids, K-means'e benzer biçimde çalışmaktadır ancak küme merkezi olarak kümeye ait gözlemlerin ortalamaları yerine kümenin en merkezi konumundaki gözlem (medoid) kullanılmaktadır. Bu sayede ayrık gözlemlere olan direnç artırılmıştır.

Büyük veri kümelerinde işlem yükünü azaltmak amacıyla rastgele alt örnek kümelerinin kullanıldığı MiniBatch K-means ve doğrusal olarak ayrılamayan kompleks özellik uzaylarını yüksek boyutlu bir çekirdek (kernel) uzayına dönüştürerek kümeleri bu uzayda oluşturan Kernel K-means de yine literatürdeki K-means varyantlarına örnek olarak gösterilebilir.

4.2 Hiyerarşik Kümeleme

Hiyerarşik kümeleme gözlemler arasındaki ilişki örüntülerinin kılavuzsuz şekilde ortaya çıkarıldığı bir başka makine öğrenimi yaklaşımıdır. Bu algoritmada gözlemler arasında hiyerarşik bir şema (veri ağacı benzeri bir yapı) ortaya çıkartılarak veri gruplarının birbiri ile ilişkilendirilmesi sağlanmaktadır (Murtagh ve Contreras, 2012). Buna göre gözlemler arasındaki benzerlikler dikkate alınarak veri kademeli biçimde kümelere ayrılmakta veya birleştirilmektedir. Önemli avantajlarından birisi K-means yaklaşımında olduğu gibi küme sayısının önceden belirtilmesini gerektirmemesidir. Küme aidiyeti kararı elde edilen hiyerarşik yapı üzerinde belirli bir kesme noktası seçilerek gerçekleştirilmektedir. Algoritmanın oluşturduğu hiyerarşik yapı genellikle dendrogram adı verilen diyagram ile gösterilmektedir. Şekil 4.3'de örnek bir dendrogram yapısı gösterilmiştir. Dikey eksen kümeler arasındaki uzaklığı, yatay eksen ise gözlemleri temsil etmektedir. Ağaç belirli bir yükseklikten kesilerek farklı sayıda küme elde edilebilir.

Şekil 4.3: Örnek bir dendrogram grafiği görülmektedir. Dikey eksen kümeler arası uzaklığı, yatay eksen gözlemleri göstermektedir.

Hiyerarşik kümeleme iki farklı perspektifle veriye yaklaşabilmektedir. Bunlardan birisi birleştirici (agglomerative) yöntemdir. Bu yöntemde başlangıçta gözlemler birer küme olarak kabul edilir ve her adımda en benzer iki küme birleştirilir. Iterasyonlara tüm gözlemler tek bir kümede toplanıncaya kadar devam edilmektedir. Diğer yaklaşım olan bölücü (divisive) yöntemde ise tüm gözlemleri içeren tek bir kümeden başlanarak kümeler kademeli biçimde alt

kümelere farklılıklarına göre ayrılır. Her iki yaklaşım da aynı hiyerarşik yapıyı farklı perspektiflerden inşa etmektedir.

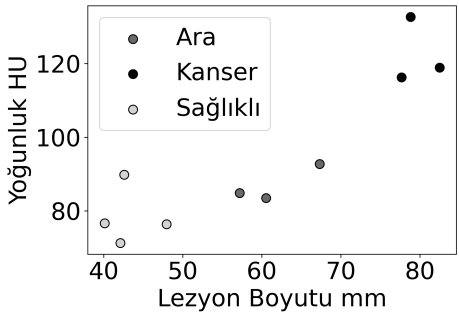
Benzerlik ölçütü olarak genellikle Öklidyen veya Manhattan uzaklıkları kullanılmaktadır. Ancak hiyerarşik kümelemede analizi etkileyen önemli ölçütlerden birisi de iki küme arasındaki uzaklığın nasıl tanımlandığını belirleyen bağlantı (linkage) ölçütüdür. Buna göre iki küme arasındaki en yakın iki gözlem dikkate alınabilir (tek bağlantı, single linkage), iki küme arasındaki en uzak iki gözlem dikkate alınabilir (tam bağlantı, complete linkage) veya kümeler arası tüm gözlem çiftlerinin ortalama uzaklığı hesaplanır (ortalama bağlantı, average linkage). Tek bağlantı yaklaşımı uzun, zincir şeklinde kümeler oluşturma eğilimindeyken, tam bağlantı yaklaşımı ile daha kompakt kümeler elde edilebilmektedir. Ortalama bağlantı yaklaşımında ise daha dengeli bir yapı oluşturulmaktadır. Bunlara alternatif olan Ward yönteminde de küme birleşiminin en az toplam hata karesine neden olacağı bağlantı seçilerek varyans tabanlı bir strateji izlenmektedir.

Hiyerarşik kümeleme veriye dair çok az ön varsayımla uygulanabilmektedir. Buna göre küme sayısının bilinmediği veya verinin farklı seviyelerdeki ilişkilerinin araştırıldığı durumlarda etkin şekilde kullanılabilmektedir. Dendrogram grafikleri verinin iç yapı örüntülerine dair bilgi sağlamakta bu sayede gözlemler arası benzerlik ilişkileri görsel olarak analiz edilebilmektedir. Bu noktada çok büyük veri kümelerinin analizi olası matris büyüklükleri nedeni ile yöntemi pratik olmaktan uzaklaştırabileceğinden yaklaşımın gerektirebileceği hesaplama maliyeti mutlaka gözönünde bulundurulmalıdır. Hiyerarşik küme yaklaşımını örneklendirmek üzere sentetik bir kanser teşhisi verisi hazırlanmıştır. 10 hasta ve 5 farklı radyolojik ölçümden oluşan veri Tablo 4.1'de verilmiştir.

Tablo 4.1: Kanser teşhisine dair oluşturulan sentetik veri ve atanan kanser sınıfları verilmiştir.

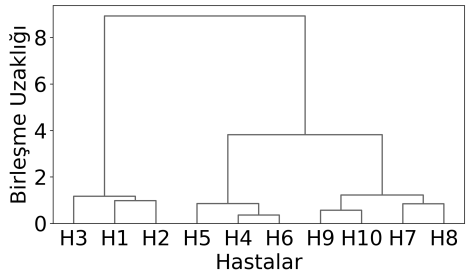
Hasta	Boyut (mm)	Yoğunluk	Kontrast	Pürüzlülük	Vaskülarite	Gerçek Küme
H1	82.48	118.89	0.83	0.78	0.89	Kanser
H2	78.83	132.63	0.84	0.68	0.92	Kanser
H3	77.68	116.27	0.81	0.60	0.83	Kanser
H4	57.19	84.87	0.57	0.45	0.53	Ara
H5	67.33	92.74	0.55	0.43	0.57	Ara
H6	60.55	83.49	0.57	0.47	0.59	Ara
H7	42.59	89.82	0.35	0.31	0.44	Sağlıklı
H8	40.12	76.67	0.25	0.30	0.41	Sağlıklı
H9	47.95	76.37	0.34	0.34	0.33	Sağlıklı
H10	42.12	71.31	0.40	0.36	0.31	Sağlıklı

Tablo 4.1'de sunulan veride yer alan yoğunluk ve lezyon boyutu özellikleri ile örnek gözlem uzayının bir saçılım grafiği hazırlanarak Şekil 4.4'de gösterilmiştir.



Şekil 4.4: Kanser teşhisi sentetik verisinden iki özellikle oluşturulmuş saçılım grafiği görülmektedir.

Şekil 4.4'de verinin oluşturduğu doğal kümelerin özellikle kanserli ve sağlıklı hasta gruplarında birbirinden ayrıştığı görülmektedir. Hiyerarşik kümeleme işlemi öncesinde veri öncelikle standardize edilerek özelliklerin benzer ölçekte işleme alınabilmesi sağlanmıştır. Ardından Ward yaklaşımı ile kümeler arası uzaklıklar bulunmuş, bu uzaklıklar kullanılarak dendrogram diyagramı oluşturulmuştur. Diyagram aşağıda gösterilmiştir (Şekil 4.5).



Şekil 4.5: Kanser teşhisi sentetik verisinden elde edilen dendrogram grafiği görülmektedir.

Bu örnek uzayda kanser grubuna ait ölçümlerin belirgin biçimde yüksek olması nedeniyle bu gözlemler hiyerarşik yapıda ayrı bir dal olarak ayrışmış ve dendrogramda açık biçimde gözlenebilir hale gelmiştir.

4.3 Yoğunluk Tabanlı Kümeleme

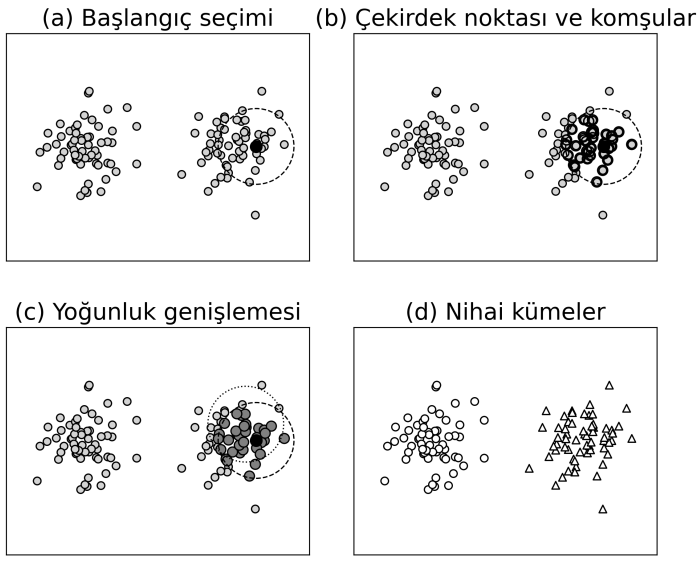
4.3.1 DBSCAN

Yoğunluk tabanlı kümeleme doğrudan merkez tabanlı yaklaşımlara bir alternatif olarak gözlemlerin yoğunlaşmalarını tespit ederek bu yoğunlaşmayı takip etmeyi amaçlamaktadır. Bu sayede ilişkili gözlemleri küme merkezinin etrafındaki dağılım geometrilerinden farklı olarak aynı kümeye atayabilmektedir.

Yaygın kullanım bulmuş yoğunluk tabanlı kümeleme yaklaşımlarından birisi de DBSCAN (Density-Based Spatial Clustering of Applications with Noise) yaklaşımıdır. Bu yöntem, önceden belirlenen bir yarıçap(ε) içindeki gözlemleri komşu kabul etmektedir. Buna göre kümeleme komşu yoğunluğuna dayanılarak gerçekleştirilmektedir (Khan vd., 2014). Böylece küre biçiminde olsa da olmasa da kümeleme etkin şekilde gerçekleştirilebilmektedir. Bu algoritma gözlemlerin dağılım karakterlerine göre üç tipten birisi olduğunu varsaymaktadır. Bu varsayıma göre çekirdek noktalar ε yarıçapı içinde en az bir parametre olarak belirlenen limit miktarda (MinPts) komşusu bulunan noktalardır. Sınır noktalar bir

çekirdek noktaya komşu olmasına rağmen limit miktarda komşusu bulunmayan noktalardır. Gürültü noktalar ise herhangi kümeye atanmamış izole haldeki gözlemlerdir.

Algoritma, rastgele bir noktadan başlayarak komşuları inceler ve çekirdek olup olmadığına karar verir. Eğer nokta çekirdek olarak kabul edildiyse ε yarıçapı içindeki tüm komşuları aynı kümeye dahil edilmektedir. Aynı işlem kümelenmemiş tüm noktalar için tekrarlanarak kümeleme tamamlanır. Böylece küme sayısının önceden kestirilmesine gerek kalmadan esnek bir kümeleme gerçekleştirilmektedir. Bu yapısı sayesinde izole, ayrık gözlemleri dışlayabilmektedir ancak ε ve MinPts parametrelerinin verinin yapısına uygun seçilmesi işlemin başarısı için kritik olabilir. Çok küçük bir ε değeri fazla sayıda küçük küme ve gürültü üretirken, çok büyük bir değer kümeleri birleştirerek ayrışmayı azaltabilir. Şekil 4.6'da DBSCAN algoritması ile gerçekleştirilen yoğunluk tabanlı bir kümeleme örneği gösterilmiştir.



Şekil 4.6: DBSCAN algoritması ile gerçekleştirilen yoğunluk tabanlı kümeleme örneği.

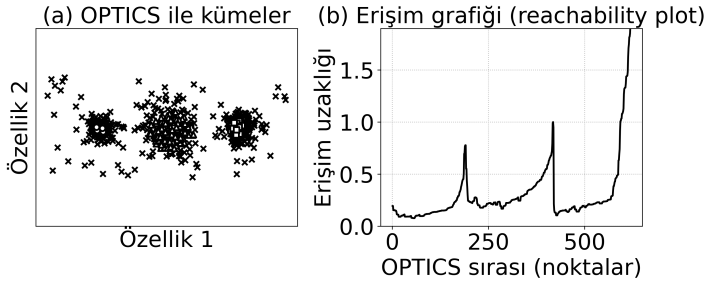
4.3.2 OPTICS

DBSCAN algoritması verinin yoğunluk dağılımını takip ederek farklı geometrilerde yayılmış aynı kümeye ait gözlemleri yakalayabilmektedir ancak bir limitasyonu sabit yarıçap ile iterasyonlara devam etmesidir. Veri kümesinde bulunması muhtemel farklı yoğunluklardaki kümeler etkinliği azaltabilmekte bazı kümeler ya birleşmekte ya da gürültü olarak tanımlanmaktadır. Bu noktada OPTICS (Ordering Points To Identify the Clustering Structure) algoritması, farklı yoğunluk seviyelerindeki kümelerin tanımlanabilmesi adına DBSCAN algoritmasını genelleştirmiştir (Ankerst vd., 1999). Bu alternatif yaklaşım yoğun kümelenmelerde küçük ε, daha seyrek kümelenme bölgelerinde ise görece daha büyük ε değerlerinin kullanılabilmesini sağlayarak yöntemin veri yoğunluğuna adapte olmasını sağlamaktadır.

OPTICS algoritması adaptasyonu sağlamak için her bir gözlem noktasında çekirdek uzaklığı ve erişim uzaklığı olmak üzere iki temel kavram hesaplamaktadır. Çekirdek uzaklığı, bir noktanın MinPts'inci en yakın komşusuna olan mesafeyi tanımlar. Yani bir noktanın çekirdek olabilmesi için gereken en küçük yarıçaptır. Erişim uzaklığı ise bir noktanın başka bir çekirdek noktaya yoğunluk tabanlı erişilebilirlik mesafesini ifade etmektedir. Bu uzaklık, ilgili çekirdek noktanın çekirdek uzaklığından küçük olmamalıdır.

Algoritma, erişim uzaklıklarına göre sıralanmış noktalardan bir erişim grafiği çıkartmaktadır. Bu grafik, farklı yoğunluklarda kümelerin hiyerarşisini görsel olarak araştırmayı mümkün kılmaktadır. Bu yöntemin kullanıcıdan tek bir ε değeri istememesi önemli bir avantajdır. Bunun yerine veri içindeki farklı yoğunluk seviyelerine adapte olarak farklı yoğunluklara sahip daha sofistike veri yapılarında etkinliğini sürdürebilmektedir.

Şekil 4.7'de örnek bir veri uzayı üzerinde OPTICS algoritmasının ürettiği erişim grafiği örneği gösterilmiştir. Grafikteki çukurlar kümeleri göstermektedir. Çukurların derinleşmesi küme yoğunluğunun yüksek olduğunu göstermektedir

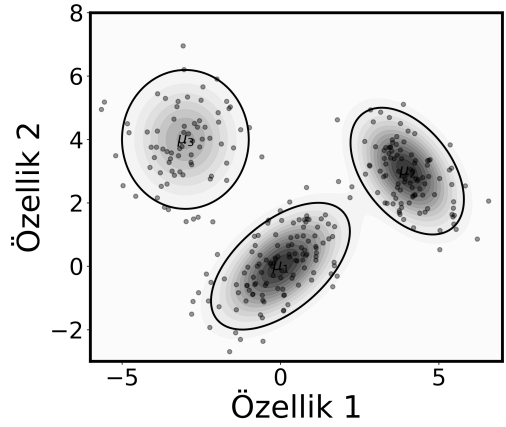


Şekil 4.7: OPTICS algoritmasının ürettiği örnek bir erişim grafiği gösterilmiştir.

4.4 Gaussian Karışım Modeli

Merkez ve yoğunluk tabanlı yaklaşımlara bir başka alternatif de gözlemlerin küme aidiyetlerini istatistiksel olarak kestirmeye çalışan yaklaşımlardır. Bu yaklaşımlar gözlemlerin konumsal yoğunluğundan ziyade olasılıksal yoğunluğunu haritalandırarak bir derecelendirme yapmaktadır. Buna göre veri kümesi olasılık modellerinin bir karışımı olarak görülmekte böylece kümeleme işlemi bu karışımın parametrelerinin kestirilmesi görevine indirgenmektedir (Wang ve Jiang, 2021). Bu yaklaşıma göre kümeleri aynı tipte dağılıma ancak farklı parametrelere sahip olasılık yoğunluk fonksiyonları ile tanımlamak mümkündür.

Gaussian Karışım Modeli (Gaussian Mixture Model, GMM) kümeleri, ortalama vektörü ve kovaryans matrisi ile tanımlanan çok değişkenli bir Gaussian dağılımı olarak modellemektedir. Veri kümesi bu dağılımların ağırlıklı toplamı olarak düşünülmelidir. Bu yaklaşımda gözlemler kümelere kesin bir biçimde atanmaz ancak bunun yerine belirli bir aidiyet olasılığı ile küme-gözlem ilişkileri kurulmaktadır. Şekil 4.8'de iki boyutlu bir veri kümesi üzerinde gözlemler arka plandaki yoğunluk alanı üzerinde konumlanmış şekilde gösterilmiştir.



Şekil 4.8: İki boyutlu bir veri kümesi üzerinde gözlemler ve yoğunluk alanları gösterilmiştir. Elipsler kümelerin kovaryans yapısını göstermektedir.

Gaussian Karışım modeli istatistiksel anlamda güçlü ve esnek bir yöntem olmasına rağmen, varsayılan dağılım biçiminin doğru seçilmemesi durumunda modelin etkinliği azalabilmektedir. Ayrıca karmaşık dağılımlar için parametre tahmini süreci yüksek işlem gücü gerektirebilmektedir. Buna rağmen özellikle biyomedikal, görüntü analizi ve gen ekspresyonu gibi alanlarda sıkça tercih edilen bir yaklaşımdır.

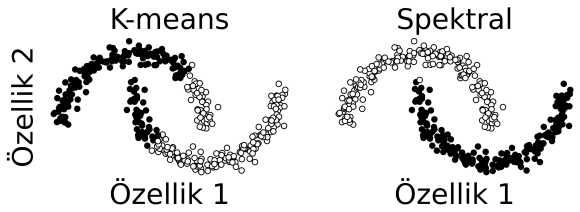
4.5 Spektral Kümeleme

Şekil olarak düzensiz, dolaşık ve sınırları belirsiz verilerde küme ilişkilerinin ortaya çıkarılması için geleneksel yaklaşımlar yetersiz kalabilmektedir. Spektral kümeleme konumsal mesafe ve merkez tabanlı yaklaşımlar yerine gözlemler arası ilişkileri farklı bir perspektiften ortaya çıkarmayı hedefleyen yaklaşımlardan birisidir.

Bu yaklaşım gözlemlerin diğerlerine olan yakınlığını ölçen bir ilişki tablosuna (benzerlik matrisi) dayanmaktadır. Bu tablo, verinin içerdiği farkedilemeyen örüntülerin ortaya çıkmasını sağlamaktadır. Bu ilişki yapısı özvektör temsiline dayalı bir dönüşüm aracılığı ile verinin farklı ama daha açık bir biçimde temsil edildiği yeni bir uzaya dönüştürülmekte bu sayede yeni uzayda belirgin

hale gelen kümesel ilişkiler klasik kümeleme algoritmalarıyla kolayca ayrılabilmektedir (Von Luxburg, 2007). Spektral kümeleme karmaşık dağılım gösteren hasta gruplarının ayrıştırılması için etkin bir yaklaşımdır.

Şekil 4.9'da yapısal olarak dolaşık bir veri kümesi görülmektedir. Klasik K-means bu iki grubu ayırmakta zorlanacaktır ancak spektral kümeleme gözlemler arasındaki ilişkilere odaklandığı için kümeleri doğru biçimde ayırabilmiştir.



Şekil 4.9: Dolaşık bir veri kümesinde klasik ve spektral kümeleme sonuçları görülmektedir.

4.6 Meta-sezgisel kümeleme

Gerçek hayat problemlerinde özellikle karmaşık gözlem uzayları yaratan tıbbi veri problemlerinde otomatik kümeleme yaklaşımlarının pek çoğu optimum çözüme ulaşmakta zorlanabilmektedir. Özellikle doğrusal olmayan, çok boyutlu ve karmaşık yapılara sahip veri kümeleri geleneksel yaklaşımların çözemeyeceği kadar sofistike yapılar yaratmaktadır. Bu noktada kümeleme problemi bir optimizasyon problemi olarak ele alındığında meta-sezgisel yaklaşımlar geleneksel deterministik yöntemlere bir alternatif teşkil etmektedir. Meta-sezgisel algoritmalar doğadan esinlenilmiş optimuma yakın çözüm arama yaklaşımlarıdır ve çok büyük alternatif çözüm uzaylarında bile yüksek kaliteli kümeleri sezgisel olarak ararlar.

Meta-sezgisel algoritmalar sınamakta oldukları çözümleri değerlendirebilmek için bir uygunluk (fitness) fonksiyonuna ihtiyaç duyarlar (Abualigah vd., 2021). Bu fonksiyonun problemin doğasına uygun ve veri yapısı ile uyumlu olması önemlidir. Oluşturu-

lan alternatif kümelerin birbirlerine olan uzaklıklarının artması ve kendi içlerindeki homojenliğin korunması uygunluk fonksiyonu sayesinde derecelendirilerek sağlanmaktadır. Böylece algoritma, en iyi kümelenmeyi sağlayacak alternatif cevapları iteratif biçimde arar. Genetik Algoritma, Parçacık Sürü Optimizasyonu (PSO), Karınca Kolonisi (ACO) ve Yapay Arı Kolonisi (ABC) yaygın kullanım bulmuş meta-sezgisel yaklaşımlardır.

Parametrelerin sayısının fazla olduğu, gürültülü veya karmaşık biyomedikal veri kümelerinde meta-sezgisel arama avantaj sağlama potansiyeline sahiptir. Örneğin gen ekspresyonu verilerinde hasta alt tiplerini belirlemek, sayısal görüntü işleme görevlerinde bölge ayrımı yapmak veya biyopotansiyel sinyallerde anomali tespit etmek gibi geniş çeşitlilikteki görevlerde klasik yöntemlerin erişemediği çözümleri yaratabilmektedirler. Ancak unutulmamalıdır ki seçilen uygunluk fonksiyonunun derecelendirme başarısı, seçilen algoritmanın parametrelerinin iyi ayarlanmış olması, yineleme sayısının yeterli olması gibi faktörler uygulanan yaklaşımın başarısı üzerinde oldukça etkili olabilirler.

Bölüm 5

Sonuç

Günden güne zenginleşen ve sayısallaşan tıbbi veri algı sınırlarımızı zorlayacak boyutlarda bilgi içermekte. Bu gömülü bilgiyi işleyebilmek demek pek çok kritik kararın alınmasını ve zor sürecin yürütülmesini hızılandırarak kolaylaştırmak demek. Hekim pratiğine herhangi açıdan temas eden kimsenin alet kutusuna veriyi kümeler üzerinden inceleyerek anlamlandırmaasını kolaylaştıracak pek çok araç eklemiş olduk. Bu araçlar sayesinde veri yapısının incelenebilmesini ve anatomisinin ortaya çıkarılabilmesini hedefledik. Unutulmamalıdır ki etkin bir analiz uygulamak için yöntemler kadar verinin yapısal özelliklerine de hakim olunması kritik önem taşımakta. Temel, ancak son derece güçlü bilgilerle tek bir tuğla olarak yerine yerleştirdiğimiz bu ilk adım, tıbbi veri madenciliğinden, bilimsel araştırmalara pek çok alanda yardımcı kaynak olmaya adaydır. Verinin nasıl görselleştirilebileceği bu görsellerin yorumlanması, yardımcı grafiklerin yaratılması ve veri yapısının numerik ölçütlerle değerlendirilebilmesi hedefenen kazanımlardandı. Bu kazanımlar sayesinde geniş bir yelpazede gelen tıbbi bilgi içinde hastaları, hastalıkları veya herhangi ilgi odağını karakterize eden özellikleri farklı perspektiflerden araştırmak mümkün olacaktır. Hekimlerin genellikle aşina oldukları araştırma teknikleri ve istatistik gibi güçlü araçlara destekleyici bir unsur olarak kitapta incelenen metrikler, küme dinamikleri, kümelerin sınıflandırılmasını sağlayan özellikleri ve otomatik kılavuzsuz kümeleme yaklaşımları da araştırmaların etkinliğini ve hızını artırmaya aday-

dır.

Verinin kümeler perspektifinden incelenmesi hızlı ve iç görü sağlayabilmekle birlikte özelliklerin doğasının anlaşılması veya özellik uzayının sadeleştirilmesi gibi pek çok destekleyici iş için küme analizleri tercih edilebilmektedir. Ancak işlenen konuların pek çoğunda veri dağılımının kritik olduğu ve bazı mecburi ön varsayımlarla analize başlanabildiği görülmektedir. Bu unsurlar elbette küme analizlerinin limitlerinde belirleyici olmaktadır. Pek çok gerçek dünya uygulamasında özellikle çok büyük boyutlu karma hasta demografilerinde veriyi sınırları çok belirli kesin geometrilere oturtmak ciddi bir zorluktur. Kılavuzsuz yaklaşımlar daha önce hiç bir eğitime tabi tutulmadan en uygun kümeleri yaratmaya çalıştıklarından çok büyük boyutlu, sofistike veri kümeleri söz konusu olduğunda beklenmedik şekilde optimum çözüme ulaşamamaları veya beklenen bilgiyi sağlayamamaları olağandır. Bu noktada araştırmacının alet kutusundaki bütün araçları çok iyi tanıması ve alternatif yaklaşımlar geliştirebilmesi gerekmektedir.

Kılavuzsuz kümeleme ve küme analizlerinin doğal sınırlarının zorlandığı noktalarda günümüzde melez yaklaşımlar öne çıkmaya başlamıştır. Bu yaklaşımlar derin öğrenme gibi büyük miktarda ağırlık matrisleri ve çok sayıda katmanla veriyi tamamen yeni bir uzaya yansıtarak yeni perspektifler aydınlatabilmektedir. Araştırmacının en temel veri anlamlandırma araçları kadar bu gibi daha karmaşık melez yapıları da tanıyarak zorlu problemlere adapte edebilme becerisini kazanması günden güne karmaşıklaşan problemleri çözüme götürebilmesini kolaylaştıracaktır.

Kaynakça

Abualigah, L., Gandomi, A. H., Elaziz, M. A., Hamad, H. A., Omari, M., Alshinwan, M., ve Khasawneh, A. M. (2021). Advances in meta-heuristic optimization algorithms in big data text clustering. *Electronics*, 10(2):101.

Agrawal, P., Abutarboush, H. F., Ganesh, T., ve Mohamed, A. W. (2021). Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019). *Ieee Access*, 9:26766–26791.

Al Shalabi, L., Shaaban, Z., ve Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, 2(9):735–739.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., ve Sander, J. (1999). Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60.

Asanza, V., Peláez, E., Loayza, F., Mesa, I., Díaz, J., ve Valarezo, E. (2018). Emg signal processing with clustering algorithms for motor gesture tasks. In *2018 IEEE Third Ecuador Technical Chapters Meeting (ETCM)*, pages 1–6. IEEE.

Berguin, S. H. ve Mavris, D. N. (2015). Dimensionality reduction using principal component analysis applied to the gradient. *Aiaa Journal*, 53(4):1078–1090.

Bezdek, J. C. ve Pal, N. R. (1995). Cluster validation with generalized dunn's indices. In *Proceedings 1995 second New Zealand international two-stream conference on artificial neural networks and expert systems*, pages 190–193. IEEE.

Brereton, R. G. (2015). The mahalanobis distance and its relationship to principal component scores. *Journal of Chemometrics*, 29(3):143–145.

Campello, R. J. (2007). A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters*, 28(7):833–841.

Cengizler, Ç., Kabakci, A. G., Bozkır, D. M., Sire Eren, D., ve Bozkır, M. G. (2023). A cluster validity index-based objective criteria for aesthetic evaluation of periorbital treatment. *Clinical, Cosmetic and Investigational Dermatology*, pages 2537–2546.

Chawla, N. V., Bowyer, K. W., Hall, L. O., ve Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Ciaramella, A., Cocozza, S., Iorio, F., Miele, G., Napolitano, F., Pinelli, M., Raiconi, G., ve Tagliaferri, R. (2008). Interactive data analysis and clustering of genomic data. *Neural Networks*, 21(2-3):368–378.

Clark, M. C., Hall, L. O., Goldgof, D. B., Clarke, L. P., Velthuizen, R. P., ve Silbiger, M. S. (2002). Mri segmentation using fuzzy clustering techniques. *IEEE Engineering in Medicine and Biology Magazine*, 13(5):730–742.

Darst, B. F., Malecki, K. C., ve Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC genetics*, 19(Suppl 1):65.

El-Darzi, E., Abbi, R., Vasilakis, C., Gorunescu, F., Gorunescu, M., ve Millard, P. (2009). Length of stay-based clustering methods for patient grouping. In *Intelligent patient management*, pages 39–56. Springer.

Färber, I., Günnemann, S., Kriegel, H.-P., Kröger, P., Müller, E., Schubert, E., Seidl, T., ve Zimek, A. (2010). On using class-labels in evaluation of clusterings. In *MultiClust: 1st internati-*

onal workshop on discovering, summarizing and using multiple clusterings held in conjunction with KDD, page 1.

Ferreira, A. J. ve Figueiredo, M. A. (2012). An unsupervised approach to feature discretization and selection. *Pattern recognition*, 45(9):3048–3060.

González-Estrada, E., Villaseñor, J. A., ve Acosta-Pech, R. (2022). Shapiro-wilk test for multivariate skew-normality. *Computational Statistics*, 37(4):1985–2001.

Grant, R. W., McCloskey, J., Hatfield, M., Uratsu, C., Ralston, J. D., Bayliss, E., ve Kennedy, C. J. (2020). Use of latent class analysis and k-means clustering to identify complex patient profiles. *JAMA network open*, 3(12):e2029068–e2029068.

Habibzadeh, F. (2024). Data distribution: normal or abnormal? *Journal of Korean medical science*, 39(3).

Higham, D. J., Kalna, G., ve Kibble, M. (2007). Spectral clustering and its use in bioinformatics. *Journal of computational and applied mathematics*, 204(1):25–37.

Huang, A. vd. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56.

Huang, L., Shea, A. L., Qian, H., Masurkar, A., Deng, H., ve Liu, D. (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics*, 99:103291.

Hussain, H. M., Benkrid, K., Seker, H., ve Erdogan, A. T. (2011). Fpga implementation of k-means algorithm for bioinformatics application: An accelerated approach to clustering microarray data. In *2011 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*, pages 248–255. IEEE.

Jan, Z. ve Verma, B. (2020). Multicluster class-balanced ensemble. *IEEE Transactions on Neural Networks and Learning Systems*, 32(3):1014–1025.

Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., ve Decker, S. (2021). Deep learning-based clustering approaches for bioinformatics. *Briefings in bioinformatics*, 22(1):393–415.

Khan, K., Rehman, S. U., Aziz, K., Fong, S., ve Sarasvady, S. (2014). Dbscan: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pages 232–238. IEEE.

Kwak, S. K. ve Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology*, 70(4):407.

Liu, P., Yuan, H., Ning, Y., Chakraborty, B., Liu, N., ve Peres, M. A. (2024). A modified and weighted gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses. *BMC Medical Research Methodology*, 24(1):305.

Malkauthekar, M. (2013). Analysis of euclidean distance and manhattan distance measure in face recognition. In *Third International Conference on Computational Intelligence and Information Technology (CIIT 2013)*, pages 503–507. IET.

Mayorga, A. ve Gleicher, M. (2013). Splatterplots: Overcoming overdraw in scatter plots. *IEEE transactions on visualization and computer graphics*, 19(9):1526–1538.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, 23(2):143–149.

Milligan, G. W. ve Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of classification*, 5(2):181–204.

Moulavi, D., Jaskowiak, P. A., Campello, R. J., Zimek, A., ve Sander, J. (2014). Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 839–847. SIAM.

Murtagh, F. ve Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 2(1):86–97.

Nersisyan, S., Novosad, V., Galatenko, A., Sokolov, A., Bokov, G., Konovalov, A., Alekseev, D., ve Tonevitsky, A. (2022). Exhaufs: exhaustive search-based feature selection for classification and survival regression. *PeerJ*, 10:e13200.

Orhan, U., Hekim, M., ve Ozer, M. (2011). Eeg signals classification using the k-means clustering and a multilayer perceptron neural network model. *Expert Systems with Applications*, 38(10):13475–13481.

Özbay, Y., Ceylan, R., ve Karlik, B. (2006). A fuzzy clustering neural network architecture for classification of ecg arrhythmias. *Computers in Biology and Medicine*, 36(4):376–388.

Rendón, E., Abundez, I., Arizmendi, A., ve Quiroz, E. M. (2011). Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5(1):27–34.

Renear, A. H., Sacchi, S., ve Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4.

Rodrigues, É. O. (2018). Combining minkowski and chebyshev: New distance proposal and survey of distance metrics using k-nearest neighbours classifier. *Pattern Recognition Letters*, 110:66–71.

Rodrigues, J., Belo, D., ve Gamboa, H. (2017). Noise detection on ecg based on agglomerative clustering of morphological features. *Computers in biology and medicine*, 87:322–334.

Scott, D. W. (2010). Histogram. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):44–48.

Shahapure, K. R. ve Nicholas, C. (2020). Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE.

Shan, J., Cheng, H., ve Wang, Y. (2012). A novel segmentation method for breast ultrasound images based on neutrosophic l-means clustering. *Medical physics*, 39(9):5669–5682.

Shannon, W., Culverhouse, R., ve Duncan, J. (2003). Analyzing microarray data using cluster analysis. *Pharmacogenomics*, 4(1):41–52.

Shapiro, S. S. ve Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Silitonga, P. (2017). Clustering of patient disease data by using k-means clustering. *International Journal of Computer Science and Information Security (IJCSIS)*, 15(7):219–221.

Strauss, J. S., Bartko, J. J., ve Carpenter Jr, W. T. (1973). The use of clustering techniques for the classification of psychiatric patients. *The British Journal of Psychiatry*, 122(570):531–540.

Suwanda, R., Syahputra, Z., ve Zamzami, E. M. (2020). Analysis of euclidean distance and manhattan distance in the k-means algorithm for variations number of centroid k. In *Journal of Physics: Conference Series*, volume 1566, page 012058. IOP Publishing.

Thomas, D. R., Pastrana, S., Hutchings, A., Clayton, R., ve Beresford, A. R. (2017). Ethical issues in research using datasets of illicit origin. In *Proceedings of the 2017 Internet Measurement Conference*, pages 445–462.

Tuan, T. M. vd. (2017). Dental segmentation from x-ray images using semi-supervised fuzzy clustering with spatial constraints. *Engineering Applications of Artificial Intelligence*, 59:186–195.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Wang, J. ve Jiang, J. (2021). Unsupervised deep clustering via adaptive gmm modeling and optimization. *Neurocomputing*, 433:199–211.

Wong, K.-P., Feng, D., Meikle, S. R., ve Fulham, M. J. (2002). Segmentation of dynamic pet images using cluster analysis. *IEEE Transactions on nuclear science*, 49(1):200–207.

Wu, J. ve Mahfouz, M. R. (2016). Robust x-ray image segmentation by spectral clustering and active shape model. *Journal of Medical Imaging*, 3(3):034005–034005.

Xiao, J., Lu, J., ve Li, X. (2017). Davies bouldin index based hierarchical initialization k-means. *Intelligent Data Analysis*, 21(6):1327–1338.

Yang, C., Fridgeirsson, E. A., Kors, J. A., Reps, J. M., ve Rijnbeek, P. R. (2024). Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *Journal of Big Data*, 11(1):7.

Zhang, L.-Q., Shiavi, R., Hunt, M. A., ve Chen, J.-J. (1991). Clustering analysis and pattern discrimination of emg linear envelopes. *IEEE transactions on biomedical engineering*, 38(8):777–784.

TIBBİ VERİ ANALİZİNDE KÜME YAKLAŞIMLARI

yaz
yayınları

YAZ Yayınları
M.İhtisas OSB Mah. 4A Cad. No:3/3
İscehisar / AFYONKARAHİSAR
Tel : (0 531) 880 92 99
yazyayinlari@gmail.com • www.yazyayinlari.com

ISBN: 978-625-8508-71-0