ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

YÜKSEK BOYUTLU BÜYÜK VERİLERDE AYKIRI DEĞERLERİ TANILAMA TEKNİKLERİ

Aslıhan Asena KÖPRÜ

İSTATİSTİK ANABİLİM DALI

ANKARA
2024

ÖZET

Yüksek Lisans Tezi

YÜKSEK BOYUTLU BÜYÜK VERİLERDE AYKIRI DEĞERLERİ TANILAMA TEKNİKLERİ

Aslıhan Asena KÖPRÜ

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Prof. Dr. Olçay ARSLAN

Gelişen teknoloji ve artan kaynaklar nedeniyle ortaya çıkan yüksek boyutlu veriler konusuna odaklanan bu tez, binlerce veya milyonlarca değişken içeren veri kümelerinin analiz ve yönetiminde ortaya çıkardığı sorunları ele almaktadır. Bu tez kapsamında, büyük veri ve yüksek boyutlu verinin tanımları, tarihsel uygulamaları ve disiplinler arası ilişkileri araştırılmaktadır. Yüksek boyutlu büyük verinin analizinde kullanılan boyut indireme tekniklerinin pratik uygulamaları amaçlanmaktadır. Tez, aykırı değer tespit yöntemleri üzerine yapılan çalışmaların yanı sıra, bu yöntemlerin test edilmesine yönelik simülasyonları ve çalışmanın temel sonuçlarından elde edilen sonuçları içermektedir.

Şubat 2024, 71 sayfa

Anahtar Kelimeler: Anomali Tespiti, Aykırı Değer, Büyük Veri, Yüksek Boyut Problemi, Yüksek Boyutlu Veri

ABSTRACT

Master Thesis

OUTLIER DETECTION TECHNIQUES IN
HIGH-DIMENSIONAL BIG DATA

Aslıhan Asena KÖPRÜ

Ankara University
Graduate School of Natural and Applied Sciences
Department of Statistics

Supervisor: Prof. Dr. Olçay ARSLAN

Focused on the issue of high-dimensional data arising from advancing technology and increasing resources, this thesis addresses the challenges encountered in the analysis and management of datasets containing thousands or millions of variables. Within the scope of this thesis, definitions of big data and high-dimensional data, their historical applications, and interdisciplinary relationships are investigated. Practical applications of dimensionality reduction techniques used in the analysis of high-dimensional big data are aimed to be explored. The thesis encompasses studies on outlier detection methods, as well as simulations conducted to test these methods, and the results obtained from the fundamental findings of the study.

February 2024, 71 pages

Keywords: Anomaly Detection, Outlier, Big Data, High-Dimensional Problem, High-Dimensional Data

TEŞEKKÜR

Çalışmalarımın her aşamasında beni yönlendiren, akademik araştırmalarımda derin bilgi ve deneyimlerini paylaşan, geleceğe dair perspektifime büyük katkılarda bulunan değerli danışman hocam Prof. Dr. Olçay ARSLAN'a (Ankara Üniversitesi İstatistik Anabilim Dalı Öğretim Üyesi),

Beni hayata dair önemli düşüncelerle donatan, prensipleri ve örnek davranışlarıyla ışık tutan, zor zamanlarımda doğru yollara yönlendirmesi ve her zaman bana olan güveni ile beni cesaretlendiren, hem insan ilişkilerimde hem de iş hayatımda, geleceğe dair bakış açımın gelişmesinde ve yetişmemde büyük katkıları olan değerli hocam Sayın Dr. Öğr. Üyesi Yetkin TUAÇ (Ankara Üniversitesi İstatistik ABD)'a,

Çalışmalarımda değerli bilgileri, zamanı ve kıymetli yardımlarını esirgemeyen Dr. Öğr. Üyesi Yeşim GÜNEY (Ankara Üniversitesi İstatistik ABD)'e,

Her zaman yanımda olan, bitmek bilmeyen enerjileri ve yüreklendirici sözleriyle beni hep ileri taşıyan kıymetli Onur Yüksek Teknoloji Ailesine ve değerli çalışma arkadaşlarıma,

Birlikte başladığımız bu yolda tüm zorlukları birlikte atlattığımız, her seferinde birbirimizi cesaretlendirdiğimiz, hem akademik hem de manevi olarak bana çok desteği olan sevgili arkadaşım İrem UYSAL'a,

Her zorluğumda yanımda olan, fedakârlıklarıyla beni güçlendiren ve her adımımda duydukları gururla destek olan kıymetli annem Hatice KÖPRÜ'ye, kıymetli babam Erdal KÖPRÜ ve biricik kardeşim Öykü Dilara KÖPRÜ'ye, Bu süreçte ve hayatımda, beni hep destekleyen, varlığıyla güç veren ve her anımda yanımda olan yol arkadaşım Umur ÇIRAK'a sonsuz teşekkürler.

Aslıhan Asena KÖPRÜ
Ankara, Şubat 2024

İÇİNDEKİLER

SİMGELER VE KISALTMALAR DİZİNİ

n	Gözlem Sayısı
p	Değişken Sayısı
Y_i	Yanıt değişkeni
X_{1i}	Bağımsız değişken
β_j	Bilinmeyen parametre
ε_i	Hata terimleri
$E(.)$	Beklenen Değer
$Var(.)$	Varyans
$cov(.)$	Kovaryans
y_i	Aykırı değer skoru

MAD	Ortanca Mutlak Sapma
MMS	Medyan Mutlak Sapma
TBA	Temel Bileşen Analizi
TB	Temel Bileşen
KMO	Kaiser Meyer Olkin
EKK	En Küçük Kareler

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Bu tez kapsamında; büyük veri, yüksek boyutlu veri ve yüksek boyutlu büyük verilerin tanımı, geçmişte yapılmış uygulamaları, diğer disiplinlerle ilişkileri ile ilgili genel bilgiler verilip, boyut indirgeme tekniklerinden bahsedilmiş, yüksek boyutlu büyük verilerde aykırı değerlerin tespit edilmesine yönelik algoritmalar belirlenmiş, gerçek ve üretilmiş bir veri seti üzerinde belirlenen yöntemlerin uygulaması yapılmış sonuçları yorumlanmış ve farklı boyutlardaki veriler için aykırı değer tanılama teknikleri simüle edilerek yorumlanmıştır.

Data kelimesi, İngilizce'de 'bilgi' veya 'veri' anlamına gelir ve geçmişi yaklaşık olarak 1640'lı yıllara dayanmaktadır. Ancak 1946 yılında, bugünkü anlamını kazanarak, aktarılabilir ve saklanabilir bilgisayar bilgisi olarak tanımlanmıştır. Veri işleme terimi ise, veri tanımının hemen ardından, 1954 yılında yaygın olarak kullanılmaya başlanmıştır. Geçen bu süre zarfında veri kavramı, bilgisayar bilimleri ve bilgi teknolojileri alanlarında önemli bir başkalaşım geçirerek, günümüzdeki kapsamlı ve çeşitli anlamlarını kazanmıştır (Yakar, 2022).

Gün geçtikçe artan veri depolama sistemleri nedeniyle, depolanan veri yığınları giderek büyümekte ve bu durum veri analizi için geleneksel yöntemlerin yetersiz kalmasına neden olmaktadır. Bu yetersizliğin üstesinden gelmek ve anlamlı ilişkileri keşfetmek için geleneksel istatistiksel yöntemlerin aksine, yeni yöntemlere ve araçlara da ihtiyaç duyulmaktadır. Bilgisayar bilimi, veri tabanı teknolojileri, istatistik ve diğer disiplinleri birleştiren veri madenciliği kavramını ortaya çıkarmıştır. Veri madenciliği de temelde, çeşitli kaynaklardan elde edilen verilerin işlenerek anlamlı hale getirilmesini amaçlamaktadır.

Dijital çağda, neredeyse her şeyin dijital bir yansıması bulunmaktadır. Ses, fotoğraf, video, belge, telefon bildirimleri, alışverişlerimiz, sık ziyaret ettiğimiz mekânlar gibi yaşamımızın bir parçası olan her şey, aslında dijital dünyada veri olarak birer karşılık bulmaktadır. Veri; günümüzde her şeyin temelini oluşturan bir unsur olmakla birlikte bir konuda doğru sonuca ulaşmak, geleceğe yönelik kararlar alabilmek için verileri

elde etmek kadar analiz etmek de önemlidir. Ancak bu noktada, veri ile bilgi arasındaki ayrımı yapmak oldukça kritiktir

Veri, sadece bilgi sahibi olma sürecinin ilk adımıdır. Veriler, işlendikçe ve analiz edildikçe bilgiye dönüşebilir. Ancak, verilerin ham olması nedeniyle işlenmesi ve anlamlandırılması gereklidir. Veriler işlenmediğinde, karar verme sürecinde rehberlik etme özelliğini kaybeder ve gerçekten bilgi sağlayıp sağlayamadığı belirsizleşir. Bu nedenle, çok miktarda veri her zaman doğru bilgiyi sağlamayabilir (Berksoy, 2019).

Araştırılan bir konunun çözümü, veri toplama teknikleri kullanılarak elde edilen bilgilerle mümkün hale gelmektedir. Bu veriler, ölçme, sayım, deney ve gözlem gibi çeşitli veri toplama teknikleriyle elde edilmektedir (Taşdöğen, 2019). Veriler genellikle niceliksel (sayısal) ve niteliksel (kategorik) olarak iki gruba ayrılır. Yapılacak araştırmanın amacına bağlı olarak doğru veri toplama tekniği, uygun veri seçimi ve uygulanacak analiz yöntemi belirleyici kriterlerdendir. Araştırmaya özgü veri türlerinin belirlenmesi, bu verilerin analizi için kullanılacak yöntemleri de belirler. Bu sayede, araştırmanın hedefine en uygun veri toplama ve analiz stratejileri belirlenebilmektedir (Öztürk, 2006).

Veriler; nicel ve nitel veriler olmak üzere iki temel türde sınıflandırılabilmekle birlikte ölçme, sayım, deney ve gözlem gibi çeşitli veri toplama teknikleri kullanılarak elde edilmektedir. Nicel veriler, aralık veya oranlama düzeyinde yapılan ölçümlerden elde edilir ve sayısal değerlere sahiptir. Örneğin, ağırlık ve boy özelliklerinin ölçümü nicel veriye örnektir. Nitel veriler ise bir kategorik yapıya sahiptir ve genellikle öznellik içerir. Hareket özelliğinin ölçümü, bir kategorik veya niteliksel veri örneğidir. Araştırmanın amacına ve elde edilmek istenen sonuca bağlı olarak, uygun veri seçimi, doğru veri toplama tekniği ve analiz yöntemi belirlenmelidir. Bu noktada, araştırmanın güvenilirliği ve sonuçların doğruluğu için dikkatli bir metodoloji seçimi önemlidir (Taşdöğen, 2019) (Öztürk, 2006).

Nitel veri analizinde sıklıkla kullanılan yöntemler deney ve gözlemdir. Yapılan gözlemler sonucunda elde edilen değerler, nitelikleri anlamak için kullanılır fakat genellikle deney ve gözlem terimleri birbirleriyle karıştırılmaktadır. Deney, belirli bir yöntem ve kurallara uygun olarak gerçekleştirilen bir işlem olup bilimsel bir gerçeği veya varsayımı göstermek, saptamak, doğrulamak veya kanıtlamak amacıyla uygulanmaktadır. Deney, araştırmacının belirli değişkenleri kontrol altında tutarak, neden-sonuç ilişkilerini anlamak veya bir hipotezi test etmek amacıyla yapılan düzenli bir test veya gözlem sürecidir. Bir deney, bir kontrol grubu ve bir deney grubu içerir; kontrol grubu değişkenlere müdahale edilmeden standart şartlarda tutulurken, deney grubu belirli bir değişikliğe maruz bırakılır. Öte yandan, gözlem, olayları, durumları, kişileri veya nesneleri dikkatlice izleyerek, kaydederek ve belirli bir plan olmaksızın veri toplama sürecidir. Gözlemlerde, araştırmacı doğrudan gözlemlenenleri kaydeder, ancak bir kontrol grubu veya deneysel müdahale bulunmaz. Temel fark, deneyin kontrollü bir ortamda yapılan sistemli bir müdahale olduğu, gözlemin ise daha geniş bir perspektiften bilgi toplama amacını güttüğüdür (Morgan, 2001) (Ebeto, 2017).

Araştırma popülasyonun tanımlanması, hangi birimlerden verilerin elde edileceğini ve araştırma sonuçlarına dayalı güncellemelerin kimleri veya neleri kapsayacağını belirlemek anlamına gelir. Popülasyonun doğru şekilde belirlenmesi, elde edilen verilerin temsil edilme düzeyini artırır ve araştırma sonuçlarının genellenmesini güçlendirir. Bu sayede, bilimsel araştırmaların daha kapsamlı ve güvenilir bir temele dayanmasına olanak tanınır (Kılıç & Ural, 2005).

İçinde bulunduğu evreni yeterli düzeyde temsil edecek şekilde oluşturulan veya seçilen kümeye "örneklem" denir ve bu kümeyi seçme işlemine genellikle "örnekleme" adı verilir. 1940'lardan itibaren örnekleme teorisi incelendiğinde, örneklemin avantajları arasında, bilgiyi örnekten elde etmenin bütün popülasyondan elde etmekten daha ekonomik olması ve elde edilecek bilginin evrene göre daha hızlı bir şekilde elde edilmesi yer alır. Bu nedenle, örneklem seçimi, bilgiyi elde etme sürecinde etkili ve pratik bir yöntem olarak tercih edilir (Yamane, 2001).

Pazar araştırmalarında, araştırmacılar çeşitli örnekleme yöntemleri kullanarak eyleme geçirilebilir içgörüler elde ederler. Tam sayım, bir grup içindeki tüm bireyler hakkında bilgi toplama işlemidir ve genellikle nüfus sayımları bu tür bir tam sayım örneğidir. Ancak, zaman ve mali kaynak sınırlamaları nedeniyle tam sayım mümkün olmayabilir. Bu noktada, evreni temsil eden bir alt küme olan örneklemden bilgi toplamak daha yaygın bir yaklaşımdır. Bu yöntem, tüm popülasyonu değil, örneğin özelliklerini ve tepkilerini yansıtan daha küçük bir gruptan bilgi toplamayı içerir. Bu da daha ekonomik, hızlı ve kolay bir yaklaşım sunar (Vaus, 1990).

Ancak, seçilen örneklemin popülasyonu doğru ve güvenilir bir şekilde temsil etmesi için uygun bir yöntemin belirlenmesi önemlidir. Seçilecek bu yöntem, minimum örneklem hatası ile maksimum temsiliyeti sağlamayı amaçlar. Uygun örneklem büyüklüğü, araştırmanın güvenilir ve doğru sonuçlar elde etme hedefiyle doğru orantılıdır. Örnekleme yaparken dikkat edilmesi gereken en kritik faktörler, uygun örneklem büyüklüğü ve uygun örneklem yönteminin seçimidir. Uygun örnekleme yöntemlerini belirlerken, olasılıklı ya da olasılık dışı örnekleme yönteminin seçilmesi, araştırmanın güvenilirliği açısından büyük bir öneme sahiptir. İyi bir örneklemin popülasyonu etkili bir şekilde temsil etmesi, örneklemde eşit şansların bulunması ve örneklem büyüklüğünün minimum örneklem hatası ile belirlenmesi, başarılı bir örnekleme yönteminin önemli unsurlarıdır. (Sarfo, Debrah, Gbordzoe, & Obeng, 2022)

Kullanılacak örnekleme yöntemleri ise, olasılıklı örnekleme ve olasılık dışı örnekleme yöntemleri olarak ikiye ayrılmaktadır.

Olasılıklı örnekleme yöntemleri,

1. Basit Tesadüfi Örnekleme
2. Sistematik Tesadüfi Örnekleme
3. Tabakalı Örnekleme
4. Küme Örnekleme

Olasılık dışı örnekleme yöntemleri,

1. Monografik Örnekleme
2. Kolay Örnekleme

3. Kota Örnekleme
4. Amaçlı (Kasti) Örnekleme
5. Kartopu Örnekleme olarak alt başlıklara ayrılmaktadır.

Olasılık kuramı, belirli özelliklerin evrende normal dağıldığı ilkesine dayanır. Bu teori, olasılıklı örnekleme yöntemlerinin temelini atar ve yansızlık kuralına dayanır. Yansızlık kuralı, örneklem içinde yer alacak her birimin birbirinden bağımsız ve evren içinde eşit seçilme şansına sahip olmasını amaçlar. Bu sayede olasılıklı örnekleme, evrenin çeşitli özelliklerini daha güvenilir bir şekilde temsil etmeyi hedeflemektedir (Ural & Kılıç, 2011).

Veri analizi süreçlerinde sıkça karşılaşılan durumlardan biri, gözlem sayısının yanıt değişkeni sayısından daha fazla olduğudur. Ancak, araştırmalarda katılımcılardan yeterli sayıda yanıt alınamadığında, maliyet ve lojistik zorluklar nedeniyle örneklem çapının yeterli büyüklüğe ulaşamadığı durumlarda, gözlem sayısı yanıt değişkeninden daha küçük olabilir. Bu durumda, yüksek boyutlu veri analizine yönelme ihtiyacı ortaya çıkmakla birlikte, yüksek boyutlu verilerle çalışmak, çeşitli zorlukları beraberinde getirir. Özellikle bağımsız değişken ile yanıt değişkeni arasındaki karmaşık ilişkileri tanımlayabilecek ve öğrenebilecek yeterli bir model bulmak zor olabilir. Bu nedenle, boyut küçültme teknikleri kullanılarak, veri setindeki değişken sayısını azaltmak ve model eğitimini kolaylaştırmak amaçlanır.

İstatistiksel analiz literatüründe, yüksek boyutlu verilerde boyut indirgeme için birçok yaklaşım bulunmaktadır. Bu yöntemler arasında faktör analizleri (FA), kümeleme analizleri, temel bileşen analizleri (TBA) gibi teknikler bulunmaktadır. Gün geçtikçe yüksek boyutlu verilerin daha yaygın hale gelmesi, bu verilere etkili bir şekilde yaklaşmak ve analiz yapmak için boyut indirgeme tekniklerinin önemini artırmaktadır.

Yüksek boyutlu veri ve büyük veri, veri analitiği alanlarında ortak çalışma alanlarına sahiptir. Yüksek boyutlu verilerde genellikle gözlem sayısı, yanıt değişkenine eşit ya da daha büyük olduğu için analiz etme ve veriyi görselleştirme zorlukları ortaya

çıkabilmektedir. Ancak büyük verilerde, hem gözlem sayısı hem de yanıt değişkeni oldukça yüksektir. Bu iki kavram, veri setlerinin boyutları ve özellikleri açısından farklılık gösterse de, her ikisi de büyük miktarda verinin depolanması, yönetilmesi ve analiz edilmesi ile ilgili zorlukları içermektedir.

Veri depolama ve analiz konusundaki ilerlemelere tarihsel bir perspektiften bakıldığında, veri yönetimi ve analizinin köklerinin MÖ 18.000 yılına kadar uzandığı görülmektedir. Ishango Kemiği gibi arkeolojik buluntular, insanların tarih öncesi dönemlerde bile veri depolama ve takip etme ihtiyacını anlamış olduklarını gösterir. Günümüzde, yüksek boyutlu veri ve büyük veri kavramları, teknolojik gelişmelerle birlikte daha fazla önem kazanmış ve bu alandaki analitik yöntemlerin evrimine öncülük etmiştir (Firican, 2021).

Son yıllarda büyük veri; akademik çalışmalarda, sanayinin büyümesinde, hükümetlerin alacağı kararlarda, hemen hemen her yapılanmada kullanılan ve bu hususta gelişen küresel bir güç haline gelmektedir (Manyika, ve diğerleri, 2011). Geleneksel ölçme ve gözlem sistemlerine ek olarak, çeşitli sensörler ve sosyal ağlardan aktarılan büyük verilerin hızla çoğalması, günlük hayatımızın önemli bir parçasını oluşturmaktadır. Büyük verilerin temel özellikleri; verilerin kendisi, veri analitiği ve analizlerin yorumlanmasıdır (Gantz & Reinsel, 2012).

Eğer bir veri kümesi n gözlem ve p değişken (veya özellik) içeriyorsa, bu veri p-boyutlu olarak adlandırılır. Boyutların sayısı arttıkça, veri analizi daha karmaşık hale gelir. Bu durum bazen "boyutlanma laneti" olarak adlandırılır. Yani, veri seti boyutları arttıkça analizdeki zorluk ve karmaşıklık da artar. Bu durum, büyük veri setlerinin yüksek boyutluluğuyla sıkça ilişkilidir. Yüksek boyutlu veri analizi, veri setlerinin boyutlanma laneti etkisini minimize etmek ve anlamlı bilgiler çıkarmak için özel analiz yöntemlerini gerektirebilir (Shin, Hooi, Kim, & Faloutsos, 2017) (Thudumu, Branch, Jin, & Singh, 2019).

Büyük ve/veya yüksek boyutlu veri kümelerinde, genellikle veri setinin genel özelliklerine uymayan ve diğer gözlemlerden önemli ölçüde farklı olan veri

noktalarına "aykırı değer" denir. Aykırı değerler, istatistikte bir veri noktasının diğerlerinden önemli ölçüde farklı olduğunu belirtir. Elde edilen verilerin analiz edilmesi ve sonuç çıkarımı yapılması açısından, aykırı değerler önemli bir rol oynar. Aykırı değerler, tahmincilerin ve bu tahmincilere dayalı testlerin etkinliğini olumsuz etkileyebilir. Ancak unutulmaması gereken önemli bir nokta; her aykırı değerin bir hata olmadığı ve bu değerlerin yüksek varyansın bir işareti olabileceğidir. Aykırı değerler, verinin üretim ve ölçüm süreçlerinden, örneklem hatalarından ve örneklem dağılımına ilişkin varsayımlardan kaynaklanabilmektedir (Aggarwal C. C., 2016).
Aykırı değerlerin belirlenmesi sürecinde değişken sayısının yanı sıra veri hacmi, dağılımı ve değişken sayısı da etkili olmaktadır. Bu unsurlara bağlı olarak, aykırı değerleri tespit etmek için genellikle grafiksel veya istatistiksel yöntemler tercih edilir. Aykırı gözlemlerin belirlenmesi için kullanılacak yöntemler verilerin hacmi, dağılımı ve değişken sayılarına bağlı olarak değişiklik göstermektedir. Ayrıca regresyona dayalı yöntemler, Cook uzaklığı, uyum fark istatistiği, kaldıraç değerleri gibi tekniklerle de aykırı değer tespiti mümkündür. Özellikle çok değişkenli ve yüksek hacimli veri setlerinde, sıfıra yakın veya negatif değerli gözlemlerin bulunması durumunda, gözlemler arası ilişkilere dayalı Mahalanobis ölçütü önerilmektedir (Johnson & Wichern, 2002).

Yüksek boyutlu verilerde ve yüksek boyutlu büyük aykırı değer tespiti yapılırken öncelikli olarak n<p sorununun ortadan kaldırılması gerekir. Fakat burada dikkat edilmesi gereken nokta ise, n<p probleminden kurtulmak için kullanılacak boyut indirgeme yönteminin, aykırı değerler bakımından minimum hata kaybıyla bu işlemleri gerçekleştirebiliyor olmasıdır.

2. YÜKSEK BOYUTLU VERİ

Yüksek boyutlu istatistikler, klasik çok değişkenli analizde genellikle düşünülen boyuttan daha büyük veriyi inceleyen istatistik teorisinde bir alandır. Bu alan, veri vektörlerinin boyutunun örneklem büyüklüğüyle karşılaştırılabilir veya hatta daha büyük olabileceği birçok modern veri setinin ortaya çıkması nedeniyle ortaya çıkmıştır (Wainwright, 2019).

Yüksek boyutlu verilerin ortaya çıkması, değişken sayısının veya özelliklerin sayısının önemli olduğu veri setlerinin zorluklarına yanıt olarak gelişmiştir. Bu gelişmelerle birlikte yapılan çalışmalar, yüksek boyutlu verilerin benzersiz özellikleri ve karmaşıklıklarıyla başa çıkmak için özel metodolojilere ihtiyaç duyan yeni istatistiksel yöntemlerin ve yaklaşımların geliştirilmesine yol açmıştır (Lederer, 2022). Özetle, yüksek boyutlu istatistikler, boyutun önemli bir kriter olduğu, geleneksel istatistiksel tekniklerin kapsamının ötesinde özel metotların gerektiği durumlarda veriyi analiz etmek ve yorumlamak için bir çerçeve sağlar.

Son dönem biyoteknolojileri, özellikle DNA mikroarray teknolojisi ve ileri nesil dizileme cihazları gibi gelişmeler, bireyler üzerinde yüksek boyutlu veri toplama imkanı sunmaktadır. Bu teknolojiler sayesinde, aynı anda on binlerce genin transkripsiyon seviyeleri ölçülebilmektedir. Genomun hemen hemen her bölgesinin transkripsiyon seviyelerini tespit edebilen ileri nesil dizileme cihazları da bu alanda büyük öneme sahiptir. Bu biyoteknoloji verileri, biyolojik düzenleme mekanizmalarını anlamak, hastalıkların genetik temellerini keşfetmek ve yeni ilaçlar geliştirmek açısından büyük potansiyel taşımaktadır. Tıbbi görüntüler, astrofizik gözlemleri, video gözetim görüntüleri gibi farklı kaynaklardan oluşan büyük veri tabanları, genellikle binlerce hatta milyonlarca veri içermektedir. Özellikle tıbbi görüntüler, hastalıkların tanısından tedaviye kadar birçok alanda kullanılmaktadır. Web siteleri gibi platformlar, kullanıcıların tercihleri ve alışkanlıkları hakkında büyük miktarda veri toplar. Bu veriler, pazarlama stratejileri, kişiselleştirilmiş öneriler ve fiyatlandırma politikaları için analiz edilir. Kullanıcıların davranışlarına dair elde

edilen büyük veri setleri, işletmelerin daha etkili stratejiler geliştirmesine yardımcı olabilir.

Lojistik şirketleri ve büyük işletmeler, iç ve dış verileri etkili bir şekilde kullanarak gelecekteki talepleri tahmin etmeye çalışırlar. Bu iş verileri genellikle binlerce değişken içerir ve bu da veri setinin yüksek boyutlu olmasına neden olur. Örneğin, eBird programı, kuş gözlemcilerinin çevrimiçi olarak kuş gözlemlerini kaydetmelerini sağlar ve Kuzey Amerika'daki kuş yoğunluğunu izlemeyi amaçlar. Bu kapsamlı veri setleri, doğa gözlemcilerinin ve bilim insanlarının ekosistemlerdeki değişiklikleri anlamalarına yardımcı olabilir (Donoho, 2000).

Yüksek boyutlu istatistik, bilinmeyen parametre sayısı p'nin örneklem büyüklüğü n'den çok daha büyük olduğu durumlardaki (p > n) istatistiksel çıkarımı ifade ettiği önceki bölümlerde belirtilmiştir. Aşağıda, verilerdeki yüksek boyutun neden bir problem olduğunu kısa bir örnek ile açıklanmıştır;

Bir lineer model olan regresyon modeli,

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \tag{2.1}$$

olmak üzere;

Y_i: Bağımlı değişkeni yani yanıt değişkenini,

X_{1i}:Bağımsız değişkeni,

β_j: Bu değişkenin bilinmeyen parametreleri,

ε_i: Hata terimlerini

ifade etmektedir.

Denklem (2.1) ile verilen modelin matris gösterimi ise,

$y = X\beta + \varepsilon$ biçimindedir.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

y, n x 1 boyutlu gözlem vektörü; X, n x p boyutlu bağımsız değişken matrisi, β, p=k+1 olmak üzere p x 1 boyutlu regresyon katsayılarını ifade eden vektörü, ve ε, (n x 1) boyutlu resgele hataların vektörüdür.

$X'X\,\hat{\beta} = X'y$ olmak üzere, β'nın EKK tahmin edicisinin elde edilme aşamasında eğer, $(X'\,X)$ matrisinin tersi mevcut ise β'nin EKK tahmin edicisi aşağıdaki formül ile elde edilmektedir.

$$\hat{\beta} = (X'\,X)^{-1}X'y \qquad (2.2)$$

Fakat; denklem (2.2)'nin sonucunun bulunması için yapılacak p x p boyutlu $X'X$ matrisinin tersi alınması sırasında n<p durumundan kaynaklı, ilgili matrisin tersi alınamadığı için, n<p problemi bu yöntem için elverişsiz olmaktadır.

Bu tür problemlerin yol açtığı durumlarda geleneksel istatistiksel yöntemler ve veri analizi teknikleri, yüksek boyutlu verilerle başa çıkmak için uygun olmayabilir, çünkü artan öznitelik sayısı bir dizi zorluğa neden olabilir. Boyut sayısının artması, verilerin birbirinden uzaklaşmasına yol açar. Bu nedenle yüksek boyutlu verilerle çalışırken, boyutu küçültme çalışmalarına odaklanmak, daha dayanıklı ve güvenilir sonuçlara ulaşmayı destekleyecektir.

3. YÜKSEK BOYUTLU VERİLERDE BOYUT İNDİRGEME TEKNİKLERİ

Yüksek boyutlu veriler, genellikle birçok değişken içerdiğinden analiz ve görselleştirme açısından zorluklar yaratabilmektedir. Yüksek boyutlu veriler için kullanılan boyut indirgeme yöntemleri, bu tür veri setlerinin işlenmesinde kullanılan güçlü araçlardır. Bu yöntemler, veri setindeki değişken sayısını azaltarak anlamlı bilgileri korumayı hedeflemektedir. Kısaca boyut indirgeme yöntemleri, yüksek boyutlu veri setlerindeki karmaşıklığı azaltarak, önemli bilgileri anlamamıza ve görselleştirmemize yardımcı olur. Bu yöntemler, özellikle makine öğrenimi, veri madenciliği ve görselleştirme gibi alanlarda geniş uygulama alanına sahiptir.

Kullanılan boyut indirgeme teknikleri, analiz yeteneğine, verinin boyutuna, bilgi kaybına, çözümlenebilirliğine göre değişkenlik göstermektedir. Bu yöntemlerden biri olan TBA, veri setindeki değişkenler arasındaki en büyük varyansı ortaya çıkarmayı hedefleyen bir yöntemdir. TBA, veri setini yeni bir koordinat sistemine dönüştürerek, verideki temel desenleri ve ilişkileri anlamak için yaygın olarak kullanılmaktadır.

Bir diğer boyut indirgeme tekniği ise FA'dır. İngiliz psikolog Charles Spearman, 1904 yılında genel zekâ çalışmaları sırasında zihinsel yeteneklerin temelinde yatan zekâ faktörünü inceleyerek FA kavramını geliştirmiş ve Spearman tarafından yapılan bu çalışma, faktör analizi çalışmalarının başlangıcı olmuştur. 1930'ların ortalarında Karl Pearson ile devam eden FA gelişimi, 1950'lerden itibaren bilgisayar teknolojilerinin de gelişmesiyle hız kazanmıştır (Güzeller, Eser, & Aksu, 2017). 20. yüzyılın başlarından günümüze kadar birçok istatistiksel çalışmada kullanılan Faktör Analizi, ilişkili değişkenleri, yani faktörleri, belirli başlıklar altında bir araya getirerek anlamlı yeni değişkenler elde etmeyi amaçlar (Chris, 1997). FA, tüm değişkenlerin birbiri arasındaki korelasyon ve kovaryans matrisinin oluşturulmasıyla başlamaktadır. Faktörlerin sayısı, korelasyon matrisindeki değerlere dayalı olarak belirlenir, böylece değişkenlerin birbirleriyle olan ilişkilerini en iyi şekilde açıklayan faktörler belirlenmiş olur. (Wichern & Johnson, 2007). Kline'a göre faktör analizi, analizin amacına göre açımlayıcı ve doğrulayıcı olmak üzere iki ana kategoriye ayrılır. Açımlayıcı FA'da, değişkenler arasındaki ilişkilerden yola çıkarak faktörleri

bulmaya ve ölçek oluşturmaya çalışılır. Doğrulayıcı faktör analizinde ise açımlayıcı faktör analizi ile elde edilen faktörlerle seçilen ölçeğin yapısının, seçilen veri setine uygunluğu araştırılır (Kline, 1994).

Yüksek boyutlu verilerde FA uygulamak, özel zorluklar ve dikkate alınması gereken önemli noktalar içerir. İlk olarak, faktör sayısının doğru bir şekilde belirlenmesi gerekmektedir. Yetersiz faktör sayısı, veriyi yeterince iyi açıklayamazken, fazla faktör sayısı aşırı uyuma (overfitting) yol açabilmektedir. Elde edilen faktörlerin hangi değişkenleri ne kadar iyi açıkladığını anlamak, analizin etkili bir şekilde yorumlanması ve adlandırılması ve faktörlerin anlamını doğru bir şekilde kavramak için önemlidir.

TBA ve FA dışında, Kümeleme Analizi; benzer özelliklere sahip gözlemleri gruplamak ve veri setindeki yapıları ortaya çıkarmak için etkili bir yöntemdir. Diskriminant analizi, önceden belirlenmiş sınıflar arasındaki farkları vurgulamak ve sınıflandırma problemleri için kullanışlıdır. t-SNE yöntemi ise özellikle görselleştirme yardımıyla kullanılan bir boyut indirgeme tekniğidir, benzer örnekleri bir araya getirir ve farklı örnekleri uzaklaştırarak veri setindeki ilişkileri daha iyi anlamak için kullanılır. Her biri belirli bir analiz hedefine hizmet eden bu yöntemler, veri setinin özellikleri ve analiz bağlamına göre tercih edilmektedir.

3.1 Temel Bileşen Analizi

TBA, çok boyutlu bir veri setinin daha düşük boyutlu bir uzaya izdüşümünü bulma yöntemidir. Bu yöntem, verinin varyansını maksimize etmeyi amaçlamaktadır (Alpaydın, 2010). TBA, birbirine bağlı birçok değişken içeren veri setlerinin daha küçük boyutlara indirme amacını taşıyan bir dönüşüm tekniğidir. Bu yöntem, veri içerisindeki değişiklikleri mümkün olduğunca korumayı ve mevcut verinin minimum sayıda değişkenle ifade edilmesi için en iyi dönüşümün belirlenmesini hedefleyerek veriyi yeni bir koordinat sistemi içinde ifade ederek, veri setindeki değişkenler arasındaki bağımlılığı azaltmaya çalışır (Çilli, 2007).

TBA ile yapılan bu dönüşüm sonucunda elde edilen değişkenlere, mevcut değişkenlerin temel bileşenleri denir. Varyans değeri en yüksek olan bileşenden başlamak üzere, elde edilen temel bileşenler, bu değere göre sıralanır (Sütçüler, 2006). Özetle TBA yönteminin temel amacı, verinin çeşitliliğini daha iyi yakalayacak yeni bir boyut kümesini bulmaktır (Berkhin, 2006). TBA, boyut indirgeme amacının ötesinde, bir dizi farklı alanda kullanılan çok yönlü bir istatistiksel yöntemdir. Veri görselleştirmesi, gürültü azaltma (gereksiz varyansı çıkarma), veri depolama, veri sıkıştırma, veri setindeki değişkenlerin önem sıralamasının yapılması, veriler arasındaki ilişkilerin anlaşılması, büyük boyutlu giriş verilerini daha küçük boyutlu bir uzaya dönüştürerek sinir ağlarının eğitimi, veri madenciliği, görüntü ve örüntü tanıma uygulamalarının geliştirilmesi gibi birçok farklı alanda kullanılmaktadır. Ancak, TAB'nın, ana bileşenlerin varyanslarını maksimize etmek amacıyla veriyi dönüştürdüğüdür. Bu nedenle bazen bazı detayların kaybedilebileceği ve bu durumun analiz sonuçlarını etkileyebileceği unutulmamalıdır.

3.1.1 Temel bileşen analizi uygulaması

2012 yılında St. Lawrence Üniversitesi Jeoloji Bölümü'nden Dr. Jeff Chiarenzelli tarafından Environmental Earth Sciences dergisinde yayımlanan araştırmada, Adirondack bölgesinin kuzeyindeki nehrin kimyasındaki çeşitliliği ve kayaç tamponlaması ile ilişkilendirilmiş element kimyasındaki değişiklikleri kapsamlı bir şekilde incelemektedir (Chiarenzelli, Lock, Cady, Bregani, & Whitney, 2012).

Bu veri seti, Kuzey New York nehirlerinin su kimyasını anlamak amacıyla Grasse, Oswegatchie, Raquette ve St. Regis nehirlerinde belirlenen üç farklı konumda (yukarı akış, orta akış ve aşağı akış) su örnekleri toplanmıştır. Bu çalışmanın temel amacı, yapılacak Temel Bileşen Analizi ile, su örnekleri içerisinde yer alan elementlerin, kaç tanesi bütün bir çalışmayı açıklamaya yetecektir. Bu veri seti; 12 gözlem ve 26 değişkenden oluşmaktadır.

1960 yılında Kaiser, özvektör (açıklanan varyans) değeri 1'in üzerinde olan tüm faktörlerin seçilmesi gerektiğini savunmaktaydı. Diğer bir deyişle, Kaiser'e göre, her bir faktör en azından birim varyansı açıklamalıdır (Kaiser, 1960). Ancak, Jollife, Kaiser'in kriterinin çok katı olduğunu düşünerek, özvektör (açıklanan varyans) değeri 0.7'den fazla olan tüm faktörlerin seçilmesinin diğer yönteme göre daha uygun bir yaklaşım olduğunu savunmaktadır (Jolliffe, 1972).

TBA'da kullanılan ve örneklem yeterliliği ölçüm kriteri olarak adlandırılan KMO (Kaiser-Meyer-Olkin) kriteri (Olkin, 1974) 'e göre KMO katsayısı şu şekilde değerlendirilmektedir;

- 0-0.5 arası: Kabul edilemez
- 0.5: Yeterli
- 0.5-0.7 arası: Orta
- 0.7-0.8 arası: İyi
- 0.8-0.9 arası: Çok iyi
- 0.9 ve üstü: Mükemmel

Bu katsayı, FA için uygunluğu değerlendirmek adına kullanılmaktadır. KMO değerinin yüksek bulunması, veri setinin faktör analizi için uygun olduğunu ifade etmektedir. Yapılan analiz sonucuna göre, ilgili veri seti için KMO değerinin 0.5 olduğu görülmüştür. KMO kriteri dışında, veri setinin TBA için uygunluğunun bir diğer teyidi ise Bartlett testidir.

Bartlett testi ise, bir veri setinin korelasyon matrisinin birim matrise eşit olup olmadığını değerlendiren bir istatistik testidir. Bu test, değişkenler arasındaki korelasyonların istatistiksel olarak anlamlı olup olmadığını belirlemekte kullanılmaktadır.

Testin temel hipotezleri şu şekildedir;

H_0: Değişkenler arasındaki korelasyonlar istatistiksel olarak anlamsızdır; korelasyon matrisi birim matrise eşittir.

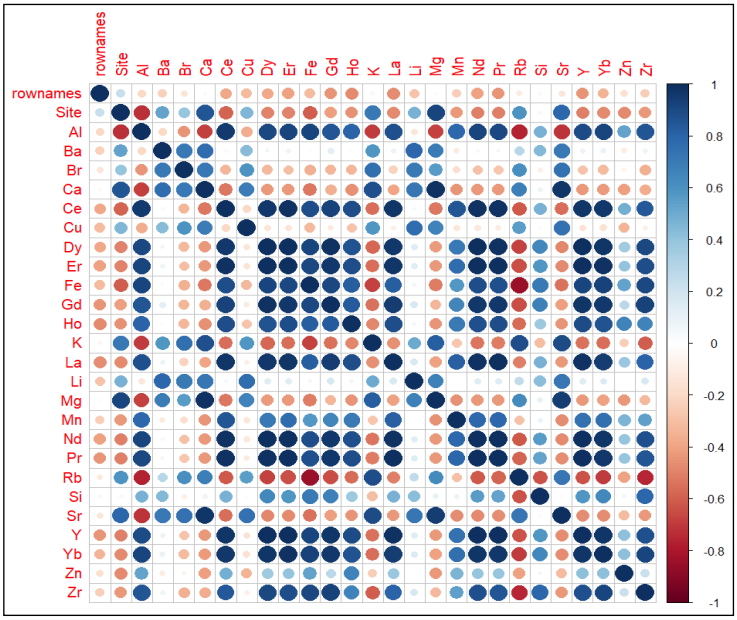
H_1: Değişkenler arasındaki korelasyonlar istatistiksel olarak anlamlıdır; korelasyon matrisi birim matrise eşit değildir.

İlgili hipotez testi, özellikle TBA gibi çok değişkenli istatistiksel teknikler uygulamadan önce değişkenler arasındaki potansiyel bağlantıları değerlendirmek için kullanılır. TBA, değişkenler arasındaki korelasyonları açıklamaya çalıştığı için, bu tür bir bağlantının varlığını kontrol etmek oldukça önemlidir.

Bartlett testinin p değeri, gözlemlenen korelasyon matrisinin birim matrisinden istatistiksel olarak farklı olup olmadığını belirler. Eğer p değeri düşükse, alternatif yani H_0 hipotezi reddedilir ve değişkenler arasında anlamlı bir korelasyon olduğu tespit edilir.

Yapılan test sonucunda ilgili veri seti için $p = 4.633e - 16$ olduğu görülmüştür. H_0 hipotezi reddedilerek, değişkenler arasındaki korelasyonlar istatistiksel olarak anlamlı olduğu ve korelasyon matrisinin birim matrise eşit olmadığı anlamında gelmektedir. Verinin, TBA yapmak için uygun olduğu durumu ifade etmektedir.

Verinin uygunluğu teyit edildikten sonra, değişkenler arasındaki ilişkilerin görselleştirilmesi maksadıyla veriye ilişkin korrelogram grafiği Şekil 3.1'de verilmiştir.



Şekil 3.1 River Elements Veri Seti Korelogram Grafiği

Çizelge 3.1’de, yapılan TBA sonuçlarını göstermektedir. Analizin temel bileşenlerin (TB1, TB2, … ,TB12) standart sapmalarını, toplam varyansın yüzdesini ve kümülatif varyansı açıklamaktadır.

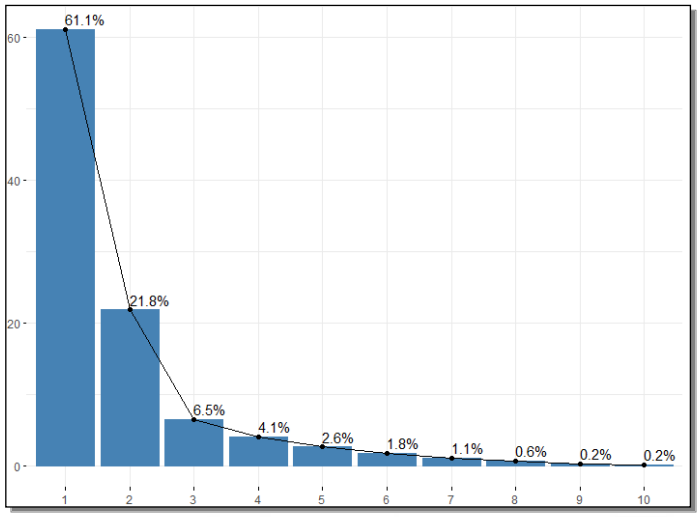
Çizelge 3.1 River Elements Veri seti temel bileşen analizi sonuçları

	TB1	TB2	TB3	TB4	TB5	TB6
Standart sapma	3.97	2.37	1.30	1.03	0.83	0.68
Varyans Oranı	0.62	0.21	0.06	0.04	0.02	0.01
Kümülatif Varyans	0.60	0.82	0.89	0.93	0.95	0.97

	TB7	TB8	TB9	TB10	TB11	TB12
Standart sapma	0.53	0.41	0.24	0.20	0.16	0
Varyans Oranı	0.01	0	0.0	0.0	0.0	0.0
Kümülatif Varyans	0.98	0.99	0.99	0.99	1	1

Çizelge 3.1’de yer alan sonuçlara göre; her bir temel bileşenin standart sapması, o bileşenin veri setindeki değişkenliği ölçmektedir. Örneğin, TB1'nin standart sapması 3.97'dir, bu da TB1'in veri setindeki değişkenliğin diğer temel bileşenlere büyük olduğunu göstermektedir. Aynı şekilde, TB12'nin standart sapması çok düşüktür (yaklaşık *4.633e-16*), bu da bu bileşenin pratikte önemsiz olduğunu ve analizden çıkarılabileceğini göstermektedir.

Elde edilen öz vektörler, yani her bir temel bileşen tarafından açıklanan varyanslar Şekil 3.2’de verilmiştir. Grafikte görüldüğü üzere 3. Temel bileşen bükülme noktasıdır. Bükülme noktasının solunda kalan ilk iki bileşenin,”özdeğer (açıklanan varyans) değeri 1’in üzerinde olan bütün faktörlerin alınmasını savunmaktadır” (Kaiser, 1960). Şekil 3.2 üzerinde yapılan değerlendirmeye göre, %10’un üzerinde olan ilk iki faktör bükülme noktasının solunda yer almaktadır. Bu nedenle, analiz için en uygun olan iki faktörü seçmek, Kaiser'ın kriterlerine uygun bir yaklaşım olacaktır.



Şekil 3. 2 River Elements Veri Seti Öz Değer Grafiği

Her bir temel bileşenin toplam varyans içindeki oranı, o bileşenin toplam değişkenliği ne kadar açıkladığını gösterir. Örneğin, TB1 toplam varyansın %60.87'sini açıklarken, TB12'nin varyans oranı çok düşük ve bu bileşenin önemsiz olduğunu gösterir. Bu durum, analizdeki diğer önemli temel bileşenlere odaklanılması gerektiğini işaret etmektedir.

Kümülatif varyans, belirli bir temel bileşene kadar olan toplam açıklanan varyansı gösterir. Örneğin, TB1 ve TB2 bir arada toplam varyansın %82.65'ini açıklarken, TB1'den TB12'ye kadar olan temel bileşenler kümülatif olarak toplam varyansın %100'ünü açıklamaktadır. Bu da analizdeki genel varyansın ne kadarının açıklandığını göstermektedir. Bu değerlendirmeler ışığında, analizde odaklanılması gereken en önemli iki temel bileşenin TB1 ve TB2 olduğu sonucuna varılabilmektedir.

3.1.2 Temel bileşen analizi simülasyonu

Bu bölümde, temel bileşen analizi yöntemini kullanarak farklı boyutlardaki veri setlerinin analiz edilmesi, rastgele örneklenmiş farklı boyutlardaki matrisler üzerinde

TBA uygulayarak, her bir matrisin kaç temel bileşene ihtiyaç duyduğunun belirlenmesi ve bu işlemi 100 kez tekrarlayarak, farklı matrislerin farklı temel bileşen sayılarına ihtiyaç duyup duymadığını değerlendirmek ve bu analizlerin ortalamasını çıkarmak hedeflenmiştir. Ayrıca, varyansın %70'i açıklanana kadar kaç temel bileşene ihtiyaç duyulduğunu belirleyerek, TBA'nın veri setindeki önemli özellikleri nasıl vurguladığını ve boyut azaltma sürecinde kaç bileşenin kullanılması gerektiğini anlamak amaçlanmaktadır.

Yapılan simülasyon $n = \{10,100,1000\}$ ve $p = \{100,1000,10000\}$ olmak üzere normal dağılımdan rasgele sayı üreterek ikili kombinasyonlar şeklinde uygulanmış ve temel bileşenlerin %70 varyans açıklama oranını elde edildiği temel bileşen sayılarının ortalamaları alınmıştır.

Çizelge 3.3 Farklı n ve p kombinasyonlarına göre TBA simülasyonu sonucu ortalama temel bileşen sayıları

n	p	Ortalama Temel Bileşen Sayısı
10	100	5.87
100	1000	6.02
1000	10000	6.17

TBA simülasyonu sonuçlarına göre, farklı gözlem (n) ve değişken (p) sayılarındaki kombinasyonlara bağlı olarak ortalama temel bileşen sayılarında belirgin bir artış gözlemlenmektedir. Gözlem sayısı ve değişken sayısı arttıkça ortalama temel bileşen sayıları da artmaktadır. Ancak, bu artışın görece düşük olduğu ve ekstra gözlem veya değişken eklemenin temel bileşen sayısını artırmada daha az etkili olduğu görülmektedir. Bu durum, büyük ve yüksek boyutlu veri setlerinin analizinde optimal temel bileşen sayısının seçiminde dikkatli olunması gerektiğini vurgulamaktadır. Genel olarak, simülasyon sonuçları, analizin karmaşıklığı ve veri setinin özelliklerine bağlı olarak temel bileşen sayısının belirlenmesi gerektiğini göstermektedir.

4. BÜYÜK VERİ

Geleneksel istatistiksel yöntemlerin, analiz etmek için yetersiz kaldığı ve karmaşık bir yapıya sahip büyük miktardaki veri setleri için büyük veri terimi kullanılmaktadır (Chen, Chiang, & Storey, 2012). Büyük veri uygulamalarındaki temel amaç, yüksek hacimli ve yüksek boyutlu verilerin yardımıyla kestirimde bulunmak için uygulanabilir çözümler sunmak, değişkenler arasındaki ilişkileri anlamak ve önemli alt popülasyonların benzer olduğu durumları belirlemektir (Badaoui, Amar, Hassou, Zoglat, & Okou, 2017).

Büyük veri, geleneksel veri türlerinden farklı olarak kendine özgü özelliklere sahiptir. Bu özellikler, birbirleriyle ortak özelliklere sahip olmayan ve farklı alt gruplardan gelen verilerin birleşimini içerir (Taylor, Cowls, Schroeder, & Meyer, 2014) . "Büyük verinin temel özelliklerini açıklamak için Laney tarafından 2000li yılların başında öne sürülen "hacim (volume), çeşitlilik (variety), ve hız (velocity)" üçlüsü yaygın olarak bilinmekteydi. Bu üç özelliğin yanı sıra, günümüzde "değişkenlik (variability), doğruluk (veracity) ve değer (value)" özellikleri de eklenerek 6V boyutu ortaya çıkmıştır (Schroeck, Shockley, Smart, Morales, & Tufano, 2013) (Katal, Wazid, & Goudar, 2013)."

- Hacim (Volume): Büyük verinin temel özelliğidir ve veri miktarını ifade eder. Terabayt, petabayt veya ekzabayt gibi büyüklüklerle ölçülür. Bu, verilerin kullanıcılar veya makineler tarafından üretilip üretilmediğine bakılmaksızın büyük miktarlarda olabileceği anlamına gelir.
- Çeşitlilik (Variety): Bu özellik, büyük verinin içerisinde farklı kaynaklardan gelen ve birbirinden farklı yapısal özelliklere sahip veri türlerini içermesini ifade eder. Bu çeşitlilik, farklı veri türlerinin işlenmesi için daha yüksek veri işleme kapasitesine ihtiyaç duyulmasına neden olabilir.
- Hız (Velocity): Verinin işlenmesi gereken hızı ifade eder. Büyük veri, daha önce görülmemiş bir hızda üretilebilecek, depolanabilecek veya yenilenebilecek verileri içerir. İşlenmesi gereken veri miktarının hızlı değişimi bu özelliği belirler.

- Değişkenlik (Variability): Bu özellik, verideki yapısak heterojenliği belirtir. Belirsiz veya tahmin edilemeyen şartlar altında bile verilerin işlenebilir güvenilir olup olmadığını belirlemeye yardımcı olur.
- Doğruluk (Veracity): Verinin güvenilirlik düzeyini ifade eder. Güvenilirlik, hatasızlık ve kesinlik gibi özelliklerle ilişkilidir. Bu özellik, belirli veri türleriyle ilişkili güvenilirlik düzeyini ifade eder.
- Değer (Value): Büyük veriden elde edilen bilgilerin, iş süreçlerine ve karar alma süreçlerine gerçek değer katan bilgiler olup olmadığını ifade eder. Veri analitiği sürecinin sonucunda elde edilen bilgilerin iş süreçlerine değer katma potansiyeli önemlidir (Schroeck, Shockley, Smart, Morales, & Tufano, 2012).

Büyük veri, sadece hacim, çeşitlilik ve hız gibi temel özelliklere değil, aynı zamanda geçerlilik, görselleştirme, güvenlik açığı ve karmaşıklık gibi daha spesifik özelliklere de sahiptir. Bu özellikleri inceleyecek olursak;

- Geçerlilik (Validity): Bu özellik, verilerin hedeflenen kullanım için doğru ve yeterli olduğunu belirtir. Verilerin anlamlı, doğru ve geçerli olması, doğru analizler ve sonuçlar elde etmede önemlidir (Firican, 2017).
- Görselleştirme (Visualization): Bu özellik, büyük veri içindeki karmaşık yapıları ve ilişkileri anlamak için grafik ve görsel araçları kullanma yeteneğini ifade eder. Büyük veri setlerinin karmaşıklığını anlamak ve analiz etmek için görselleştirme önemlidir (Owais & Hussein, 2017).
- Güvenlik Açığı (Vulnerability): Güvenlik açığı, bir kaynağın bilinen zayıflığına işaret eder. Büyük veri setlerinin güvenliği, bu verilere yetkisiz erişimi önlemek ve veri bütünlüğünü korumak için güvenlik önlemlerini içerir (Owais & Hussein, 2017).
- Karmaşıklık (Complexity): Bu özellik, büyük verinin işlenmesindeki zorlukları ve karışıklığı ifade eder. Yüksek boyutlu, çeşitli kaynaklardan gelen ve hızlı değişen büyük veri setleri genellikle karmaşıktır. Veri karmaşıklığı, analiz ve yorumlamada ek zorluklar yaratabilir (Kaisler, Armour, Espinosa, & Money, 2013).

Bu özellikler, büyük veri analitiğinde karşılaşılan zorlukları anlamak ve bu zorluklara karşı etkili önlemler geliştirmek için araştırmacılara rehberlik etmektedir.

Büyük veri uygulamalarındaki başlıca zorluklar; verilerin etkili bir şekilde işlenebilirliği ve depolanması ile ilişkilidir. Verilerin muhafaza edilmesi problemi, gerekli olan donanımın ve veri çözümlemelerinin yetersizliğinden kaynaklanmakla birlikte veri işleme sürecinde de karşılaşılan problemler, bellek sınırları gibi hafıza yetersizliği gibi durumların da, bu verilerin çözümleme sırasındaki zaman kısıtlarının da giderilmesi önem arz etmektedir (Jin, Wah, Cheng, & Wang, 2015). Büyük verilerin belirgin özellikleri, büyük hacmi ve yüksek boyutluluk özellikleridir (Prakasa Rao, 2014). Mevcut özellikler, istatistiksel yaklaşımların yetersiz kalmasına yol açmış ve haliyle yeni istatistiksel analiz yöntemlerinin gelişmesinde önemli rol oynamıştır (Bickel, ve diğerleri, 2009).

5. YÜKSEK BOYUTLU BÜYÜK VERİ

Yüksek boyutluluk problemi, veri kümelerinin analiz için kullanılabilir bağımsız değişkenlerin, bileşenlerin, özelliklerin veya niteliklerin büyük bir sayısına sahip olduğu veri kümelerini ifade etmektedir (Thudumu, Branch, Jin, & Singh, 2019). Boyut sayısının artması, veri analizi karmaşıklığını artırmakla birlikte veriyi işlemek için de daha karmaşık yöntemleri beraberinde getirmektedir. Veri kümeleri, örnek boyutu açısından büyüdükçe, boyut açısından daha da büyümektedir. Aynı zamanda, büyük veri çağında yaygın olarak yüksek boyutlu olarak yanlış anlaşılır, çünkü binlerce boyut çok yaygındır. Bir veri kümesi n gözleme ve p değişkene sahipse, veri p boyutlu olarak adlandırılabilir. Genel olarak, veri kümesi, boyutların sayısı boyutlanma laneti etkisine yol açtığında yüksek boyutlu olarak adlandırılabilir.

Bu durum, veri kümelerinin analiz için kullanılabilir bağımsız değişkenlerin, bileşenlerin, özelliklerin veya niteliklerin büyük bir sayısına sahip olduğu durumu ifade eder. Boyut sayısının artması, veri analizi karmaşıklığını artırırken aynı zamanda daha karmaşık analiz yöntemlerini gerektirir. Özellikle büyük veri çağında, binlerce boyutun yaygın olduğu durumlarla karşılaşılabilir.

Yüksek boyutlu büyük veri, büyük veri kavramının yüksek boyutlu özelliklere sahip verileri içeren bir alt kümesini ifade eder. Büyük veri, genellikle geleneksel yöntemlerle işlenemeyecek kadar büyük, karmaşık ve hızla büyüyen veri kümesini ifade eder. Yüksek boyutlu veri ise, verinin çok sayıda özellik veya değişken içerdiği bir bağlamı ifade eder. Özellikle veri madenciliği ve makine öğrenimi alanlarında kullanılır ve çok sayıda özellik içeren veri kümelerini tanımlar. Özellikle görüntü işleme, metin analizi, biyoinformatik gibi alanlarda karşılaşılır. Bu tür verilerin analizi, özellik seçimi, boyut azaltma ve veri madenciliği yöntemlerini gerektirebilir. Yüksek boyutlu büyük veri analizi, daha fazla bilgi sağlama avantajına sahiptir, ancak boyutun artması analiz süreçlerini karmaşıklaştırabilir.

Boyutsallık laneti terimi, öncelikle Bellman (2013) tarafından kullanılmış olup, boyut veya giriş değişkenlerinin sayısındaki artışın neden olduğu sorunu tanımlamak için

kullanılır. Bu lanet, yükselen boyutlara bağlı olarak anormal durumların düzeyinin gereksiz nitelikler tarafından gölgelenebileceği veya hatta gizlenebileceği bir durumu ifade eder. Anormallik tespiti tekniklerini etkileyen boyutsallık laneti, özellikle aykırı değerlerin seyrek bölgelerdeki veri örnekleri olarak tanımlandığı durumlarda belirgin hale gelir. Yüksek boyutlu uzayın hemen hemen aynı derecede seyrek olduğu bölgelerde, yetersiz bir ayırıcı bölge gözlemlenir. Boyutların artması, veri noktalarını daha dağılmış ve izole hale getirir, bu da veri kümesinin genel optimumlarını bulmayı zorlaştırır (Zhang, Lin, & Karim, 2017) (Aggarwal C. , 2017). Her eklenen boyut, bir veri kümesine yanlış pozitifler getirme eğilimindedir. Bu durum, boyutların artmasıyla birlikte analiz ve modelleme süreçlerini karmaşıklaştırır ve yanlış sonuçlara yol açabilir. Boyutsallık laneti, yüksek boyutlu veri setlerinde anormallik tespiti ve veri analizi konularında dikkate alınması gereken önemli bir zorluk olarak ortaya çıkar (Thudumu, Branch, Jin, & Singh, 2020).

İstatistiksel modeller, artan boyutluluk sorununu aşmak amacıyla çeşitli yöntemler kullanır. Bu yöntemlerden bazıları, işleme süresini azaltmak için önceden belirlenmiş önemli örnekleri seçmeyi içerir (Datta & Kibler, 1995). Boyutsallık laneti, yüksek boyutlu verilerin analizi sırasında ortaya çıkan bir zorluk durumunu ifade eder. Bu durumun temel özelliği, boyut arttıkça uzayın hacminin orantılı olarak büyümesi ve bu durumun, birçok istatistiksel teknik için zorlayıcı bir seyreklik yaratmasıdır. Yüksek boyutlu verilerle başa çıkmak, mesafe ölçüleri gibi istatistiksel yaklaşımların boyutsallık laneti nedeniyle birbirine neredeyse eş uzaklıkta olan noktalardan daha az etkili hale geldiği durumları içerir (Zimek, Schubert, & Kriegel, 2012).

Yüksek boyutlu veri, büyük bir hesaplama belleği ve yüksek bir hesaplama yükü gerektirir. Yüksek boyutlu veri setlerinde kullanışlı bilgileri veya desenleri tanımak karmaşık ve zor hale gelir. Boyutsallık sorununu ele almanın en basit yolu, özellikleri en aza indirmektir. Yüksek boyutlu veriler, genellikle veri madenciliği ve özellikle anormallik tespiti alanında özel zorluklarla karşılaşır. Boyutsal artışa sahip veri kümelerinde anormallik tespiti zordur ve birçok geleneksel veri madenciliği tekniği için ciddi sorunlar doğurur. Boyutsallık lanetinden kaynaklanan sorunlar sadece anormallik tespitiyle sınırlı değildir; aynı zamanda birçok diğer veri madenciliği

tekniğine de uygulanabilir. İstatistiksel modeller, boyutsallık sorununu ele almak için çeşitli yöntemler kullanır, ancak boyut sayısı arttıkça daha fazla hesaplama karmaşıklığı getirir, veri örnekleri dağılmış ve daha az yoğun hale gelir, bu da veri dağılımını etkiler (Hodge & Austin, 2004).

Yüksek boyutlu uzayda seyrek verileri ele almanın daha etkili yollarından biri, veri noktalarının benzerliklerine dayalı fonksiyonları kullanmaktır. Daha büyük bir veri kümesinin küçük kısımlarının değerlendirilmesi, aynı veri kümesinin tamamını incelemek yerine aksi takdirde diğer anormallikler tarafından gizlenmiş olabilecek anormallikleri tanımaya yardımcı olabilir. Bir veri örneğinin veri kümesi içinde diğerlerine benzerliğini ölçmek, düşük boyutlu anormallik tespit prosedürlerinin kritik bir parçasıdır. Bu, olağandışı bir veri noktasının, veri kümesinde benzer örneklerin yetersiz olduğu anlamına gelir.

6. YÜKSEK BOYUTLU BÜYÜK VERİLERDE AYKIRI DEĞER

Aykırı değerler genellikle diğer gözlemlerden belirgin şekilde farklı oldukları için dikkat çekmektedir. Örneğin Hawkins (1980), aykırı değerleri diğer gözlemlerden önemli ölçüde sapma gösteren veriler olarak tanımlar ve bu durum, farklı bir mekanizma tarafından üretildiğine dair şüpheler uyandırabileceğini söyler. Barnett ve Lewis (1978) ise aykırı değerleri, belirli bir popülasyondan alınan ortalama büyüklükte bir örneklem içerisinde, ana gruptan oldukça uzakta bulunan bir veya iki değer olarak tanımlanmaktadır. Aguinis vd. (2013) aykırı değerleri, aynı yapıdaki diğer veri değerleri ile karşılaştırıldığında oldukça büyük veya küçük olan veri değerleri olarak nitelendirir aykırı değerleri, Leys vd. (2013), ise çok büyük artık değerlere sahip veri noktaları olarak belirtir ve bu artık değer, gözlenen değer ile istatistiksel modeller tarafından tahmin edilen değer arasındaki farkı ifade etmekte olduğunu dile getirir.

Veri analizi ve modelleme sürecine başlamadan önce, aykırı gözlemlerin tanımlanması ve ele alınması büyük bir öneme sahiptir. Aykırı gözlemler, bir veri setinde diğer gözlemlerden belirgin şekilde farklı olan gözlemlerdir. Bu farklılık, hatalı ölçüm, beklenmeyen olaylar veya verinin anormal davranışı gibi çeşitli sebeplerle ortaya çıkabilir. Genellikle aykırı değerler hata veya gürültü olarak kabul edilse de, bu değerler içerdikleri bilgiler açısından önemli olabilirler.

En etkili istatistiksel teknikler genellikle otomatik olarak önemli özniteliklere odaklanabilir ve işlenebilir bir süre içinde daha fazla boyutu işleyebilir. Ancak, k-NN, sinir ağları, Minimum Hacimli Elips veya Konveks Soyma gibi birçok teknik boyutlanma lanetine duyarlıdır. Bu yaklaşımlar, genellikle önce belirgin öznitelikleri seçmek için ön işleme algoritmaları kullanır ve ardından aykırı değerleri normal veri noktalarının ana kümesine odaklar.

Aykırı bir gözlem için kesin bir tanım, genellikle veri yapısı ve uygulanan tespit yöntemi ile ilgili gizli varsayımlara bağlıdır. Hawkins (1980), aykırı bir gözlemi diğer gözlemlerden o kadar sapmış bir gözlem olarak tanımlar ki, farklı bir mekanizma

tarafından oluşturulduğu şüphesini uyandırır. Ancak, Bartnett ve Lewis (1994) bir aykırı gözlemin diğer örneklem üyelerinden önemli ölçüde sapmış gibi göründüğünü belirtir, Johnson ve Wichern (2002) ise, aykırı bir gözlemi veri kümesinde uyumsuz olduğu görünen bir gözlem olarak tanımlamaktadır.

Aykırı değer tespiti, bir veri setinde genellikle beklentilere uymayan veya standartlardan sapmış olan özel durumları belirleyen bir yöntemdir. Bu tespit teknikleri, kredi kartı dolandırıcılığı tespiti, klinik denemeler, oy hileleri analizi, veri temizleme, karşı koruma, şiddetli hava tahmini, coğrafi bilgi sistemleri, sporcu performans analizi ve diğer veri madenciliği görevleri gibi birçok uygulama için önerilmiştir. Aykırı değerler, genellikle normallerden sapmış veya istatistiksel olarak anlamlı ölçülerle belirlenmiş anormal veri noktalarını ifade eder. Bu sayede, anormalliklerin tespiti, çeşitli sektörlerde güvenlik, kalite kontrolü ve veri analizi gibi önemli alanlarda kullanılabilecek değerli bilgiler sağlamaktadır (Hawkins, 1980).

Aykırı değer tespit yöntemlerinde öne çıkan teknikler arasında, en önemli boyutları belirleyen değişken seçim yöntemleri ve boyut indirgeme yöntemleri bulunmaktadır (Van Der Maaten, Postma, & Van Der Herik, 2009). Ancak, bu yöntemler, büyük veri setleri ve hızla gelen veri akışlarıyla ilgili zorluklar nedeniyle zorlaşabilmektedir.

Aykırı değerlere yönelik araştırmalar, aykırı değerlerin tespiti ve bunlarla başa çıkma yöntemlerini geliştirmeye odaklanmıştır. Aykırı değerler, genellikle metodolojik veri sorunları olarak kabul edilir ve nasıl ele alınması gerektiği konusunda net bir yöntem bulunmamaktadır. Aguinis ve ekibi (2013) aykırı değerleri sınıflandırarak daha detaylı bir bakış açısı sunmuşlardır. Bu sınıflandırmaya göre, aykırı değerler üç ana kategoriye ayrılmıştır: "hata aykırı değerler", "tahmin aykırı değerler" ve "amaçlanan aykırı değerler". Bu sınıflandırma, aykırı değerlerle ilgili tek bir tanımın veya çözüm yolunun bulunmasının zorluğunu vurgulamaktadır.

Yüksek boyutlu problemler, anormallikleri tespit etmekte zorluklar yaratmanın yanı sıra, verinin hızla arttığı ve sınırsız veri akışları olarak geldiği ortamlarda ek zorluklar da içermektedir. Bu tür problemleri ele almanın yaygın stratejileri, değişken seçim

yöntemi olan en önemli boyutları (yani özelliklerin bir alt kümesini) belirlemek veya boyut azaltma yöntemi olarak bilinen yeni değişkenleri küçük bir veri kümesine dönüştürmektir (Van Der Maaten, Postma, & Van Der Herik, 2009). Ancak, büyük veri setleri ve hızla üretilen veri akışları gibi zorluklar nedeniyle bu stratejiler, anormallikleri tespit etme konusunda karmaşık ve etkisiz hale gelebilmektedir.

Aykırı değerler genellikle metodolojik veri sorunları olarak kabul edilir ve nasıl ele alınması gerektiği konusunda belirsizlikler içerir. Ancak bazı araştırmacılar, aykırı değerleri olağandışı fenomenler olarak görmekte ve bunları anlamak için farklı yaklaşımlar benimsemektedir. Aykırı değerlerin tanımlanması, belirlenmesi ve ele alınması konusundaki karmaşıklığı vurgulayan bir meta-analiz çalışması, aykırı değerlere dair tek bir tanımın veya çözüm yolunun bulunmasının zorluğunu ortaya koymuştur. Aguinis ve ekibi (2013), yaptıkları çalışma kapsamında 46 farklı metodolojik kaynağı ve 232 örgütsel bilim dergisi makalesini inceleyerek 14 farklı aykırı değer tanımı, 39 farklı aykırı değer tespit yöntemi ve 20 farklı aykırı değerlerle başa çıkma yolunu belirlemişlerdir. Bu çeşitlilik, aykırı değerlerin farklı bağlamlarda farklı şekillerde ele alınması gerektiğini göstermektedir.

Veri boyutu arttıkça, yüksek boyutlu problemi ele almak anormallik tespiti konusunda çeşitli zorlukları beraberinde getirir. Anormallikler, yüksek boyutlu uzayda gizlenir ve çoklu yüksek alt uzay projeksiyonlarının içinde saklanabilir. Doğru alt uzayı seçmek, özellikle veri arttıkça, kritik ve karmaşık hale gelir. Bu nedenle, bazı araştırmacılar, alt uzayları tahmin etmek ve yüksek projeksiyonlarının skorlarını hesaplamak gibi yöntemlerle geleneksel anormallik sıralamalarının kalitesini artırmak amacıyla yeni stratejiler geliştirmişlerdir.

Aykırı değer tespiti, veri setinde normal olmayan davranışlar sergileyen gözlemleri tanımlar. Bu tür değerler; anormal değer, aykırı değer, uyumsuz gözlem, hatalar ve benzeri birçok terimle adlandırılabilmektedir. Aykırı değer tespiti, kredi kartı sahtekârlığı, sigorta dolandırıcılığı, vergi kaçakçılığı gibi alanlarda kullanılır ve siber güvenlik, askeri istihbarat, kritik sistem arızalarının tespiti gibi pek çok alanda yaygın olarak uygulanan önemli bir problemi ifade eder. Bu alandaki araştırmalar, aykırı

değerlerin tespiti ve bunlarla başa çıkma yöntemlerini geliştirmeye odaklanmıştır ve bu çalışmalar, çeşitli endüstriyel ve güvenlik uygulamalarında büyük öneme sahiptir. Bu alanlarda, aykırı değerlerin doğru bir şekilde tanımlanması ve analizi, olası risklerin önlenmesine veya anormal durumların belirlenmesine yardımcı olabilir (Chandola, Banerjee, & Kumar, 2009).

Bir veri noktasının neden aykırı bir değer olarak kabul edilmesi gerektiğini belirlemek, analiste özel senaryolarda gerekli teşhis konusunda daha fazla ipucu sağlamak, bir veri noktasının neden aykırı değer olarak kabul edilmesinin önemini vurgular. Yani, belirli durumlarda neden bu veri noktasının normalden sapma gösterdiğini anlamak, analistin daha derinlemesine bir teşhis yapabilmesine ve özel durumları daha iyi anlamasına yardımcı olur. Bu süreç, aykırı değerleri keşfetme süreci olarak adlandırılır veya aykırı değer tespiti ve açıklama süreci olarak bilinir. Farklı aykırı değer tespit modelleri, çeşitli yorumlanabilirlik seviyelerine sahiptir. Genellikle, orijinal özniteliklerle çalışan ve daha az veri dönüşümü kullanan modeller daha yüksek yorumlanabilirliğe sahiptir. Veri dönüşümleri genellikle yorumlanabilirlik maliyetiyle ilişkilidir ve bu nedenle aykırı değer analizi için model seçerken, aykırı değerlerle normal veri noktaları arasındaki farkı artırma etkisi göz önünde bulundurmak kritiktir. Aykırı değer tespiti yöntemlerini literatürde karşılaştıran birçok çalışma bulunmaktadır. Bu yöntemlerin birçoğu, benzerlik ölçüsü olarak Manhattan veya Öklidyen mesafesini kullanır. Ancak, Öklidyen mesafenin benzerliği ölçmek için kullanılması durumunda, sonuçlar birden fazla boyuttan istenmeyen en yakın komşular nedeniyle anlamsız olarak kabul edilebilir (Sadik & Gruenwald, 2014).

Büyük veri kullanımının artması, genellikle veri setlerinin karmaşıklığını ve boyutunu beraberinde getirir. Bu durumda, anormallik tespiti tekniklerinin işleme verimliliğini sürdürmek daha zor hale gelir. Özellikle temel olasılık dağılımının bilinmemesi ve büyük veri kümesinin karmaşıklığı, hesaplama gereksinimlerini artırarak bu tekniklerin daha karmaşık bir hale gelmesine neden olabilir (Gadepally & Kepner, 2014).

Büyük veri kümesiyle çalışırken anormallik tespiti tekniklerinin başlıca zorluğu performanstır. Veri boyutu büyüdükçe, sınırlı hesaplama kapasitesi ve ilişkili faktörler nedeniyle anormallik tespiti teknikleri etkili olmayabilir. Bu bağlamda, aykırı değerlerin doğru bir şekilde tespit edilmesi ve yönetilmesi, istatistiksel analizlerin güvenilirliğini artırmak için kritik bir öneme sahiptir. Aykırı değerleri ele almanın çeşitli yöntemleri arasında düzeltme, dışlama veya özel işleme stratejileri bulunmaktadır. Bu stratejiler, veri setindeki anormallikleri azaltarak, istatistiksel sonuçların daha doğru ve güvenilir olmasına katkıda bulunmaktadır.

Aykırı değer tespiti yöntemleri, tek değişkenli ve çok değişkenli yöntemler olmak üzere genellikle iki ana kategoride sınıflandırılır. Bunun yanı sıra, parametrik ve parametrik olmayan yöntemler olarak da bir başka sınıflandırma yapılmaktadır (Ben-Gal, 2005). İstatistiksel parametrik yöntemler, genellikle belirli bir dağılım varsayımına dayanır ve bu varsayımların en yaygın olanı, gözlemlenen verilerin normal dağılıma sahip olduğu varsayımıdır. Bu normalite varsayımı, istatistiksel analizlerin, regresyon, varyans analizi ve çok değişkenli analiz gibi klasik yöntemlerin temelini oluşturur Ancak, pratikte çalışılan verilerin her zaman normal dağılıma sahip olmadığı durumlarla karşılaşılır. Bu durumda, aykırı değerler ortaya çıkabilir ve normallik varsayımı altında klasik istatistiksel yöntemler üzerinde büyük bir bozucu etkiye sahip olabilir (Maronna, Martin, & Yohai, 2006). Bu varsayımları sağlayamayan gözlemleri aykırı değer olarak tanımlar. Ancak, bu yöntemler çok boyutlu veri setleriyle veya dağılımı bilinmeyen veri setleriyle çalışmada sınırlılıklar gösterebilmektedir.

Verilerde aykırı değer olması durumunda ya da normallik varsayımının sağlanamaması durumunda dayanıklı tahmin ediciler kullanılmaktadır. Dayanıklı tahmin ediciler, verilerin homojen dağılmadığı durumlarda aykırı değerlerin etkilerini azaltmak amacıyla kullanılmaktadır. Kısaca Dayanıklı tahmin ediciler, aykırı değerlerden etkilenmeyen tekniklerdir.

Veri madenciliği alanında, parametrik olmayan aykırı değer tespit yöntemleri genellikle uzaklık temelli yöntemler olarak adlandırılır. Bu yöntemlerde, genellikle

yerel mesafe ölçümleri kullanılır ve bu nedenle büyük veri setlerinde etkili bir şekilde uygulanabilirler. Bir diğer sınıflandırma, küçük boyutlu kümelerde bulunan kümelenmiş aykırı değerlere odaklanan kümeleme tekniklerini içerir. Bu yöntemler, belirli bir kümenin içinde bulunan aykırı değerleri tanımlamak için kullanılır. Uzaysal aykırı değerleri tespit etmek için kullanılan bir başka sınıflandırma ise, tüm veri kümesinden belirgin bir şekilde farklı, ancak komşu değerlere göre uç veya dengesiz olarak nitelendirilen gözlemleri tanımlayan yöntemleri içerir (Ben-Gal, 2005).

Dayanıklı yöntemler, aykırı değerlere karşı daha dirençli olabilir. Eğer veri seti aykırı değerler içermiyorsa, dayanıklı yöntemler klasik yöntemle benzer sonuçlar verebilir. Ancak, az miktarda dahi aykırı değer mevcutsa, dayanıklı yöntemler aykırı değer içermeyen verilere uygulanan klasik yöntemle benzer sonuçlar verebilir (Maronna, Martin, & Yohai, 2006).

Çok değişkenli verilerde, aykırı değerlerin tespit edilmesi için çeşitli yöntemler kullanılabilmektedir. Bu yöntemler arasında Mahalanobis uzaklıklarının hesaplanması, Chernoff yüzleri veya ikon grafikleri gibi grafiksel yöntemler, genelleştirilmiş varyans oranına dayalı kararlar almak gibi yaklaşımlar bulunmaktadır (Alpar, 2013). Bu tür aykırı değer tespit yöntemleri, veri setlerinin doğru analizini ve modelleme süreçlerini sağlamak için önemlidir. Aykırı değerlerin doğru bir şekilde ele alınması, istatistiksel sonuçların güvenilirliğini artırabilir. Aykırı değer tanılama yöntemleri, gelişen ve değişen teknoloji ve alternatif yöntemlerle birlikte parametrik ve nonparametrik (paratmetrik olmayan) olarak ikiye ayırabilmekteyiz.

Parametrik aykırı değer tanılama yöntemleri, verinin normal dağılıma sahip olduğu varsayımına dayanarak belirli bir dağılım şekline göre parametre tahminleri gerçekleştiren analiz yöntemleridir. Bu yöntemler, genellikle veri setinin ortalaması ve standart sapması gibi parametrelerle çalışırlar. En yaygın yöntemlerden biri olan Z Değeri (Z-Score), bir veri noktasının ortalama değerden ne kadar uzak olduğunu standart sapma değeriyle ifade eder ve belirli bir eşik değeri aşıldığında aykırı olarak kabul edilir. Kullanılan yaygın yöntemlerden bir diğeri ise Tukey testidir. Tukey, çeyrekler arası aralıkları kullanarak alt ve üst sınırları belirler; bu sınırların dışındaki

değerler aykırı olarak tanımlanır. Mahalanobis Uzaklıkları, çok değişkenli veri setlerinde aykırı değerleri tespit etmek için kullanılan ve değişkenler arasındaki korelasyonu dikkate alan uzaklığa dayalı bir yöntemdir. Grubbs Testi, potansiyel aykırı değerleri belirlemek için kullanılır ve bir değerin diğerlerine göre uzaklığını değerlendirir. Dixon Testi ise, bir veri setindeki en küçük veya en büyük değerin diğerlerine göre aykırılığını ölçer ve eşik değeri aşıldığında aykırı olarak kabul edilir. Bu parametrik yöntemler genellikle küçük veri setleri ve az sayıda değişken içeren durumlar için kullanışlıdır. Aykırı değer tespiti, istatistiksel analizin güvenilirliği açısından önem taşır ve dikkatlice ele alınmalıdır.

Bu yaklaşımların seçimi, veri setinin özelliklerine ve varsayımlarına bağlıdır. Parametrik yöntemler, genellikle belirli bir dağılımın parametrelerini içeren bir model üzerinde çalışırlar ve bu modeller, genellikle varsayımlara dayanarak verilerin belirli bir dağılıma uyması gerektiği öngörüsünde bulunurlar. Öte yandan, nonparametrik yöntemler belirli bir dağılıma bağlı olmadan verileri analiz etmeye çalışır ve bu esnek yaklaşım, belirli bir dağılım hakkında önceden bilgiye ihtiyaç duymaz. Ancak, bu esneklik bazen verilerdeki düzeni belirlemekte veya belirli özellikleri ortaya çıkarmakta daha az etkili olabilir. Eğer veri seti belirli bir dağılıma uyuyorsa ve parametre tahminleri yapmak isteniyorsa, parametrik yöntemler tercih edilebilir. Ancak, dağılıma ilişkin varsayımlar sağlanmıyorsa veya mevcut dağılım hakkında bir bilgi yoksa nonparametrik yöntemlere başvurmak daha uygundur.

Nonparametrik yöntemler, genellikle daha esnek olmalarına rağmen, belirli dağılım bilgisini kullanmadıkları için daha az hassas olabilirler. Bu durum, analizin veya tahminin belirsizlik derecesinin yüksek olabileceği ve elde edilen sonuçların daha fazla değişebileceği anlamına gelir. Kısacası, belirli bir model veya dağılım hakkında bilgi eksik olduğu için analizde daha fazla belirsizlik olabilmektedir. Nonparametrik yöntemler, özellikle büyük ve yüksek boyutlu veri setlerinde kullanılan esnek analiz araçlarıdır, bu nedenle özellikle makine öğrenmesi algoritmalarında sıkça tercih edilmektedir. Parametrik yöntemlerin ihtiyaç duyduğu spesifik dağılım varsayımları olmaksızın geniş bir uygulama alanına sahiptirler.

K-Ortalamalar (K-Means) yöntemi, belirli sayıda küme oluşturarak veriyi gruplamak için kullanılır. Her kümenin merkezi, o kümenin ortalamasıdır ve veri noktaları bu merkezlere olan uzaklıklarıyla gruplandırılır. Kümeleme yöntemleri, benzer özelliklere sahip veri noktalarını gruplamak için kullanılır ve bu gruplar belirli bir ölçüye göre benzerlik gösteren veri noktalarını içerir (Edward, 1965).

Hiyerarşik kümeleme, veriyi hiyerarşik bir yapıda kümelere ayırmak için kullanılır, kümeleme adımları sırasında benzer veri noktaları bir araya getirilir (Nielsen, 2016).

Spektral Kümeleme, veriyi özellik uzayında temsil ederek ve ardından bu uzayda kümeleme yaparak çalışır. Genellikle graf teorisine dayalı bir yaklaşım kullanırken, Kernel Kümeleme özellik uzayını genişleterek non-lineer ilişkileri keşfetmeyi amaçlar. Çekirdek fonksiyonları kullanarak veriyi yüksek boyutlu uzaya taşır. Wards Kümeleme ise hiyerarşik kümeleme yöntemlerinden biridir ve kümeleme adımlarında küme içi varyansı minimize etmeye çalışır (Luxburg, 2007).

Bu nonparametrik yöntemler, büyük veri setleri üzerinde esnek ve güçlü analizler yapabilme kabiliyetine sahiptir ve genellikle veri içindeki yapıları anlama ve özellik çıkarımı yapma amacına hizmet ederler. Bu yöntemlerin avantajları, parametrik yöntemlere kıyasla daha geniş bir uygulama alanını kapsamaları ve dağılım varsayımlarına ihtiyaç duymamalarıdır.

Çok değişkenli veri setlerinde aykırı değerlerin varlığı, temel veri kümesini değiştirerek anakütle parametre tahminini güçleştirmekte ve kullanılan istatistiksel testin gücünü düşürerek hata varyansını artırmaktadır. Bu nedenle, aykırı değerleri ele almak, veri analizi ve istatistiksel modelleme süreçlerinde kritik bir adımdır. Aykırı değerler, genellikle beklenmeyen veya diğer gözlemlerden belirgin bir şekilde farklı olan veri noktalarını temsil eder. Bu tür değerleri tespit etmek ve uygun bir şekilde ele almak için çeşitli yöntemler bulunmaktadır.

Gözlem sınırlandırma yöntemi, veri setindeki en küçük ve en büyük değerleri belirli bir yüzdeyle sınırlamayı içerir. Bu, veri setindeki aşırı değerleri azaltabilir ve genel

dağılımı düzeltebilir. Bir diğer yaklaşım ise dönüşümlerdir. Logaritmik, karekök veya ters dönüşümler gibi matematiksel dönüşümler, veri setinin dağılımını düzeltebilir ve aykırı değerlerin etkisini azaltabilir.

Winsorizing yöntemi, istatistiksel verilerdeki aşırı değerleri sınırlayarak muhtemel yanlış aykırı değerlerin etkisini azaltma işlemidir. Bu yöntem, belirli bir yüzde dilim içindeki değerleri korurken, diğer değerleri en yakın sınıra getirir. Böylece, aykırı değerlerin etkisi azalırken, diğer veri değerleri korunur. Medyan tabanlı yöntemler, medyan ve MAD (Ortanca Mutlak Sapma) gibi istatistikleri kullanarak aykırı değerlere karşı daha dirençli olabilir (Winsor, Tukey, Mosteller, & Hasting Jr, 1947).

Aykırı değerleri inceleme yöntemi, görselleştirmeler ve özet istatistikleri kullanarak bu değerleri anlamak ve veri setindeki özel durumları keşfetmeyi amaçlar. Hangi yöntemin kullanılacağı, veri setinin özelliklerine, analiz hedeflerine ve aykırı değerlerin ortaya çıkma nedenlerine bağlı olarak değişebilir. Unutulmaması gereken bir diğer önemli nokta, aykırı değerleri sadece temizlemek değil, aynı zamanda veri setinin genel yapısını anlamak için değerlendirmek olduğudur.

Bu tez kapsamında, yüksek boyutlu veri analizine odaklanılırken temel bileşen analizinin önemli bir rol oynadığı belirtilmiştir. Ancak literatürde, yüksek boyutlu veri setlerinin karmaşıklığını daha iyi ele alabilmek ve non-lineer ilişkileri keşfetmek amacıyla farklı boyut indirgeme yöntemlerine giderek daha fazla ilgi gösterilmektedir. Özellikle t-SNE, LLE, İzdüşüm Arama ve Autoencoders gibi non-lineer boyut indirgeme yöntemleri, yüksek boyutlu veri setlerinde daha etkili sonuçlar elde etmek üzere tasarlanmıştır. Bu yöntemler, veri setindeki özellikleri daha iyi yorumlama yeteneği sayesinde temel bileşen analizinden farklı bir bakış açısı sunar. t-SNE, veri setindeki benzer örnekleri bir araya getirme eğilimindeyken, LLE lokal lineer bağlantıları korumaya odaklanır. İzdüşüm Arama ise veriyi bir graf üzerinde temsil ederek benzer örnekleri birbirine yaklaştırmayı amaçlar. Autoencoders ise veriyi kendi temsilini öğrenen yapay sinir ağlarıdır ve non-lineer ilişkileri modellleme konusunda güçlüdürler. Bu çeşitlilik, yüksek boyutlu veri setlerini analiz etmek isteyen araştırmacılara, veri setinin doğasına uygun bir boyut indirgeme yöntemi

seçme esnekliği sağlar. Her bir yöntemin avantajları ve sınırlamaları dikkate alındığında, analistler veri setinin özelliklerine en iyi şekilde uyacak olanı seçebilmektedir. Bu da daha dayanıklı, anlamlı ve detaylı sonuçlar elde etmeye olanak tanımaktadır.

6.1 Aykırı Değer Tespit Yöntemleri

Aykırı değerler, bir veri setindeki diğer değerlerle karşılaştırıldıklarında belirgin bir şekilde daha büyük veya daha küçük olan değerlerdir. Bu değerler, istatistiksel analizlerde belirgin bir etkiye ve haliyle değişime neden olamadıkları gibi, bazen de önemli ölçüde değişim yaratabilmektedir. Veri analizi için bu değerlerin tespiti ve anlaşılması, veri analizinde önemli bir adımdır.

Aykırı değerlerin belirlenmesi, istatistiksel analizlerin doğruluğunu ve güvenilirliğini önemli ölçüde etkileyebilmektedir. Bu değerler, genellikle bir veri grubunun genel eğiliminden sapmış olan veri noktalarıdır. Aykırı değerlerin tanımlanması, veri setindeki anormal durumları tespit etmek ve anlamak için kullanılır. Aykırı değerlerle başa çıkma, öncelikle bu değerlerin neden olabileceği etkileri anlamayı içerir. Eğer aykırı değerler hatalı verilere dayanmıyorsa, bu durum bize süreçlerimizi nasıl iyileştirebileceğimize dair değerli ipuçları verebilmektedir. Aykırı değerleri etkileyen diğer değişkenlerin incelenmesi, maliyet düşürücü veya gelir artırıcı fırsatları ortaya çıkarabilir. Bu nedenle, aykırı değerlerin sadece bir hata olarak değil, süreçlerin anlaşılması ve geliştirilmesi için birer işaret olarak ele alınması önemlidir (Gürsakal, 2001).

Yapılan birçok veri analizinde, büyük bir veri seti oluşturulur, elde edilir veya örneklenir. Tutarsız gözlemlerin, aykırı değerlerin ya da anomalilerin tespiti, tutarlı bir analiz elde etme yolunda atılan ilk adımlardan biridir. Aykırı değerler genellikle bir hata veya gürültü olarak düşünülse de, önemli bilgiler içerebilmektedirler. Modelin yanlış belirlenmesi, yanlış parametre tahmini ve dolayısıyla karşılaşılacak yanlış sonuçlar verinin içerisinde yer alan potansiyel aykırı değerler adaylarıdır. Bu nedenle, modelleme ve analiz öncesinde bunları tanımlamak önemlidir (Williams, Baxter, He, Hawkins, & Gu, 2002) (Liu, Parsons, & Haque, 2004).

Aykırı değerin tam bir tanımı, genellikle veri yapısı ve uygulanan tespit yöntemiyle ilgili gizli varsayımlara bağlıdır. Bu bağlamda, aykırılıkların özelliğini belirlemek, bir dizi tanımlamaya dayanmaktadır. Örneğin, Hawkins (1980), bir aykırı değeri diğer gözlemlerden oldukça fazla olan sapma olarak tanımlar ki bu durumda da bu verinin farklı bir mekanizma tarafından oluşturulmuş olma şüphesini uyandırır. Barnett ve Lewis (1994), aykırı bir gözlemin, diğer örnek üyelerinden belirgin bir şekilde sapma eğiliminde olduğunu belirtir. Diğer durumlarda ise özel tanımlamalar ortaya çıkar. Aykırı değer tespiti yöntemleri, kredi kartı dolandırıcılığı tespiti, klinik denemeler, oy düzensizliği analizi, veri temizleme, ağ sızma, şiddetli hava tahmini, coğrafi bilgi sistemleri, sporcu performans analizi ve diğer veri madenciliği görevleri gibi birçok uygulama için önerilmiştir. Yani, aykırı değer tespiti, geniş bir yelpazedeki analiz ve uygulamalarda önemli bir rol oynamaktadır. Bu yöntemler, veri setlerindeki anormal davranışları belirleyerek, güvenlik, kalite kontrolü, dolandırıcılık tespiti ve benzeri konularda etkili çözümler sunmaktadır (Ruts & Rousseeuw, 1996), (Fawcett & Provost, 1997), (Johnson & Wichern, 1998), (Kay & Jolliffe, 2001), (Acuna & Rodriguez, 2004).

Aykırı gözlemlerin tespit edilmesinde kullanılan yöntemler, uzaklık tabanlı yöntemler ve yansıma yöntemleri olarak 2 ana başlıkta incelenebilmektedir. Uzaklık tabanlı yöntemler, bir noktanın veri kümesinin merkezinden ne kadar uzakta olduğunu ölçerek aykırı gözlemleri tespit etmeye çalışır. Bir veri noktasının aykırılığını ölçmenin yaygın bir yolu, dayanıklı bir Mahalanobis uzaklığı versiyonudur.
Bu uzaklık tabanlı yöntemde, herhangi bir $x_i$veri noktası için "aykırılık" ölçümü (6.1)'de verilmiştir.
$x_i \in R^p, i = 1, \dots, n$ olmak üzere,

$$RD_i = \sqrt{(x_i - T)^T C^{-1} (x_i - T)} \tag{6.1}$$

Burada;
T:veri kümesinin dayanıklı bir konum ölçüsünü yani konum parametresini,
C: varyans-kovaryans matrisinin dayanıklı tahmin edicisini ifade etmektedir.

Fakat uzaklık tabanlı yöntemlerde, güvenilir bir C tahmini elde etmek ve aykırı olarak değerlendirilecek RD_i noktasının ne kadar büyük olması gerektiği gibi problemlerle karşılaşılmaktadır. Bu ölçütler, aykırı değer tespitinde, aykırı değer ile dayanıklı olarak tahmin edilen noktanın arasındaki ilişkiyi için önemlidir. Aykırı değer tespitinde kullanılan dayanıklı konum ve ölçek parametreleri istatistiksel analizlerde aykırı gözlemlere karşı daha dirençli ve güvenilir sonuçlar elde etmek için kullanılır. Bu kısımda bazı önemli dayanıklı konum ve ölçek tahmin yöntemleri ele alınmıştır.

6.1.1 Dayanıklı (Robust) konum tahmini

Bir veri setindeki merkezi eğilimi belirleme amacıyla kullanılan istatistiksel bir yöntemdir. Dayanıklı konum tahmini, özellikle veri setindeki aykırı gözlemlere karşı dirençli olmasıyla bilinir. Aykırı gözlemler, veri setinde diğer gözlemlerden önemli ölçüde farklı olan veya beklenen değerden sapmış olan değerlerdir.

6.1.1.1 Koordinat düzeyi medyan tahmini

Bu yöntem, veri setindeki her bir boyut için medyanı ayrı ayrı hesaplayarak konumu belirler. Koordinat düzeyi medyan, ortogonal dönüşümlere karşı denklemeyen (equivariant) bir özellik göstermez. Yani, veri setindeki bir dönüşüm sonrasında konum tahmini değişebilir. p boyutlu bir veri kümesi için formül (6.2)'de verilmiştir.

$$\hat{\mu}_{coord}(X) = (med(X_1), (med(X_2), \dots, (med(X_p)) \tag{6.2}$$

Koordinat Düzeyi Medyan Tahmininde dayanıklılık değeri veri setindeki aykırı değerlere karşı direnci ölçen bir kavramdır. Koordinat düzeyi medyanın dayanıklılık değeri düşüktür, yani veri setindeki birkaç aykırı değer, medyan tahminini oldukça etkileyebilmektedir.

6.1.1.2 L1 medyan tahmini

Bir veri kümesinin merkezi konumunu dayanıklı bir şekilde tahmin etmek için kullanılan bir istatistiksel yöntemdir. Bu yöntem, veri kümesindeki gözlemlerin merkeze olan mutlak mesafelerini minimize ederek konum tahmini yapar. L1 medyan aynı zamanda "Manhattan Medyanı" olarak da adlandırılır.

L1 medyan, veri setindeki her bir gözlemin konumdan merkeze olan mutlak mesafesini ölçen L1 normunu kullanır. Merkezi konum, bu mutlak mesafelerin toplamını minimize eden değerdir. Formülle ifade edildiğinde:

$$\hat{\mu}_{L1}(X) = argmin_{\mu \epsilon R} \sum_{i=1}^{n} \|x_i - \mu\| \qquad (6.3)$$

Bu minimizasyon problemi, her bir gözlemin tahmin edilen konumdan L1 normu temelinde uzaklığının toplamını minimize ederek en uygun konumu bulmaya çalışır. Burada $\|x_i - \mu\|$, x_i ile μ arasındaki Euclidean normu ifade etmektedir. L1 medyan, özellikle aykırı değerlere karşı dayanıklı olma özelliği ile bilinir. Bu, L1 medyanın, veri setindeki herhangi bir gözlemi dikkate alırken mutlak farklar yerine karesel farkları kullanması nedeniyledir. Aykırı değerlerin etkisi, L1 medyanın tahminini daha dayanıklı hale getirir (Hossjer and Croux, 1995).

6.1.2 Dayanıklı (Robust) ölçek tahmini

Veri setindeki yayılımın, özellikle aykırı değerlerin etkilerini minimize ederek dayanıklı bir şekilde tahmin edilmesini ifade eder. Bu tahmin, veri setindeki ölçek parametresini temsil eder.

6.1.2.1 Ortanca mutlak sapma

Bir verinin ortalama ve standart sapma değerleri hesaplarken kullanılan yöntemlerden olan aykırı değer olması ya da normallik varsayımının sağlanamaması durumunda; ortalama için medyan değeri, standart sapma için ise Ortanca Mutlak Sapma (MAD) değeri kullanılabilmektedir.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6.4}$$

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6.5}$$

$$\tilde{X} = \begin{cases} \dfrac{n+1}{2}, n\ tek \\ \dfrac{\frac{n}{2} + (\frac{n}{2}+1)}{n}, n\ çift \end{cases} \tag{6.6}$$

$$\text{MAD}(x_1, \ldots, x_n) = med(|x_i - med(x)|)1.4826 \tag{6.7}$$

Burada; 1.4826 normal bir dağılımda medyanın ortalama ile olan ilişkisini düzeltmek için kullanılan bir düzeltme faktörüdür ve $|x_i - med(X)|$, x_i ve $med(X)$ arasındaki farkın mutlak değerini ifade etmektedir. Dayanıklı ölçek tahminleri, özellikle veri setinde aykırı değerlerin bulunduğu durumlarda etkili olabilir, çünkü medyan ve medyan mutlak sapmalarının medyanı gibi ölçümler aykırı değerlere karşı dayanıklıdır ve bu durumlarda ortalama ve standart sapma gibi tahminlerden daha güvenilir olabilirler (Rousseeuw & Croux, 1993).

Ortalama ve standart sapma değerlerinin aykırı değerlerden ne kadar etkilendiğini ufak bir örnek ile açıklamak gerekirse; 2, 3, 5, 4, 8, 6, 5, 1, 7, 5, 6, 4, 45, 4, 6, 9, 3, 12, 11, 44, 75, 4, 5, 2, 8, 6, 4 değerleri ile oluşturulan bir veri seti için sırasıyla ortalama, standart sapma, medyan ve MAD değerleri çizelge 6.1.de verilmiştir.

Çizelge 6.1 Örnek veri seti Ortalama, Standart Sapma, Medyan ve MAD Değerleri

	Değer
Ortalama	*10.88*
Standart Sapma	*16.70*
Medyan	*5*
MAD	*2*

Çizelge 6.1'de Ortalama ile medyan arasındaki fark, veri setinin dağılımında aykırı değerlerin etkisi olduğunu düşündürebilir. Eğer ortalama medyandan daha büyükse, bu aykırı değerlerin sağa çekici bir etkiye sahip olduğunu gösterir. Standart sapmanın MAD değerine oranla oldukça yüksek olması, veri setindeki değerlerin geniş bir aralığa yayıldığını ve bu yayılımın içinde aykırı değerlerin etkisinin bulunduğunu gösterir. MAD değerinin standart sapmadan çok daha küçük olması, ortalama ile medyan arasındaki farka rağmen aykırı değerlerin ortalama üzerinde büyük bir etkisi olmadığını gösterir.

Bu karşılaştırmalar, veri setindeki aykırı değerlerin etkilerini değerlendirmeye yardımcı olur. Standart sapma, genel veri setinin yayılımını ifade ederken, MAD değeri, ortalamadan ne kadar uzaklaşan bir örnek değerinin tipik olarak ne kadar olduğunu gösterir. Bu değerlerin karşılaştırılması, veri setinin istatistiksel özellikleri ve dağılımı hakkında daha fazla anlayış sağlar. Ortanca Mutlak Sapma ölçütünün formülasyonu (6.8)'de yer almaktadır.

$$OMS(x_1, \dots, x_n) = med(|x_i - med(X)|)1.4826 \quad (6.8)$$

Burada; 1.4826 normal bir dağılımda medyanın ortalama ile olan ilişkisini düzeltmek için kullanılan bir düzeltme faktörüdür ve $|x_i - med(X)|$, x_i ve $med(X)$ arasındaki farkın mutlak değerini ifade etmektedir. Dayanıklı ölçek tahminleri, özellikle veri setinde aykırı değerlerin bulunduğu durumlarda etkili olabilir, çünkü medyan ve medyan mutlak sapmalarının medyanı gibi ölçümler aykırı değerlere karşı dayanıklıdır ve bu durumlarda ortalama ve standart sapma gibi tahminlerden daha güvenilir olabilirler (Rousseeuw & Croux, 1993).

6.1.2.2 Kantil normalleştirme ile ortanca mutlak sapma

İstatistikte kantil normalleştirme, iki veya daha fazla dağılımın istatistiksel özelliklerini birbirine uyumlu hale getirmek için kullanılan bir tekniktir. MAD'ın ölçeklendirilmiş bir versiyonudur ve özellikle medyan temelli tahminlerle kullanılır. Burada; herhangi bir giriş vektörü x 'in girişin sıralanmış sırasıyla aynı dağılıma sahip olması amaçlanmaktadır. Dağılımları sıralayarak ve değerleri düzenleyerek, ortak bir desene uydurmayı içerir. İki dağılım arasında uygulandığında, test dağılımındaki her bir değer, referans dağılımındaki karşılık gelen değerin sırasına göre ayarlanır. Daha fazla dağılım içeren durumlarda ise tüm dağılımlardaki değerler, karşılıklı ortalaması alınarak düzeltilir. Referans dağılım genellikle standart istatistiksel dağılımlardan biri olarak seçilir ve bu süreç, özellikle mikroarray veri analizinde farklı deneyler arasındaki varyasyonları azaltmak ve verileri tutarlı hale getirmek için sıkça kullanılır. Bu yöntemi basitçe aşağıdaki şekilde ifade edilmektedir.

$\forall\, x \in R^p$ için $f = (f_1, f_2, \dots, f_p)^T$ geçerli bir kantil ise yani x değerleri artan şekilde sıralanmış ise, bu yöntem x vektörünü formül (6.9)'da verildiği gibi şekilde normalize etmektedir (Cabrera & Amaratunga, 2001), (Bolstad, Irizarry, Åstrand, & Speed, 2003).

$$\Phi_f(x) = \prod\nolimits_x f \quad (6.9)$$

6.1.3 PCOut yöntemi

2007 yılında Peter Filzmoser ve diğerleri yüksek boyutlu verilerde aykırı değer tespit yöntemleri üzerine bir çalışma yürüttü. Bu çalışma, yüksek boyutlu verilerde önemli hesaplama avantajları sağlayan, temel bileşenlerin basit özelliklerini kullanarak aykırı değerleri tanımlayan bir yaklaşımdır. Bu yaklaşım, aykırı değer tespiti için mevcut yöntemlere kıyasla önemli ölçüde daha az hesaplama süresi gerektirdiği ve büyük veri setlerinde kullanılmaya uygunluğu açısından elverişlidir (Filzmoser, Maronna, & Werner, 2007).

Algoritmanın temelini iki ana bölüm oluşturur: ilk adım, konum aykırı değerleri tespit etmeye yönelik, ikinci adım ise, saçılım aykırı değerleri tespit etmeye yöneliktir. Saçılım aykırı değerler, geri kalan veriden farklı bir saçılım matrisine sahipken, konum aykırı değerleri farklı bir konum parametresi ile tanımlanır. Algoritmaya başlamak için, her bir bileşeni, koordinat düzlemindeki medyan ve Medyan Mutlak Sapma (MMS) kullanarak dayanıklı bir şekilde ölçeklendirmek veya normalize etmek faydalıdır. Bu işlem, verinin her boyutunu, medyan ve MAD kullanılarak standartlaştırarak, veri setini dairesel bir yapıya sahip ve istatistiksel olarak dayanıklı hale getirir.

(i) Algoritmanın ilk adımında, her özellik(sütun) için koordinat bazında medyan ve MAD değerleri kullanılarak ölçeklendirilmektedir.

$$x_{ij}^{*} = \frac{x_{ij} - med(x_{1j},\dots,x_{nj})}{MAD(x_{1j},\dots,x_{nj})}, \qquad j = 1, \dots, p \tag{6.10}$$

Ölçekleme veya küreselleme işlemi, her bileşenin değerini o bileşenin medyan değeri ile ölçeklemek veya küreselleştirmek demektir. Koordinat bazında medyan ve MAD, aykırı değerlere karşı dayanıklı istatistiklerdir. Her öğe için, o öğenin medyanından çıkarılır ve bu fark, o özelliğin MAD değerine bölünür. Bu tür, ölçekleme ya da küreselleştirme yapılan adımların amacı, veri setindeki aykırı değerlere karşı daha dayanıklı olmaktır. Medyan ve MAD kullanılarak yapılan ölçekleme işlemi, veri

setindeki aykırı değerlere karşı daha az duyarlıdır, bu da daha güvenilir sonuçlar elde etmeyi amaçlar. Bu tür bir ön işleme, aykırı değerlerin analizini ve tespitini daha güvenilir hale getirebilmektedir.

Eğer ölçekleme işlemi yapılırken, bir boyutun MAD değeri sıfıra eşit çıkar ise, o boyut çıkarılır veya başka bir ölçek ölçüsü kullanılır.

(ii) Adım (i) de belirtilen işlem sonucunda elde edilen veri, x_{ij}^{*} ağırlıklı bir kovaryans matrisiyle ölçeklenir. Kovaryans matrisinin her bir elemanı, ilgili iki özellik arasındaki ilişkiyi ölçer. ağırlıklandırılmış kovaryans matrisinden öz değer ve öz vektörler hesaplanır. Hesaplanan bu öz değer ve öz vektörler kullanılarak yarı-dayanıklı bir temel bileşen analizi yapılır. Özdeğeler, verinin hangi yönlere ne kadar yayıldığını gösterir. Öz vektörler, bu yayılmanın hangi yönlere olduğunu gösterir.

Burada dikkat edilmesi gereken en önemli nokta; yapılan temel bileşen analizi sonucunda elde edilecek temel bileşenlerin, yani $p^{*}{}'$ların toplam varyansa katkılarının en az %99 olması gerektiğidir. Bu temel bileşen analizi sayesinde yüksek boyutluluk yani $p \gg n$ sorunu ortadan kalkmış olacak ve oluşturulan yeni matrisin boyutu $p < n$ haline gelecektir.

(iii) Bir önceki adımda elde edilen öz vektörlerden $p^{*} \times p^{*}$ boyutunda yeni bir veri matrisi oluşturulmaktadır. Bu veri matrisi ise V olmak üzere;

$$Z = X^{*}V \tag{6.11}$$

(iv) X^{*}matrisi, x_{ij}^{*} olmak üzere, her bir temel bileşeni tekrar medyan ve MAD değerlerini kullanarak ölçeklendirme işlemleri yapılmaktadır ve aşağıdaki denklemle ifade edilmektedir,

$$z_{ij}^{*} = \frac{z_{ij} - med(z_{1j},\ldots,z_{nj})}{OMS(z_{1j},\ldots,z_{nj})}, \quad j = 1,\ldots,p^{*} \tag{6.12}$$

(v) Yukarıdaki ön işleme adımlarının ardından, konum aykırı değer aşaması, her bileşen için bir dayanıklı kurtosis ölçümünün mutlak değerini hesaplayarak başlatılır. Kurtosis, bir değişkenin uç noktalara ne kadar duyarlı olduğunu ölçen bir istatistiksel ölçüdür.

(vi) Dayanıklı kurtosis ölçüsü, aykırı değerlere karşı dayanıklı bir tahmindir. (Pena & Prieto, 2001). Kurtosis katsayısının mutlak değeri, hem küçük hem de büyük kurtosis değerlerinin aykırı değerlere işaret edebileceği gerçeğini yansıtır. Bu nedenle, değerlerin negatif ve pozitif etkilerini birleştirmek için mutlak değer kullanılır.

$$w_j = \left| \frac{1}{n}\sum_{i=1}^{n} \frac{z_{ij}^{*} - med(z_{1j}^{*},\ldots,z_{nj}^{*})^4}{MAD(z_{1j}^{*},\ldots,z_{nj}^{*})^4} - 3 \right|, \; j = 1,\ldots,p^{*} \tag{6.13}$$

Bir veri seti içerisinde aykırı bir değer bulunmuyorsa, beklenen kurtosis değeri sıfıra yakın olacaktır. Ağırlıklar, her bir bileşenin beklenen kurtosis değeri ile orantılı olarak normalleştirilir. Bu, beklenen kurtosis değeri sıfıra yakın olan bileşenlere daha düşük ağırlıklar atanmasını sağlamaktadır.

(vii) Bu adımda her bir temel bileşene, kurtosis katsayısının mutlak değeri ile orantılı yani aykırı değerleri ortaya çıkarma olasılıklarına göre ağırlıklar atanır. Bu temel bileşenlere göre ağırlıklar, $w_j / \sum_i w_i$ formülü ile atanır. Bu atama, her bir temel bileşenin, genel toplam içerisindeki ağırlığını ifade etmektedir ve atanan her bir ağırlık her bir bileşenin aykırı değerleri belirleme konusundaki etkinliğini yansıtır.

Mahalanobis mesafesi, bir nokta P ile bir dağılım D arasındaki mesafeyi ölçen bir metriktir ve Mahalanobis (1936) tarafından ölçümlere dayalı olarak kafataslarının benzerliklerini belirleme probleminden kaynaklanmıştır. Bu mesafe, P 'nin D 'nin

ortalamasından kaç standart sapma uzaklıkta olduğunu ölçme fikrinin çok boyutlu bir genelleştirmesidir. Eğer bir veri noktası yani P, bir veri kümesinin yani D 'nin ortalamasında bulunuyorsa, Mahalanobis mesafesi sıfır olacaktır. Biraz daha basit bir ifadeyle, Mahalanobis mesafesi, bir veri noktasının merkezi konumuyla ilgili bir ölçüdür. Eğer bu mesafe sıfır ise, demek ki o veri noktası, veri kümesinin merkezine, yani ortalamasına eşittir.

Bununla birlikte, Maronna ve Zamar (2002)'de yaptığı çalışmalardan esinlenerek, 2007 yılında Peter Filzmoser vd, dayanıklı mesafelerin kullanılmasının bir yöntemi olduğunu, bu yöntemin; Mahalanobis mesafelerini daha dirençli bir şekilde kullanarak, aykırı değerlerle başa çıkma ve doğru bir sınıflandırma elde edileceğini ifade etmektedirler.

Bu işlemler, aykırı değerlere karşı dirençli bir mesafe ölçüsü ve ağırlık atama yöntemi oluşturmayı amaçlar.

(viii) Bu adımda, her bir veri noktası için ampirik mesafe (6.14)'de belirtildiği şekilde hesaplanır.

$$d_i = RD_i . \frac{\sqrt{\chi^2_{p^*,0.5}}}{med(RD_1,\dots,RD_n)} \quad i = 1,\dots,n \tag{6.14}$$

Burada;

d_i: Dönüştürülmüş mahalanobis uzaklığını,

RD_i: İlk aşamada (6.1) belirtilen dayanıklı Mahalanobis uzaklığını ifade etmektedir.

Ampirik mesafelerin medyanının, teorik mesafelerle aynı olmasına yardımcı olacak bir düzenleme yapılır. Bu, $\chi^2_{p^*}$, $\chi^2_{p^*,0,5}$ 'e yani $\chi^2_{p^*}$'ın %50'lik kantiline ($\chi^2_{p^*}\ 50^{th}$)karşılık geldiği bir ölçümle yapılır. Ağırlıklar, her bir gözleme ayrı ayrı atanır ve bu ağırlıklar aykırılığı ölçen bir metrik olarak kullanılır. Gözlemlere ağırlıklar atanırken, çevrilen biweight fonksiyonu kullanılır (Rocke, 1996). Tukey'nin biweight fonksiyonuna benzerdir ancak ψ fonksiyonu, orijinden belirli bir

M noktasından itibaren yükselmeye başlar. Yani, konum tahminine olan ölçülü mesafeden daha yakın olan gözlemler, 1 ağırlığını alır.

(ix) Aşağıda, çevrilen biweight fonksiyonunun ψ fonksiyonu ve buna karşılık gelen ağırlık fonksiyonu verilmiştir

$$\psi(\mathrm{d};\mathrm{c},\mathrm{M}) = \begin{cases} d, & 0 \le d \le M \\ d\left(1-\left(\dfrac{d-M}{c}\right)^2\right)^2, & M \le d \le M+c \\ 0, & d > M+c \end{cases} \qquad (6.1.5)$$

d, gözlemlerin bir noktaya olan mesafesi,
M, orijinden başlayarak ψ fonksiyonunun yükselmeye başladığı nokta,
c, bir ölçek parametresidir.

Bu fonksiyonlar, gözlemlerin bir noktaya olan mesafesine bağlı olarak ağırlıkların nasıl değiştiğini tanımlar. İki durumda da (0 ile M arasında ve M ile $M + c$ arasında), ağırlıkların belirli bir desende değiştiği görülmektedir.

Aykırı olmayan verilere tam ağırlıkların 1 (bir) atanması, bilinen aykırı değerlere 0 (sıfır) ağırlıkların atanması, tahmin edicilerin etkinliğinin artmasına ve aynı zamanda hesaplama hızının artmasına neden olur (eğer sınıflandırmalar doğru ise), ancak bu iki uç nokta arasında, genellikle kullanılan biweight fonksiyonuna benzer ağırlıklar alan bir alt küme de bulunmaktadır. Yüksek bir direnç düzeyini sürdürmek için, ağırlıkların bir kısmına bir atama yapmak en iyisidir, çünkü bir veya daha fazla ağırlığı bir (veya yakınında) olan aykırı değerler, maskelenme etkisi nedeniyle diğer aykırı değerlerin tespitini zorlaştırabilir.

$$\omega(\mathrm{d};\mathrm{c},\mathrm{M}) = \begin{cases} d, & 0 \le d \le M \\ \left(1-\left(\dfrac{d-M}{c}\right)^2\right)^2, & M \le d \le M+c \\ 0, & d > M+c \end{cases} \qquad (6.16)$$

İki aşırı durum arasında, genellikle kullanılan biweight fonksiyonuna benzer ağırlıklar alan bir alt küme bulunmaktadır. Yüksek bir direnç seviyesi korunmalıdır, çünkü ağırlıkların bir kısmına yüksek değerler atandığında, bu durum diğer aykırı değerleri tespit etmeyi zorlaştırabilir. Temel bileşenlerin median ve MAD tarafından ölçeklendirilmiş olduğu hatırlanmalıdır, bu nedenle bu dayanıklı mesafeler, median'a MAD'nin dönüştürülmüş birimlerini kullanarak) olan ağırlıklı bir mesafeyi ölçer. İyi bir deneysel sonuç elde etmek için, en küçük dayanıklı mesafeye sahip olan noktalara bir ağırlık olarak 1 atamak olacaktır.

(x) Ağırlıklandırma şemasının diğer ucunda, c'ye göre ağırlığı sıfır olarak atılan noktalara atanır.

$$c = med(d_1, \dots, d_n) + 2.5\ OMS(d_1, \dots, d_n) \tag{6.17}$$

Bu denklemde $\{d_1, \dots, d_n\}$ gözlemlerin dayanıklı mesafelerini temsil etmektedir.

Geleneksel aykırı değer sınırlarına yaklaşık olarak karşılık gelen bir c değişkeni kullanılarak, her bir gözlem için çevrilen biweight fonksiyonu tarafından hesaplanan ağırlıklardan bahsedilmektedir. Bu ağırlık atama şeması, aykırı olmayan verilere tam ağırlıklar, aykırı değerlere sıfır ağırlıklar ve bu iki uç nokta arasındaki noktalara çevrilen biweight fonksiyonu tarafından belirlenen ağırlıklar arasında bir geçiş sağlamaktadır.

(xi) Ağırlıklar, formül (6.18)'de belirtildiği gibi hesaplanır.

$$\omega_{1i} = \begin{cases} 0 & d_i \geq c \\ \left(1 - \left(\frac{d_i - M}{c - M}\right)^2\right)^2, & M \leq d_i \leq c \ \ d > M + c \\ 1, & d \leq M \end{cases} \tag{6.18}$$

Burada $i = 1, \dots, n$'inci gözlemin bir noktaya olan mesafesini $\{d_1, \dots, d_n\}$ temsil eder, M mesafelerin 3. çeyreği $33\frac{1}{3}^{rd}$ ve c bir ölçek parametresidir. Bu ağırlık atama şemasının avantajı, aykırı olmayan noktalara tam ağırlıklar, aykırı değerlere sıfır ağırlıklar ve aradaki noktalara çevrilen biweight fonksiyonu tarafından belirlenen ağırlıklar arasında geçiş yapabilmesidir. Bu sayede, muhtemelen aykırı değerler içeren bir alt küme sıfır ağırlıklar alabilir ve potansiyel aykırı değerlerin istenmeyen etkilerini önleyebilir. Denklem (6.18) $\{\omega_{1i}\}$ saklanır ve algoritmanın sonunda tekrar kullanılacaktır.

Algoritmanın ikinci aşaması, birinci aşamaya benzer ancak kurtosis ağırlıklandırma şemasını kullanmaz. Temel bileşenler, büyük varyansa sahip olan yönlere odaklanır, bu nedenle bu bölümün başında tanımlanan yarı-dayanıklı temel bileşen uzayında dağılım aykırılarını aramak için iyi sonuçlar bulduğumuzu görmek belki de şaşırtıcı değildir. Önceki gibi, temel bileşen uzayındaki veriler için Öklidyen normu hesaplamak, orijinal veri uzayında Mahalanobis mesafesine eşdeğerdir, ancak hesaplaması daha hızlıdır.

(xii) Son olarak, bu 2 adımdan elde edilen ağırlıkları final ağırlıkları ω_i hesaplamak için (6.19)'da belirtilen hale getirilmektedir.

$$\omega_i = \frac{(\omega_{1i} + s)(\omega_{2i} + s)}{(1 + s)^2} \tag{6.19}$$

Burada ölçek sabiti $s = 0{,}25$ olarak ayarlanır. s'nin tanıtılma nedeni, bazen iki adımdan birinde çok sayıda aykırı olmayan değerlerin ağırlığını 0 aldığı; $s \neq 0$ ayarlanarak, son ağırlığın yalnızca her iki adımda da düşük bir ağırlık atandığında $\omega_i = 0$ olmasına yardımcı olur. Aykırı değerler, bu durumda ağırlığı $\omega_i < 0.25$olan noktalar olarak sınıflandırılır. Bu değerler, ω_{1i}veya ω_{2i}'den biri 1'e eşitse, diğerinin 0.0625 değerinden küçük olması gerektiğini ima eder. Veya $\omega_{1i} = \omega_{2i}$ ise, bu ortak değerin 0.375 değerinden küçük olması gerekmektedir ki x_i aykırı olarak sınıflandırılsın.

6.1.4 Sign Yöntemi

Spatial Sign terimi, Türkçe'de Uzamsal İşaret veya Uzamsal Belirteç olarak ifade edilebilir. Bu yöntem, değişkenli verilerdeki gözlemlerin yönünü belirlemek için kullanılan bir tekniktir. Bu yöntem, veriyi normalize ederek (küreselleştirerek) ve ardından her bir gözlem için bir spatial sign vektörü oluşturarak çalışır. Özellikle aykırı değer tespiti, temel bileşen analizi ve benzeri çok değişkenli analiz uygulamalarında kullanılır. Normalize edilmiş ve spatial sign vektörleri, verideki aykırı gözlemlere karşı daha dirençli olabilir.

Spatial signs yöntemi, temelde çok değişkenli verilerdeki aykırı değerlere karşı dirençli bir analiz yapmak amacıyla geliştirilmiştir. Ancak, genellikle spesifik bir tarihçesi ya da belirli bir kişinin isminin bu yöntemle özdeşleştiği bir bilgiye rastlamak zordur

Spatial signs yöntemi, çok değişkenli veri analizi alanındaki literatürde özellikle 1990'lı yıllardan itibaren yaygın olarak kullanılmaya başlanmıştır. Bu yöntemin gelişimi ve uygulama alanları, farklı istatistikçiler ve araştırmacılar tarafından bir araya getirilen farklı matematiksel kavramların birleşimi olarak görülebilir.
Sign yöntemi, dayanıklı temel bileşen analizi temel alır. Veriyi bir küre üzerine projekte ederek (veya ölçekleri farklı ise bir elips üzerine) dayanıklı konum ve yayılım tahminleri elde eder. Bu sayede aykırı gözlemlerin etkisi sınırlanır ve elde edilen ortalama ve kovaryans matrisi dayanıklı bir matris haline gelir.

PCOut yöntemi ile benzer şekilde, Sign yöntemi de bir tür dayanıklı temel bileşen analizine dayanmaktadır. Veriyi bir küre üzerine (veya ölçekler farklıysa bir elips üzerine) projekte ederek konum ve yayılımın dayanıklı tahminlerini elde eder. Bu şekilde, aykırı gözlemlerin etkileri sınırlıdır çünkü bunlar elipsoidin sınırına yerleştirilir ve elde edilen ortalama ve kovaryans matrisi dayanıklıdır. Standart temel bileşen analizi, bu nedenle küreselleştirilmiş veriler üzerinde herhangi bir tek bir noktanın (veya küçük bir nokta kümesinin) aşırı etkisi olmadan gerçekleştirilebilir.

Bu yöntem temelde 4 adımdan oluşmaktadır. İlk adım veriyi ölçekleme(normalize etme), ikinci adım uzamsal belirteçlerin tespit edilmesi, üçüncü adım aykırı değerlerin tespit edilmesi ve son olarak eşik değerine göre karar verme adımıdır. Bu adımları sırasıyla inceleyecek olursak;

(i) Her bir değişkeni veri setinin ortalamasından çıkartarak ve standart sapması ile bölerek normalize etmektir. Bu, veri setini merkezileştirir ve ölçekler, böylece aşırı değerlerin etkisi azalır.

$$\mathcal{Z}_i = \frac{x_i - \bar{x}}{\|x_i - \bar{x}\|}, i = 1, \dots, n \tag{6.20}$$

$\mathcal{Z}_i$: i-inci gözlemin spatial sign vektörü,

x_i: i-inci gözlem,

$\bar{x}$: Veri setinin ortalaması,

$\|x_i - \bar{x}\|$: i-inci gözlem vektörünün Euclidean normunu ifade etmektedir.

(ii) Normalize edilmiş vektör, birim çember üzerinde bir noktayı temsil eder. Bu noktanın konumu, spatial sign olarak adlandırılır ve bu, orijinden bu noktaya çizilen vektördür.

$$\mathcal{Z}_i = \frac{z_i}{\|z_{i_i}\|}, i = 1, \dots, n \tag{6.21}$$

(iii) Bu adımda ise aykırı değerleri tespit etmek için aşağıdaki formül kullanılmaktadır. A, ilgili verinin kovaryans matrisi olmak üzere;

$$y_i = \mathcal{Z}_i{}^T A \, \mathcal{Z}_i, i = 1, \dots, n \tag{6. 22}$$

y_i: i-inci gözlemin aykırı değer skorunu ifade etmektedir.

Daha sonra, bu skorlar ki-kare dağılımının kritik değerleriyle karşılaştırılır. Aykırı değer tespiti için bir eşik belirlenir. Genellikle, kritik değer belirleme adımı şu formülle ifade edilir, $\mathcal{X}^2_{p^*,0,5}$ ve bu kritik değer, ki-kare dağılımının belirli bir yüzde noktasındaki değeri temsil eder ve belirli bir güven düzeyine dayanarak aykırı değerleri belirlemede kullanılır.

7. UYGULAMA

Bu bölümde, önceki bölümde anlatılan iki yöntemin (PCOut ve Sign) algoritmalarına ilişkin uygulamalar gerçek bir veri seti ve rasgele olarak normal dağılımdan üretilen bir veri seti üzerinde yapılmıştır. Bu çalışmalar, R Core Team 2013'te (Filzmoser & Gschwandtner, 2021) tarafından 2021 yılında geliştirilen yapılmış, mvoutlier paketi içerisindeki PCOut ve sign1 komutları ile elde edilmiştir. Elde edilen sonuçlar ayrıntılı bir şekilde incelenerek yorumlanmıştır.

7.1 Gerçek Veri Seti Uygulaması

Bu bölümde, 3.1 bölümünde incelenen Elements veri seti üzerinde n=12 ve p=26 boyutlarındaki veri için temel bileşen analizi yapılmıştır. Aykırı değer durumlarını oluşturmak amacıyla, 12x26 boyutunda 6 adet vektör elde edilmiş ve her biri mevcut veride yer alan ilk 6 gözlemin değerleri ile çarpılarak oluşturulmuştur. Bu vektörler daha sonra, Elements veri setine eklenmiştir.

Sonrasında, elde edilen n=18 ve p=26 boyutlu veri seti üzerinde PCOut ve Sign yöntemleri uygulanmıştır. Ancak, boyutsallık problemini aşmak için önce temel bileşen analizi yapılmıştır. İlk 8 temel bileşen ile %99 varyans açıklama oranına ulaşılmıştır. Daha sonra, bu temel bileşenlere göre PCOut ve Sign yöntemleri uygulanmıştır.

Sonuçlar elde edilen temel bileşenlere dayalı olarak yorumlanmıştır. Her iki yöntemin de uygulanması ve sonuçların değerlendirilmesi, temel bileşenlerin verideki önemli varyansı nasıl yakaladığını ve aykırı değerleri nasıl etkilediğini göstermektedir.

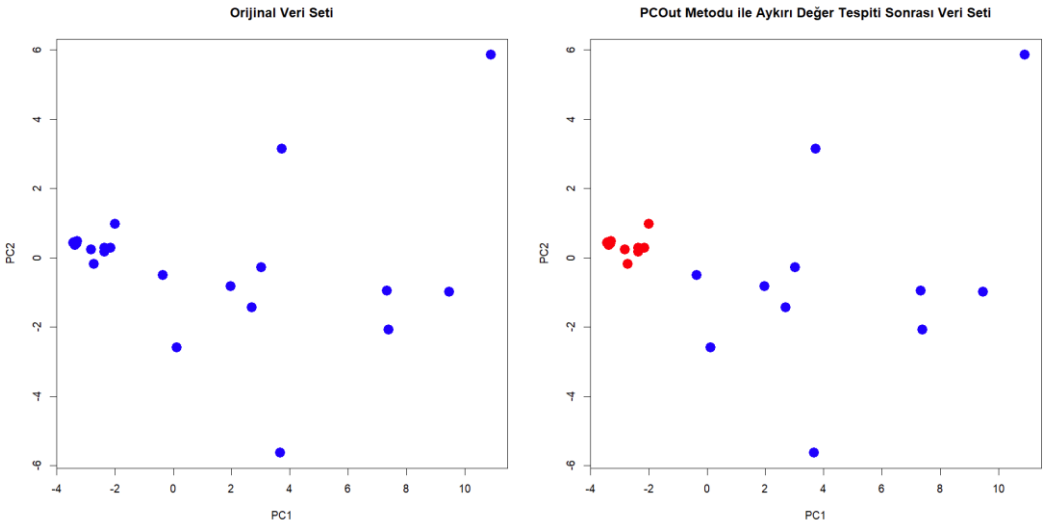
7.1.1 PCOut Yöntemi

PCOut yöntemi için, temel bileşen seçimindeki önemli bir kıstas olan, varyans açıklama oranının %99 olması gerektiği bölüm 6'da belirtilmişti. Bu kritere göre, veri seti için varyans açıklama oranlarına istinaden ilk 8 temel bileşenin alınmış ve algoritmalar, bu temel bileşenlerin oluşturduğu veri matrisi üzerinden ilerletilmiştir.

Yapılan analiz sonucunda ise; her bir gözlem için nihai ağırlıkları gösteren sayısal vektör aşağıdaki gibi bulunmuştur.

$$\omega_i = [1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$$

ω_i vektörü aykırı değer olarak işaretlenen gözlemlerin bulunduğu bir ikili vektördür (1: Aykırı Değer, 0: Aykırı Değer Değil). Yani, hangi gözlemlerin aykırı değer olarak tespit edildiğini gösterir.

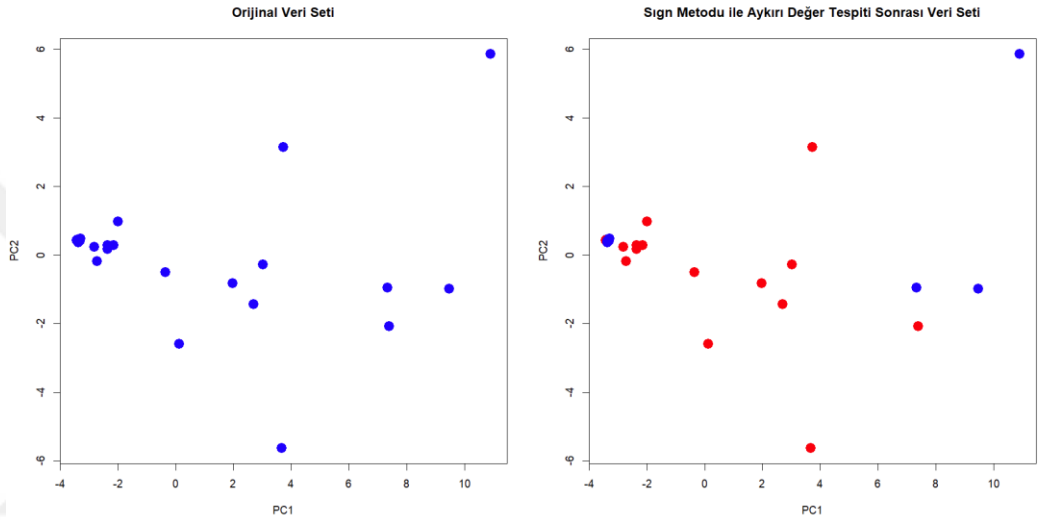


Şekil 7.1 Elements veri seti için, PCOut yönteminden önce ve sonra aykırı değer ve normal değerlerin saçılım grafiği

7.1.2 Sign Yöntemi

Yapılan analiz sonucunda; her bir gözlem için nihai ağırlıkları gösteren sayısal vektör aşağıdaki gibi bulunmuştur.

$$x_i = [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$$

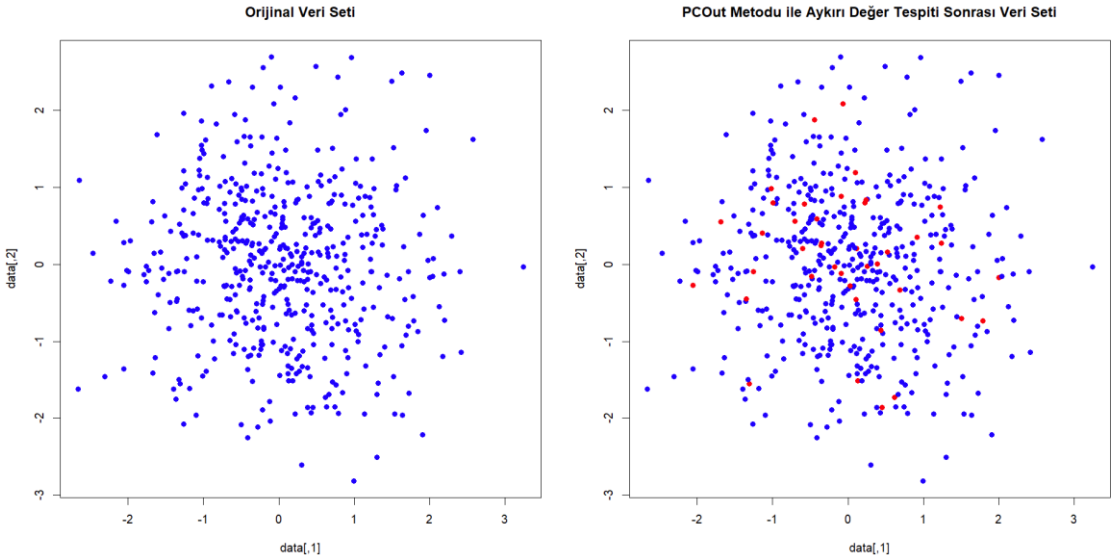


Şekil 7.2 Elements veri seti için, Sign yönteminden önce ve sonra aykırı değer ve normal değerlerin saçılım grafiği

Sign yöntemi, tüm gözlemleri aykırı değer olarak sınıflandırmış gibi görünmekle birlikte, bu durumun belirli bir eşik değeri kullanılarak aykırı değerleri tanımlamada bir esneklik eksikliği olabileceğini göstermektedir. PCOut yöntemi ise Sign yöntemine göre daha dengeli bir yaklaşım sergilemektedir. Bu, PCOut'un aykırı değerleri belirleme konusunda daha kontrollü bir yaklaşım sergileyebileceğini düşündürebilmektedir.

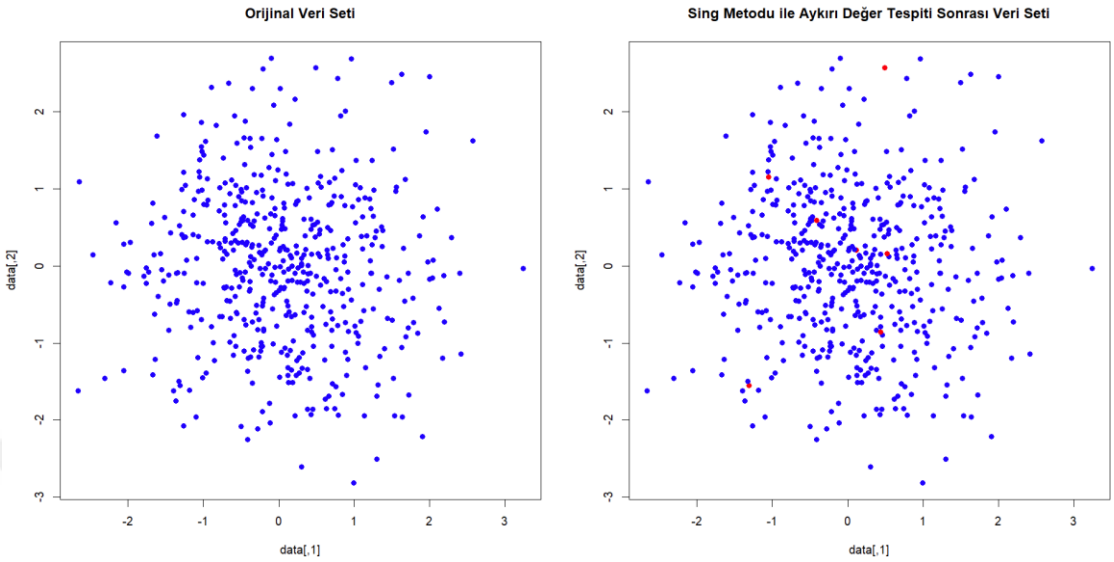
7.1.3 Üretilmiş Veri Seti Uygulaması

Bu bölümde, önceki bölümlerde anlatılmış ve gerçek veri seti üzerinde uygulaması yapılmış olan PCOut ve Sign yöntemleri için, 100 değişken ve 1000 gözlemden oluşan rasgele sayılardan üretilmiş, bu gözlemler içerisinden rasgele olarak 50 tanesi seçilmiş ve kendi değeri üzerine 5000 eklenerek bir aykırı değer haline getirilmiştir. Her iki yöntem için de ilk olarak temel bileşen analizi uygulanmış ve %99 varyans açıklama oranına sahip temel bileşenler ile uygulamalar yapılmış ve sonuçlar karşılaştırılmıştır. PCOut yöntemi kullanılarak yapılan uygulama sonucunda ω_i vektörüne ilişkin saçılım grafiği aşağıda verilmiştir.



Şekil 7.3 Üretilmiş veri seti için, PCOut yönteminden önce ve sonra aykırı değer ve normal değerlerin saçılım grafiği

Sign yöntemi kullanılarak yapılan uygulama sonucunda ω_i vektörüne ilişkin saçılım şekil 7.4'de verilmiştir.



Şekil 7.4 Üretilmiş veri seti için, Sign yönteminden önce ve sonra aykırı değer ve normal değerlerin saçılım grafiği

Her iki grafikte de mavi renkli değerler, normal değerleri kırmızı olan değerler ise, ilgili yöntem sonucunda aykırı olarak tespit edilen değerleri ifade etmektedir. İki grafik de incelendiğinde, PCOut yöntemi, aykırı değerleri tespit etme konusunda Sign yöntemine göre daha başarılı olduğu düşünülmektedir. Bu yöntemlerin performansı sonrası bölümde ele alınmıştır.

8. SİMÜLASYON ÇALIŞMASI

Bu çalışmanın temel amacı; yüksek boyutlu ve büyük veri setlerindeki aykırı değerlerin tespiti için kullanılan yöntemlerin performansını değerlendirmektir. Özellikle, bu yöntemlerin gerçek aykırı değerleri doğru bir şekilde tespit edip edemediği ve aykırı olmayan ancak yöntem sonucunda aykırı olarak tespit edilen değerlerin oranının belirlenmesi üzerine odaklanmaktadır.

Yüksek boyutlu büyük verilerde, boyut indirgeme algoritması olarak kullanılan TBA, aykırı değerlerin etkisini incelemek amacıyla gerçekleştirilmiştir. İlk aşamada, temel bileşenlerin belirlenmesi için yapılan simülasyon, rasgele şekilde ve her biri N(μ, σ^2) dağılımından olacak şekilde sırasıyla, $p = \{150,250,550\}$ ve $n = \{100,200,500\}$ için; n adet $px1$ boyutlu vektörler oluşturulmuş ve bu vektörler matris haline getirilmiştir. Aykırı değer tespiti yöntemlerinin performansını değerlendirmek amacıyla simülasyon için belirlenen veri setlerine aykırı değerler eklenmeden önce her bir gözlem birimine ait p sayıda değişken 1 ile 1000 arasında rasgele sayılardan oluşturulmuştur. Aykırı değerlerin eklenmesi için, her bir özellikte rasgele sayılar kullanılarak aykırı değerler üretilmiştir. Bu aykırı değerler, her bir gözlem birimi için belirlenen oranda (%0.1) rasgele olarak eklenmiştir. Elde edilecek temel bileşenlerdeki kısıt, %99 varyans açıklama oranı olup n ve p kombinasyonları için 100 kez simülasyon çalışması yapılmış ve temel bileşen sayılarının ortalaması alınmıştır. n ve p kombinasyonları ile elde edilen bu simülasyon sonuçlarına istinaden p =5 olarak belirlenmiştir. İkinci aşamada ise, TBA sonucu %99 varyans açıklama oranı elde edilen ilk 5 temel bileşen esas alınarak, $n = \{100,200,500\}$ olacak şekilde sayı üretilmiş, üretilen üç farklı örneklem hacmi ve p=5 için bir tasarım matrisi ve aykırı değer durumlarını oluşturmak için (0.1)n boyutunda aykırı değerler eklenmiş, $\mu = \{0,2,5,10\}$, $s^2 = \{0.1,0.5,1,2,5\}$ olarak tüm kombinasyonlar için 100 kez simülasyon çalışması yapılmış, Yüksek boyutlu ve büyük veri setlerinde aykırı değer tespiti yöntemleri kullanılarak elde edilen sonuçlar, belirli metriklerle değerlendirilmiştir.

Bu metrikler arasında, F.P (False Positive) ve F.N (False Negative) değerleri bulunmaktadır. F.P, aykırı değer olmamasına rağmen aykırı değer olarak tespit edilen değerleri, F.N ise aykırı değer olmasına rağmen aykırı değer olarak tespit edilemeyen değerlerin yüzdelerini ifade etmektedir. Bu değerler, kullanılan aykırı değer tespit yöntemlerinin doğruluk payını belirlemede kullanılmıştır. Sonuçlar, yöntemlerin performansını değerlendirmek ve kullanılabilirliğini anlamak amacıyla karşılaştırılmıştır.

Çizelge 8.1 n=100 için PCOut ve Sign yöntemleri için simülasyon sonuçları

n=100											
		s^2=0,1		$s^2 = 0,5$		$s^2 = 1$		$s^2 = 2$		$s^2 = 5$	
Yöntem	μ	%FP	%FN	%FP	%FN	%FP	%FN	%FP	%FN	%FP	%FN
PCOut	0	100	5,11	99,6	11,89	93,6	8,97	99,5	12,11	28,5	3,39
Sign	0	100	5,86	99,6	3,41	90,4	3,13	64,3	2,27	30	1,47
PCOut	2	51	8,17	47,4	6,6	55,5	6,42	47,1	3,78	27,1	3,22
Sign	2	45,5	2,02	52	2,13	47,5	2,04	44,6	1,91	30,8	1,58
PCOut	5	0	3,25	0	3,19	0	3,14	7,6	2,58	17,3	2,56
Sign	5	4	1,27	6,1	1,16	1,5	1,44	9,4	1,55	20	1,58
PCOut	10	0	3,01	0	3,22	0	3,1	0	1,98	3,7	2,42
Sign	10	0	1,06	2	1,09	1,4	0,84	1,1	1,05	5,3	1,27

Çizelge 8. 2 n=200 için PCOut ve Sign yöntemleri için simülasyon sonuçları

n=200											
		s^2=0,1		$s^2 = 0,5$		$s^2 = 1$		$s^2 = 2$		$s^2 = 5$	
Yöntem	μ	%FP	%FN	%FP	%FN	%FP	%FN	%FP	%FN	%FP	%FN
PCOut	0	100	8,68	99,8	6,58	96,5	4,99	71,6	4,32	29	3,97
Sign	0	100	7,1	99,8	6,36	91	5,1	64,5	4,3	28,9	3,69
PCOut	2	27	3,75	50,3	3,98	54,5	3,58	48	3,5	27,1	3,64
Sign	2	23,7	3,67	36,9	4,05	46,3	4,22	46,1	3,94	28,2	3,62
PCOut	5	0	3,73	0	3,26	0,1	3,49	6,6	3,69	21,1	3,17
Sign	5	1	3,42	1,5	3,21	4,8	3,07	6,9	3,19	16,5	3,65
PCOut	10	0	3,32	0	3,12	0	2,93	0	3,08	4,7	3,9
Sign	10	1	2,99	0	3,24	0	3,29	0,9	2,98	7	3,21

Çizelge 8. 3 n=500 için PCOut ve Sign yöntemleri için simülasyon sonuçları

n=500											
		s^2=0,1		$s^2 = 0,5$		$s^2 = 1$		$s^2 = 2$		$s^2 = 5$	
Yöntem	μ	%FP	%FN	%FP	%FN	%FP	%FN	%FP	%FN	%FP	%FN
PCOut	0	100	6,41	100	6,17	96,7	5,39	75,1	5,78	31	4,91
Sign	0	100	13,81	99,8	12,47	92,1	10,67	62,3	11,59	28,6	10,29
PCOut	2	28	7,02	56,4	5,83	59,9	4,53	49,4	6,03	32,2	4,86
Sign	2	20,9	10,52	39,2	10,32	44,2	10,84	44,5	10,76	26,1	10,72
PCOut	5	0	4,68	0	5,1	0,1	4,01	8	5,45	20,2	5,06
Sign	5	6	9,52	1	11,11	0,9	9,41	6,7	9,64	19,2	10,09
PCOut	10	0	4,98	0	4,46	0	4,83	0	4,79	4,2	4,03
Sign	10	0	9,35	0,1	9,49	0	9,56	0,4	9,88	4,2	9,48

8.1 Simülasyon Sonuçları

Genel eğilim, gözlem sayısı (n) arttıkça ve değişken sayısı (p) azaldıkça % F.N oranının düştüğü görülmektedir. Bu, daha fazla veriye sahip olmanın ve daha düşük boyutlu veri setlerinin, aykırı değer tespitini artırabileceğini gösterir. % F.P oranları, özellikle p'nin düşük olduğu durumlarda artma eğilimindedir. Bu durum, düşük boyutlu veri setlerinde aykırı değer tespitinin daha zor olabileceğini gösterir. Farklı n ve p kombinasyonları için % F.N ve % F.P oranlarında belirgin farklılıklar vardır. Bu, aykırı değer tespit performansının veri setinin özelliklerine ve boyutlarına bağlı olarak değişebileceğini göstermektedir.

Çizelge 8.1, 8.2 ve 8.3'ten de anlaşılacağı üzere, PCOut yöntemi, düşük μ değerleri için genellikle düşük %FP değerleri gösterirken, yüksek μ değerleri için %FP değerlerinde azalış görülmektedir. Özellikle s^2 arttıkça %FP değerleri belirgin bir şekilde azalmaktadır. Sign yöntemi ise genel olarak düşük %FP ve %FN değerlerine sahiptir, ancak yüksek s^2değerlerinde %FN değerleri azalmaktadır.

Sonuçlar, aykırı değer tespiti yöntemlerinin performansının kullanılan parametre değerlerine ve veri seti özelliklerine bağlı olduğunu göstermektedir. Yüksek s^2değerleri, genellikle %FN değerlerinde artışa ve %FP değerlerinde azalmaya yol açmaktadır. Bu nedenle, aykırı değer tespiti yöntemlerinin doğruluğunu değerlendirirken parametre değerlerini dikkate almak önemlidir.

n=100 için; PCOut yöntemi için varyans seviyesi s^2=0,1 için %FP oranı %100 iken, %FN oranı oldukça düşüktür (%5,11). Bu, modelin verileri yanlış pozitiflerle sınıflandırmakta oldukça başarılı olduğunu ancak yanlış negatiflerle sınıflandırmada daha az başarılı olduğunu gösterir. Varyans seviyesi arttıkça (%FP oranı düşerken) %FN oranı da genellikle artar. Bu, modelin düşük varyanslı verilere daha iyi uyum sağladığını ancak yüksek varyanslı verilere karşı daha az performans gösterdiğini gösterebilir.

n=200 için; PCOut ve Sing yöntemleri, n=100 durumuna benzer şekilde davranır. Ancak, daha büyük bir gözlem sayısıyla birlikte %FP ve %FN değerlerinde daha stabil bir performans elde edildiği görülmektedir. Artan μ değerleri, PCOut yönteminde %FP değerlerinde azalışa neden olurken, Sing yöntemi daha etkili bir performans sergiler.

n=500 için; Genel olarak, her iki yöntem de standart sapmanın artmasıyla birlikte %FP ve %FN değerlerinde bir azalma eğilimi göstermektedir. Ancak, standart sapma s^2=0 için %FP ve %FN değerleri yine de oldukça yüksektir. Sing yöntemi genellikle daha yüksek %FP ve %FN değerlerine sahiptir. Özellikle standart sapma s^2=0.1, s^2=0.5 ve s^2=1 için bu yöntemin performansı zayıftır. Standart sapma s^2=5 için PCOut yöntemi, %FP ve %FN değerlerinde düşük bir performans sergilerken, Sing yöntemi hala yüksek %FP ve %FN değerlerine sahiptir. Standart sapma s^2=0 için, her iki yöntem

de %FP değerleri çok yüksektir, bu da aykırı değerlerin yanlışlıkla tespit edilme olasılığının yüksek olduğunu gösterir. Her iki yöntemin de standart sapma s^2=2 için %FP ve %FN değerleri orta seviyededir, bu da bir denge sağladığını gösterir

Elde edilen üç tablo, farklı n (gözlem sayısı) değerleri için, çeşitli ortalama ve varyans kombinasyonları altında yapılan aykırı değer tespit yöntemlerinin performansını göstermektedir. Gözlem sayısı arttıkça (n=100'den n=500'e), her iki yöntemin performansında genel bir iyileşme gözlemlenmiştir. Daha büyük veri setleriyle birlikte yöntemlerin daha stabil bir performans sergilediği görülmektedir. Düşük μ ve s^2 değerleri, her iki yöntemin de daha iyi performans gösterdiği durumları temsil etmektedir. Bu durumlar, aykırı değerlerin daha belirgin olduğu durumları ifade edebilir. Artan μ ve s^2 değerleriyle birlikte, özellikle PCOut yönteminin %FP değerlerinde belirgin bir azalış gösterdiği gözlemlenmiştir.

100, 200 ve 500 gözlem sayısı için yapılan değerlendirmelerde, genel eğilimler gözlemlenebilir. Veri seti büyüdükçe, aykırı değer tespit yöntemlerinin performansının genellikle arttığı görülmektedir. Daha büyük veri setleri, algoritmaların genelleyebilme yeteneklerini artırabilir ve bu da doğruluk oranlarını iyileştirebilir. Büyük bir örnek büyüklüğüne sahip durumlar (n=500), genellikle daha düşük FP ve FN değerlerine sahiptir. Bu durum, daha büyük veri setlerinde aykırı değer tespit yöntemlerinin genellikle daha iyi performans gösterme eğiliminde olduğunu gösterir.

Standart sapmanın artması genellikle %FP ve %FN değerlerinde bir azalmaya neden olmaktadır. Yani, daha geniş bir veri dağılımına sahip olmak, aykırı değer tespit yöntemlerinin daha iyi performans göstermesine katkıda bulunabilir. Ancak, düşük standart sapma değerleri (örneğin, s^2=0.1) için %FP ve %FN değerleri genellikle yüksektir, bu da bu durumda aykırı değer tespit yöntemlerinin zayıf olduğunu gösterir. Genel olarak, standart sapmanın artması, aykırı değer tespit yöntemlerinin performansını olumlu yönde etkilemektedir. Ancak, düşük standart sapmalarda her iki yöntem de zayıf performans sergileme eğilimindedir.

Genel olarak, Sing yöntemi, daha düşük %FP ve %FN değerlerine sahiptir, özellikle standart sapma değerleri arttıkça bu eğilim devam etmektedir. PCOut yöntemi, standart sapma değeri arttıkça performansını iyileştirmekte, ancak Sing yöntemine kıyasla genellikle daha yüksek hatalara sahiptir. Standart sapma s^2=0 için her iki yöntemde de %FP değerleri çok yüksektir, bu durumda aykırı değerleri doğru bir şekilde tespit etme konusunda zorluk yaşandığı anlamına gelir. s^2=5 için ise PCOut yöntemi genellikle daha iyi performans gösterirken, Sing yöntemi hala yüksek hatalara sahiptir.

Her üç durumda da, standart sapmanın artması genellikle FP ve FN değerlerinde bir azalmaya neden olmuştur. Ancak, özellikle standart sapma düşükken (s^2=0.1), tüm yöntemlerde yüksek FP ve FN değerleri görülmektedir. Sing yöntemi, özellikle daha düşük standart sapmalarda daha yüksek FP ve FN değerlerine sahiptir.

PCOut yöntemi, özellikle standart sapma arttıkça, daha düşük FP ve FN değerlerine sahip olma eğilimindedir. Ancak, bu yöntemin de düşük standart sapmalarda performansının düşük olduğu görülmektedir. Sing yöntemi, genellikle daha yüksek FP ve FN değerlerine sahiptir, ancak bazı durumlarda PCOut yöntemine göre daha iyi performans gösterdiği görülmektedir.

9. TARTIŞMA VE SONUÇ

Bu çalışma, yüksek boyutlu veri analizi ve boyut indirgeme tekniklerine odaklanarak, özellikle PCOut ve Sign yöntemleri üzerinde detaylı bir inceleme sunmaktadır. Yüksek boyutlu veri setlerinin yönetimi zorluklar doğurabilir ve bu nedenle etkili boyut indirgeme teknikleri, veri setlerindeki karmaşıklığı anlamak ve ele almak için önemlidir.

Tez kapsamında incelenen PCOut yöntemi, temel bileşenlere ve ağırlık belirlemeye odaklanır. PCOut yöntemi, temel bileşenlere ve ağırlık belirlemeye odaklanır. Bu yöntem, ağırlıkları lokasyon ve saçılım aykırı değerleri için ayrı ayrı belirler ve ardından birleştirir. PCOut yöntemi, kritik değerleri kantil bazında belirler, bu sayede aykırı değerleri tespit eder. Ayrıca, çok değişkenli aykırı değerleri belirlemede Mahalanobis mesafelerini kullanır ve dayanıklı bir yöntemdir. Yani, veri setindeki aykırı gözlemlere karşı dirençlidir. Bir diğer yöntem olan Sign yöntemi ise; her bir gözlem için tek bir ağırlık belirler. Bu yöntem, ağırlıkları belirlerken Mahalanobis mesafesini kullanır ve bu sayede çok değişkenli aykırı değerleri tespit eder. PCOut yönteminden farklı olarak, Sign yöntemi kritik değerleri ki-kare dağılımından elde eder. Ayrıca, bu yöntem de dayanıklı bir yöntemdir ve veri setindeki aykırı gözlemlere karşı dirençlidir.

PCOut yöntemi, genelde aykırı değer tespitinde etkili olmakla birlikte, küçük veri setlerinde özellikle yanlış pozitif sonuçlar (FP) yüksek oranda görülmüştür. Bu yöntem, büyük veri setlerinde daha stabil bir performans sergilemiştir, ancak küçük veri setlerinde dikkatli kullanılmalıdır.

Sign yöntemi, genelde aykırı değer tespitinde etkili olsa da, özellikle aykırı olmayanları belirleme konusunda daha zorlanmıştır. Büyük veri setlerinde yanlış pozitif sonuçlar (FP) belirgin bir şekilde artmıştır. Bu yöntem, özellikle büyük veri setlerinde daha dikkatli bir şekilde değerlendirilmelidir.

Her iki yöntem de yüksek boyutlu veri setlerinde güçlü aykırı değer tespit yöntemleri olarak öne çıkmaktadır. Ancak, tercih edilecek yöntem, veri setinin özelliklerine bağlı olarak değişiklik gösterebilir. Yüksek boyutlu ve büyük veri setlerinde aykırı değer tespiti yaparken, hassasiyet kaybı ve yanlış pozitif sonuçlar dikkate alınmalıdır. Bu yöntemlerin seçimi, veri setinin büyüklüğü, dağılımı ve analiz gereksinimlerine göre yapılmalıdır. Hem PCOut hem de Sign yöntemleri, büyük veri setlerinde daha iyi performans göstermiştir. Her iki yöntem de % F.P değerlerinde belirgin bir azalış göstermiştir, bu da yanlış pozitif sonuçların yüksek olduğunu ve bu yöntemlerin bazı durumlarda hassasiyet kaybına neden olabileceğini gösterir.

Her iki yöntem de aykırı değer tespiti konusunda güçlüdür, ancak büyük veri setlerinde daha iyi performans gösterirler. Ancak, hassasiyet kaybına ve yanlış pozitif sonuçlara dikkat edilmelidir, özellikle de veri seti büyüdükçe. Her iki yöntem de seçilen metriklerde belirgin bir performans artışı sağlamıştır, ancak özellikle Sign yöntemi, aykırı olmayanları belirlemede zorlandığı durumlar bulunmaktadır.

Sonuç olarak, yüksek boyutlu veri analizi ve boyut indirgeme yöntemleri, veri setinin özelliklerine, amacına ve analiz kriterlerine bağlı olarak farklı performans sergileyebilir. PCOut ve Sign yöntemleri, aykırı değer tespiti konusunda güçlü olmalarına rağmen, büyük veri setlerinde daha iyi performans gösterdikleri ve seçilen metriklerde artış sağladıkları belirlenmiştir. Ancak, her iki yöntem de özellikle de veri seti büyüdükçe hassasiyet kaybı ve yanlış pozitif sonuçlar gibi zorluklarla çıkarabilmektedir.

Bu nedenle, bu tür analiz ve yöntemlerin uygulanacağı bağlamda dikkatlice değerlendirilmesi gerekmektedir. Veri setinin büyüklüğü, dağılımı ve analiz amaçları göz önüne alınarak, PCOut ve Sign gibi yöntemler arasında tercih yapılırken dikkatli bir seçim yapılmalıdır. Ayrıca, bu yöntemlerin performansını değerlendirmede kullanılan metriklerin yanı sıra, analizin duyarlılık ve özgüllük gibi kriterlere de odaklanılmalıdır. Sonuç olarak, her bir yöntemin avantajları ve zorlukları dikkate alındığında, doğru seçimin spesifik veri analiz ihtiyaçlarına uygun olarak yapılması önemlidir.

KAYNAKLAR

Acuna, E., & Rodriguez, C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. *Classification, Clustering, and Data Mining Applications* (s. 639-647). Berlin: Springer.

Aggarwal, C. (2017). *High-Dimensional Outlier Detection: The Subspace Method.* Springer. doi:https://doi.org/10.1007/978-3-319-47578-3_5

Aggarwal, C. C. (2016). C. C. Aggarwal içinde, *Outlier Analysis* (s. 8). New York: Springer. doi: 10.1007/978-3-319-47578-3

Aguinis, H., Gottfredson, R., & Joo, H. (2013, January 14). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods, 16*(2), s. 270-301.

Alpar, P. (2013). *Kommerzielle Nutzung des Internet: Unterstützung von Marketing, Produktion, Logistik und Querschnittsfunktionen durch Internet, Intranet und kommerzielle Online-Dienste.* Springer-Verlag.

Alpaydın, E. (2010). *Introduction to machine learning.* Cambrige: MIT Press.

Badaoui, F., Amar, A., Hassou, L. A., Zoglat, A., & Okou, C. G. (2017). Dimensionality reduction and class prediction algorithm with application to microarray Big Data.

Barnett, V., & Lewis, T. (1978). *Outliers in Statistical Data.* New York: Wiley.

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data, 3rd Edition.* Wiley.

Bellman, R. (2013). *Dynamic Programming.* Chelmsford: Courier Corporation.

Ben-Gal, I. (2005). *Data Mining and Knowledge Discovery Handbook.* Kluwer Academic Publisher.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. J. Kogan, C. Nicholas , & M. Teboulle içinde, *Grouping Multidimensional Data* (s. 25-71). Beltimore: Springer.

Berksoy, M. (2019, Ağustos 25). *Veri ve Bilgi Nedir? Arasındaki Fark Nedir?* Tekno Tower: https://teknotower.com/veri-ve-bilginin-onemi-1/ adresinden alındı

Bickel, D., Montazeri, Z., Hsieh, P.-C., Beatty, M., Lawit, S., & Bate, N. (2009). Gene network reconstruction from transcriptional dynamics under kinetic model uncertainty: a case for the second derivative. *Bioinformatics, 25*(6), s. 772-779.

Bolstad, B., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*(19), 185-193. https://doi.org/10.1093/bioinformatics/19.2.185

Büyüköztürk, Ş. (2023). Faktör Analizi. *Sosyal Bilimler için Veri Analizi El Kitabı* (s. 110-117). içinde Ankara: Pegem Akademi .

Cabrera, J., & Amaratunga, D. (2001). Analysis of Data from Viral DNA Microchips. *Journal of the American Statistical Association*, 1161-1170. doi:10.1198/016214501753381814

Chandola, V., Banerjee, A., & Kumar, V. (2009, July 30). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), s. 1-58.

Chen, H., Chiang, R. H., & Storey, V. C. (2012, December). Business Intelligence And Analytics: From Big Data to Big Impact. *Mis Quarterly*.

Chiarenzelli, J., Lock, R., Cady, C., Bregani, A., & Whitney, B. (2012). Variation in river multi-element chemistry related to bedrock buffering: an example from the Adirondack region of northern. *Environmental Earth Sciences, 67*(1), s. 189-204. doi:10.1007/s12665-011-1492-z

Chris, S. (1997, January 24). Basic Concepts and Procedures of Confirmatory Factor Analysis. *Southwest Educational Research Association*, s. 2-15.

Çilli, M. (2007). İnsan Hareketlerinin Modellenmesi ve Benzeşiminde Temel Bileşenler Analizi Yönteminin Kullanılması. Ankara.

Datta, P., & Kibler, D. (1995). Learning Prototypical Concept Descriptions. *Proceedings of the Twelfth International Conference on Machine Learning*, (s. 158-166). California.

Donoho, D. (2000). *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*. California: Stanford University.

Donoho, D., & Tanner, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences* , 367.

Ebeto, C. (2017). Biometrics & Biostatistics International Journal. *Sampling and Sampling Methods*, 5-7. doi:10.145067bbij2017.05.00149

Edward, F. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *International Biometric Society*, 761-777.

Fan, J., Han, F., & Liu, H. (2014, February). Challenges of Big Data analysis. *National Science Review*, s. 293-314.

Fawcett, T., & Provost, F. (1997, September). Adaptive Fraud Detection. *Data Mining and Knowledge Discovery* , s. 291-316.

Filiz, Z. (2003). Güvenilirlik Çözümlemesi, Temel Bileşenler ve Faktör Çözümlemesi. *Anadolu Üniversitesi Bilim ve Teknoloji Degisi*, 211-222.

Filzmoser, P., & Gschwandtner, M. (2021, 07 30). mvoutlier: Multivariate Outlier Detection Based on Robust Methods. *Fast algorithm for identifying multivariate outliers in high-dimensional and/or large datasets*. https://cran.r-project.org/web/packages/mvoutlier/index.html adresinden alındı

Filzmoser, P., Maronna, R., & Werner, M. (2007, January). Outlier identification in high dimension. *Computational Statistics & Data Analysis*, s. 1694-1711.

Firican, G. (2017, February). The 10 Vs of Big Data.

Firican, G. (2021, 03 1). *The history of big data*. Lights On Data: https://www.lightsondata.com/the-history-of-big-data/ adresinden alındı

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods.* SAGE Publications.

Gadepally, V., & Kepner, J. (2014). Big Data Dimensional Analysis. *HPEC* (s. 1-6). Massachusetts: IEEE.

Gantz, J., & Reinsel, D. (2012). The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East.

Gorban, A., Golubkov, A., Bogdan, G., Mirkes, E., & Tyukin, I. (2018). Correction of AI systems by linear discriminants: Probabilistic foundations. *Information Sciences*, 302-322.

Gürsakal, N. (2001). *Bilgisayar Uygulamalı İstatistik 1.* Ankara: Alfa Yayınları.

Güzeller, C. O., Eser, M. T., & Aksu, G. (2017). Faktör Analizi. *Açımlayıcı ve Doğrulayıcı FAKTÖR ANALİZİ ile Yapısal Eşitlik Modeli Uygulamaları* (s. 1-24). içinde Ankara: Detay Yayıncılık.

Hawkins, D. (1980). *Identification of Outliers.* London: Chapman and Hall.

Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review, 22*, s. 85-126.

Jin, X., Wah, B., Cheng, X., & Wang, Y. (2015). Significance and Challenges of Big Data Research. *Big Data Research*.

Johnson, R., & Wichern, D. (1998). *Applied Multivariate Statistical Analysis.* Prentice Hall.

Johnson, R., & Wichern, D. (2002). *Applied Multivariate Statistical Analysis.* Prentice Hall.

Jolliffe, I. (1972). Discarding Variables in a Principal Component Analysis. I: Artificial Data. *Journal of the Royal Statistical Society*, s. 160-173. doi:10.2307/2346488

Kaiser, H. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, s. 141-151. doi:10.1177/001316446002000116

Kaisler, S., Armour, F., Espinosa, J., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. *46th Hawaii International Conference on System Sciences.* Wailea. doi:10.1109/HICSS.2013.645

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big Data: Issues, Challanges, Tools and Good Practices. *Sixth International Conference on Contemporary Computing (IC3).* Noida, India: IEEE. doi:10.1109/IC3.2013.6612229

Kay, P., & Jolliffe, I. (2001, September). A Comparison of Multivariate Outlier Detection Methods for Clinical Laboratory Safety Data. *Journal of the Royal Statistical Society. Series D (The Statistician), 50*(3), s. 295-308.

Kılıç, İ., & Ural, A. (2005). Evren. *Bilimsel Araştırma Süreci ve SPSS ile Veri Analizi* (s. 29). içinde Ankara: Detay Yayıncılık.

Kline, P. (1994). *An Easy Guide to Factor Analysis.* Psychology Press.

Lederer, J. (2022). *Fundamentals of High-Dimensional Statistics: With Exercises and R Labs.* Springer. doi:10.1007/978-3-030-73792-4

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013, July). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), s. 764-766.

Linear Discriminant Analaysis. (2019). B. Everitt, & G. A. Dunn içinde, *Applied Multivariate Data Analysis.* New York: Oxford Üniversity Press.

Liu, H., Parsons, L., & Haque, E. (2004, June 1). Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter, 6*(1), s. 90-105.

Luxburg, U. V. (2007). A Tutorial on Spectral Clustering. U. V. Luxburg içinde, *Statistics and Computing* (s. 395-416). Tübingen: Springer.

Mahalanobis. (1936). On the Generalised Distance in Statistics. *Sankhya A*, 49-55. https://doi.org/10.1007/s13171-019-00164-5 adresinden alındı

Manyika, J., Chui, M., McKinsey Global Institute, Brown, B., Bughin, J., Dobbs, R., . . . Byers, A. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity.* McKinsey.

Maronna, R., & Zamar, R. (2002, November). Robust Estimates of Location and Dispersion for High-Dimensional Datasets. *Technometrics, 44*(4), s. 307-317.

Maronna, R., Martin, D., & Yohai, V. (2006). *Robust Statistics: Theory and Methods.* Wiley.

Morgan, G. A. (2001, September). Data Collection Techniques. *Article in Journal of the American Academy of Child and Adolescent Psychiatry*, s. 973-976. doi:10.1097/00004583-200108000-00020

Nielsen, F. (2016). Hierarchical Clustering. F. Nielsen içinde, *Introduction to HPC with MPI for Data Science* (s. 195-211). Switzerland : Springer.

Ohlhorst, F. (2013). *Big Data Analytics: Turning Big Data into Big Money.* New Jersey: John Wiley.

Olkin, K. M. (1974). An index of factorial simplicty. *Psychometrika*, 31-36. doi:10.1007/BF02291575

Owais, S. S., & Hussein, N. S. (2017). Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data. *International Journal of Advanced Computer Science and Applications*.

Özdamar, K. (1999). Lineer Diskriminant Analizi. *Paket Programlarla İstatistiksel Veri Analizi 2.* içinde Eskişehir: Kaan Kitapevi.

Öztürk, F. (2006, Haziran 23). *Veri.* Fikri Öztürk Web Sayfası: http://80.251.40.59/science.ankara.edu.tr/ozturk/Dersler/ist101/Ders3/Veri.pdf adresinden alındı

Pena, D., & Prieto, F. (2001). Cluster Identification Using Projections. *Journal of the American Statistical Association, 96*(456), s. 1433-1445.

Prakasa Rao, B. (2014). C. R. Rao: A Life in Statistics.

R Core Team 2013, Version: 2023.12.1+402

Rocke, D. (1996, June). Robustness Properties of S-Estimators of Multivariate Location and Shape in High Dimension. *The Annals of Statistics, 24*(3), s. 1327-1345.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 1273-1283.

Ruts, I., & Rousseeuw, P. (1996, November 15). Computing depth contours of bivariate point clouds. *Computational Statistics & Data Analysis, 23*(1), s. 153-168.

Ryan, T. (1997). *Modern Regression Methods.* Wiley-Interscience.

Sadik, S., & Gruenwald, L. (2014, March 17). Research issues in outlier detection for data streams. *ACM SIGKDD Explorations Newsletter, 15*(1), s. 33-40.

Sarfo, J. O., Debrah, P. T., Gbordzoe, N. I., & Obeng, P. (2022). Sampling methods. *European Researcher. Series A*, 55-63. doi:10.13187/er.2022.2.55

Schroeck, M., Shockley, R., Smart, J., Morales, D., & Tufano, P. (2012). *Analytics: The real-world use of big data: How innovative enterprises extract value from uncertain data.* New York: IBM.

Schroeck, M., Shockley, R., Smart, J., Morales, D., & Tufano, P. (2013). Analytics: El Uso De Big Data En El Mundo Real. *IBM.* içinde IBM Global Business Services.

Shin, K., Hooi, B., Kim, J., & Faloutsos, C. (2017, August 04). DenseAlert: Incremental Dense-Subtensor Detection in Tensor Streams. *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (s. 1057-1066). doi:https://doi.org/10.1145/3097983.3098087

Shin, K., Hooi, B., Kim, J., & Faloutsos, C. (2017, June 11). DenseAlert: Incremental Dense-Subtensor Detection in Tensor Streams. *Proceedings of the Tenth ACM international conference on web search and data mining*, (s. 761-770).

Stain, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Berkeley Symposium on Mathematical Statistics and Probability*, 197-206.

Sütçüler, E. (2006). Gerçek Zamanlı Video Görüntülerinden Yüz Bulma ve Tanıma Sistemi. *Yüksek Lisans Tezi*. İstanbul: Yıldız Teknik Üniversitesi Fen Bilimleri Fakültesi.

Taşdöğen, C. (2019, Aralık 19). *IIENSTITU*. Nitel ve Nicel Veri Toplama Teknikleri Nelerdir?: https://www.iienstitu.com/blog/nitel-ve-nicel-veri-toplama-teknikleri-nelerdir adresinden alındı

Taylor, L., Cowls, J., Schroeder, R., & Meyer, E. (2014, December). Big Data and Positive Change in the Developing World.

Thudumu, S., Branch, P., Jin, J., & Singh, J. (2019). Elicitation of Candidate Subspaces in High-Dimensional Data. *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, (s. 1995-2000). Zhangjiajie, China.

Thudumu, S., Branch, P., Jin, J., & Singh, J. (2020). Estimation of Locally Relevant Subspace in High-dimensional Data. *ACSW '20: Proceedings of the Australasian Computer Science*

Week Multiconference (s. 1-6). Melbourne: ACSW. doi:https://doi.org/10.1145/3373017.3373032

Ural, A., & Kılıç, İ. (2011). *Bilimsel Araştırma Sürevi ve SPSS ile Veri Analizi.* Ankara: Detay Yayıncılık.

Van Der Maaten, L., Postma, E., & Van Der Herik, J. (2009). Dimensionality reduction: a comparative review. *J Mach Learn Res*(10), s. 66-71.

Vaus, D. D. (1990). *Surveys in Social Research.* California: Unwin Hyman.

Wainwright, M. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge University. doi:10.1017/9781108627771

Warnick, K., & Karny, M. (1997). Curse of Dimensionality. *Computer Intensive Methods in Control and Signal* (s. 282-294). içinde Berlin: Birkhauser.

Wichern, D., & Johnson, R. (2007). *Applied Multivariate Statistical Analysis.* New Jersey, United States of America: Pearson.

Williams, G., Baxter, R., He, H., Hawkins, S., & Gu, L. (2002). A comparative study of RNN for outlier detection in data mining. *IEEE International Conference on Data Mining (ICDM).* IEEE.

Winsor, C. P., Tukey, J. W., Mosteller, F., & Hasting Jr, C. (1947). Low Moments for Small Samples: A Comparative Study of Order Statistics. *Ann. Math. Statist.*, 413-426. doi:10.1214/aoms/1177730388

Yakar, U. (2022, Temmuz 15). *Petrolden Bile Daha Değerli Olan Veri Nedir, Gelecekte Bizi Veri Savaşları mı Bekliyor?* Webtekno: https://www.webtekno.com/veri-nedir-neden-onemli-h117453.html adresinden alındı

Yamane, T. (2001). Giriş. *Temel Örnekleme Yöntemleri* (s. 6). içinde İstanbul: Literatür Yayınları.

Zhang, L., Lin, J., & Karim, R. (2017). Sliding window-based fault detection from high-dimensional data streams. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 47*(2), s. 289-303.

Zimek, A., Schubert, E., & Kriegel, H.-P. (2012). *A survey on unsupervised outlier detection in high-dimensional numerical data* (Cilt 5). Wiley. doi:https://doi.org/10.1002/sam.11161