
GAZİ ÜNİVERSİTESİ
Gazi
1926

SAĞLAMLIK ÖZELLİĞİNE DAYALI TAHMİN EDİCİLERLE DİSKRİMİNANT ANALİZİ

Gamze ŞAHİN

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANABİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ARALIK 2017

Gamze ŞAHİN tarafından hazırlanan "SAĞLAMLIK ÖZELLİĞİNE DAYALI TAHMİN EDİCİLERLE DİSKRİMİNANT ANALİZİ" adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi İstatistik Anabilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Yrd. Doç. Dr. Necla GÜNDÜZ

İstatistik Anabilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ...……

Başkan: Prof. Dr. Yılmaz AKDİ

İstatistik Anabilim Dalı, Ankara Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Üye: Yrd. Doç. Dr. Celal AYDIN

İstatistik Anabilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Tez Savunma Tarihi: 15/12/2017

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

……..…….

Prof. Dr. Sena YAŞYERLİ

Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Gamze ŞAHİN

15/12/2017

SAĞLAMLIK ÖZELLİĞİNE DAYALI TAHMİN EDİCİLERLE DİSKRİMİNANT ANALİZİ

(Yüksek Lisans Tezi)

Gamze ŞAHİN

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Aralık 2017

ÖZET

Gerçek hayatta karşılaştığımız veriler her zaman sorunsuz, ideal veri setleri olmayabilirler. Bu veri setlerinde, verinin çoğunluğunun yer aldığı bölgeden uzakta (uç noktalarda veya verinin genel görünümünden farklı davranışta) gözlemler yer alabilmektedir. Bu tarz gözlemler başka dağılımdan karışmış gözlemler olabilmektedir. Başka dağılımdan karışmış gözlemler, karışan gözlem olarak nitelendirilmektedir. Veri setinde yer alan başka dağılımdan karışmış gözlemler güvenilir olmayan analiz sonuçlarına yol açmaktadır. Böyle bir durumda başka dağılımdan karışmış gözlemlerin tahmin ediciler üzerindeki etkisini azaltarak güvenilir sonuçlar elde etmek için sağlam tahmin yöntemlerine başvurulabilir. Bu çalışmada çok değişkenli problemlerde başka dağılımdan karışmış gözlemlerin varlığının diskriminant analizi ve temel bileşenler analizi üzerindeki etkisi incelenerek sağlamlık özelliğine dayalı hesaplanan kovaryans matrisinin determinantı en küçük olan matrisin belirlenmesi ve izdüşüm arama yöntemleri üzerinde durulmaktadır.

Bilim Kodu : 20513
Anahtar Kelimeler : Diskriminant analizi, karışan gözlem, sağlamlık, en küçük kovaryans determinantı, izdüşüm arama
Sayfa Adedi : 85
Danışman : Yrd. Doç. Dr. Necla GÜNDÜZ

DISCRIMINANT METHOD WITH ROBUST BASE ESTIMATORS

(M. Sc. Thesis)

Gamze ŞAHİN

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

December 2017

ABSTRACT

In real life, data sets we encounter may not always be unproblematic and ideal data sets. The data sets include observations which can appear in far from the area where the majority of the data takes place. Such observations are characterized as contaminated observations from different distributions. Contaminated observations from different distributions can be described as contaminated observations. Contaminated observations from different distributions are in the data set cause untrusting analysis results. In such a case, the methods of estimating robustness can be applied to achieve reliable results by reducing the effect of contaminated observations from different distributions on predictors. In this thesis, by examining the effect of contaminated observations from different distributions on discriminant analysis and principal component analysis on multivariate problems, the calculated covariance matrix based on the robustness property which has the smallest determinant and projection pursuit method are emphasized.

Science Code : 20513
Key Words : Discriminant analysis, contaminated observations, robustness, minimum covariance determinant, projection pursuit
Page Number : 85
Supervisor : Assist. Prof. Dr. Necla GÜNDÜZ

TEŞEKKÜR

Tez çalışmam sırasında kıymetli bilgi, birikim ve tecrübeleri ile bana yol gösterici ve destek olan saygıdeğer danışman hocam; Yrd. Doç. Dr. Necla GÜNDÜZ'e, çalışmam süresince tüm zorlukları benimle göğüsleyen, hayatımın her evresinde bana destek olan, yaşadığı tüm zorluklara rağmen beni en iyi şekilde yetiştiren kıymetli annem Şükran ŞEKER'e ve tüm aileme ve çalışmam boyunca benden yardımlarını esirgemeyen arkadaşım Yusuf ATA'ya sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	Açıklamalar
$\mathbf{Al}$	Alüminyum
$\boldsymbol{\alpha}$	Ayırma gücü en iyi olan doğrultu
$\mathbf{A_{p*p}}$	p*p boyutlu genel kareler toplamı matrisi
$\mathbf{b_j^2}$	Bir değişkenin temel bileşenlerce açıklanmayan kısmı
$\mathbf{B_{p*p}}$	p*p boyutlu gruplar arası varyans-kovaryans matrisi
$\mathbf{C}$	Santigrat
$\mathbf{Ca}$	Kalsiyum
$\mathbf{d}$	Sıralı uzaklık
$\mathbf{d_i^F(x)}$	i. sınıfa ilişkin Fisher diskriminant fonksiyonu
$\mathbf{d_i^L(x)}$	i. sınıfa ilişkin doğrusal diskriminant fonksiyonu
$\mathbf{d_i^Q(x)}$	i. sınıfa ilişkin karesel diskriminant fonksiyonu
$\boldsymbol{\varepsilon}$	Karışma oranı
$\mathbf{F}$	F test istatistiği
$\mathbf{Fe}$	Demir
$\mathbf{F_0(x)}$	Dağılım fonksiyonu
$\mathbf{F_\varepsilon(x)}$	Karışma dağılım fonksiyonu
$\mathbf{F_k(x)}$	Karışan örnek dağılım fonksiyonu
$\mathbf{f(x)}$	Olasılık yoğunluk fonksiyonu
$\widehat{\boldsymbol{\gamma}}_{\mathbf{1}p}$	Mardia çarpıklık katsayısı
$\widehat{\boldsymbol{\gamma}}_{\mathbf{2}p}$	Mardia basıklık katsayısı
$\mathbf{h}$	MCD algoritması için seçilecek gözlem sayısı
$\mathbf{H_0}$	Yokluk hipotezi
$\mathbf{H_1}$	Alternatif hipotez
$\mathbf{h_j^2}$	Tüm temel bileşenlerin bir değişkeni açıklama oranı
$\mathbf{I_{p*p}}$	p*p boyutlu birim matris

Simgeler	Açıklamalar
$\mathbf{k}$	Önemli temel bileşen sayısı
$\boldsymbol{\kappa}$	Ölçek büyütme faktörü
$\boldsymbol{\chi^2}$	Ki-kare test istatistiği
$\mathbf{l}$	Sınıf sayısı
$\mathbf{ln}$	Logaritma fonksiyonu
$\mathbf{L_1}$	L_1 medyan istatistiği
$\mathbf{L(x)}$	Olabilirlik fonksiyonu
$\boldsymbol{\lambda}$	Özdeğer
$\mathbf{m}$	Karışan gözlem sayısı
$\mathbf{MC^{-1}}$	Box-M test istatistiği
$\mathbf{MD(x)}$	x gözleminin mahalanobis uzaklığı
$\mathbf{Mg}$	Magnezyum
$\mathbf{m_i^2}$	i. gözlemin merkeze uzaklığı
$\boldsymbol{\mu}$	Yığına ilişkin ortalama vektörü
$\hat{\boldsymbol{\mu}}$	Konum parametresi tahmini
$\mathbf{n}$	Gözlem sayısı
$\mathbf{p}$	Değişken sayısı
$\mathbf{p_i}$	i. sınıfa atanma önsel olasılığı
$\mathbf{P(\pi_i)}$	i. sınıfa atanma sonsal olasılığı
$\boldsymbol{\pi_i}$	Yığındaki i. sınıf
$\mathbf{Q}$	Konum kaydırma ölçüsü
$\boldsymbol{\rho}$	İlişki katsayısı
$\mathbf{R_{p*p}}$	p*p boyutlu korelasyon matrisi
$\mathbf{RD(x)}$	x gözleminin sağlam uzaklığı
$\boldsymbol{\sigma_{ij}}$	i. ve j. değişkenler arasındaki yığına ilişkin korelasyon
$\mathbf{s}$	Permütasyon
$\mathbf{S}$	Örnek varyans-kovaryans matrisi
$\mathbf{S_i}$	i. sınıf örnek varyans-kovaryans matrisi
$\mathbf{Si}$	Silisyum
$\boldsymbol{\Sigma}$	Yığına ilişkin varyans-kovaryans matrisi
$\hat{\boldsymbol{\Sigma}}$	Ölçek parametresi tahmini

Simgeler	Açıklamalar
$\mathbf{S_n}$	İzdüşüm endeksi
$\mathbf{S_{ortak}}$	Ortak varyans-kovaryans matrisi
$\mathbf{t}$	Özvektör
$\mathbf{T}$	Tahmin edici
$\boldsymbol{\theta}$	Herhangi bir parametre
$\boldsymbol{\hat{\theta}}$	Herhangi bir parametre tahmini
$\mathbf{T(F)}$	Fonksiyonel
$\mathbf{Ti}$	Titanyum
$\mathbf{V}$	Asıl temel bileşenler matrisi
$\boldsymbol{\upsilon}$	Konum değiştirme parametresi
$\mathbf{W}$	Yükler matrisi
$\boldsymbol{\Lambda}$	Wilks'in Lambda istatistiği
$\mathbf{w_i}$	i. gözlem ağırlığı
$\mathbf{W_{p*p}}$	p*p boyutlu grupiçi varyans-kovaryans matrisi
$\bar{\mathbf{x}}$	Örnek ortalama vektörü
$\bar{\mathbf{x}}_\mathbf{i}$	i. sınıf örnek ortalama vektörü
$\mathbf{x_k}$	Karışmalı veri seti
$\mathbf{X_{p*n}}$	p*n boyutlu veri matrisi
$\mathbf{y_i}$	i. temel bileşen vektörü
$\mathbf{Z_{p*n}}$	p*n boyutlu standart veri matrisi

Kısaltmalar	Açıklamalar
AER	Hatalı sınıflandırma oranı
DA	Diskriminant analizi
DE	Duyarlılık eğrisi
det	Determinant
IF	Etki fonksiyonu
KN	Kırılma noktası değeri
Köş	Matrisin köşegen elemanları
LR	Olabilirlik oranı

Kısaltmalar	Açıklamalar
MAD	Medyan mutlak sapma istatistiği
MCD	En küçük kovaryans determinant
med	Medyan istatistiği
min	En küçük
OGTT	Oral glikoz tolerans testi
PP	İzdüşüm arama
$\mathbf{PP_{MAD}}$	MAD istatistiği kullanılan izdüşüm arama yöntemi
SSPG	Kararlı durum plazma glikozu
TBA	Temel bileşenler analizi

1. GİRİŞ

Çok değişkenli gözlemleri sınıflandırmak, gruplara atamak için kullanılan çeşitli istatistiksel analiz yöntemleri vardır. Ancak, bu yöntemleri uygularken karşılaşılan en önemli sorun gerçek veri setlerinde başka dağılımdan karışan gözlemlerin bulunması durumudur. Başka dağılımdan karışan gözlemlerin varlığı istatistiksel analiz yöntemlerinin hatalı, güvenilir olmayan sonuçlar vermesine neden olmaktadır.

Çok değişkenli istatistiksel analiz yöntemlerinden biri olan temel bileşenler analizi (TBA), verinin kovaryans yapısını daha az sayıda bileşenler cinsinden açıklamayı amaçlar, yani bir çeşit veride boyut indirgeme yöntemidir. Bu bileşenler, orijinal değişkenlerin doğrusal bileşimidir. Farklı değişim kaynaklarının daha iyi yorumlanmasına ve anlaşılmasına olanak sağlamaktadır. Bu yüzden, temel bileşenler analizi bilginin çoğunu içeren bileşenleri bulmak için önemlidir. Temel bileşenler analizi veri analizinin ilk aşamasıdır, bunu diskriminant analizi (DA), kümeleme analizi veya diğer çok değişkenli analiz teknikleri takip etmektedir (Hubert ve Engelen, 2004). Çok değişkenli gözlemleri gruplara atamada kullanılan istatistiksel analiz yöntemlerinden birisi diskriminant analizidir. Diskriminant analizi, hangi gruptan geldiği belli olmayan bir gözlemin hangi gruba dahil edileceğini belirlemek için kullanılmaktadır. Ancak diskriminant analizinin önemli bir kısıtlayıcısı olan başka dağılımdan karışan gözlemlerin var olmaması durumu sağlanmadığında hatalı grup atamaları gerçekleştirilebilir.

Başka dağılımdan karışan gözlemlerin var olması durumunda, temel bileşenleri bulmak ve etkili bir grup ataması yapabilmek için başka dağılımdan karışan gözlemlerden etkilenmeyen sağlam tahmin ediciler tercih edilmektedir. Sağlam tahmin yöntemleri ile ilgili ilk çalışmalar, Simon Newcomb (1886) tarafından 19. yüzyılın sonlarında yapılmaya başlanmıştır. Ancak, 1960'larda, Tukey (1962) ve Huber (1964, 1967) tarafından önemli adımlar atılmıştır. Bu araştırmacılar tarafından önerilen yeni sağlam yöntemler, bilgisayar teknolojisindeki gelişim ve bilgisayarlara erişimin kolaylığı ile kolayca uygulanabilir hale gelmiştir. Son yıllarda sağlam istatistiklerin çalışma alanı, bir araştırma alanı olarak önemli derecede büyüme sağlamıştır, buna yayınlanmış makalelerin sayısının fazlalığı kanıt olarak gösterilebilir. Ayrıca, sağlam istatistiksel yöntemler ile ilgili olarak, Maronna, Martin ve Yohai (2006) ve Johnson ve Wichern (2007) tarafından etkili kitaplar yazılmıştır. Öte yandan,

son yıllarda, Hubert, Rousseeuw ve Branden (2005), Hubert, Rousseeuw ve VanAelst (2008), Todorov ve Filzmoser (2009) ve Pires ve Branco (2010) tarafından sağlam istatistiklerle ilgili önemli makaleler yayınlanmıştır.

Bu tez çalışmasında, temel bileşenler analizi ve diskriminant analizi uygulamalarında veride hatalı sonuçlar elde etmemize yol açan başka dağılımdan karışan gözlemler olması durumunda, tahmin sonuçlarını başka dağılımdan karışan gözlemlerin etkisinden kurtarabilen, kırılma noktası değeri yüksek olan sağlam tahmin ediciler incelenmektedir. Bu amaçla; ikinci bölümde, temel bileşenler analizi tanımlanmış, önemli temel bileşen sayısına karar verilmesi ve önemli temel bileşenlerin belirlenmesi anlatılmıştır. Üçüncü bölümde, diskriminant analizi tanımlanmış, diskriminant fonksiyonlarının oluşturulması ve diskriminant fonksiyonları yardımı ile gözlemlerin atanacakları sınıfların belirlenmesi anlatılmıştır. Dördüncü bölümde, başka dağılımdan karışan gözlemler ve belirlenme yolları anlatılmıştır. Beşinci bölümde, sağlamlık kavramı tanımlanarak sağlamlık özelliğine sahip tahmin edicilerin uygulanma algoritmaları anlatılmıştır. Altıncı bölümde, uygulamalarda sıklıkla kullanılan veri setleri ile diskriminant analizi ve temel bileşenler analizi uygulamaları yapılmıştır. Yedinci bölümde, sağlam tahmin ediciler ve klasik tahmin edicilerle diskriminant analizi uygulaması üzerine bir simülasyon çalışması yapılmıştır. Sekizinci bölümde, elde edilen bulgular açıklanmıştır.

2. TEMEL BİLEŞENLER ANALİZİ

20. yüzyılın ikinci yarısından itibaren bilim ve teknoloji alanındaki hızlı ilerleme ile çok fazla sayıda ve yüksek boyutlu verinin depolanması mümkün hale gelmiştir. Çok boyutlu verilerle karşılaşma olasılığı artmasına rağmen bu verilerle anlamlı analizler yapılmadığında yüksek boyutlu verileri depolamak anlamsızlaşmaktadır. Yüksek boyutlu veriler ile anlamlı analizler yapabilmek için çok değişkenli istatistiksel tekniklere başvurulmaktadır. Yüksek boyutlu verilerde değişkenler arasında yüksek korelasyon (ilişki) bulunabilmektedir. Çok değişkenli analizde gözlemlere ilişkin ölçülen p tane değişkenin (özellik) birbiriyle ilişkili ve değişken sayısının çok fazla olması durumunda analizde sorunlar yaşanabilmektedir. Bu durum, değişkenlerin bağımsızlığı kuralını zedelediği gibi, çok sayıda değişkenle çalışmak işlem yükünü arttırmakta ve elde edilen sonuçların yorumunda bazı güçlüklere neden olmaktadır. Aralarında yüksek ilişki bulunan çok sayıda değişken ile açıklanmak istenen problemin aralarında ilişki bulunmayan az sayıda değişken ile açıklanması temel bileşenler analizi (TBA) ile sağlanabilmektedir. TBA, daha az sayıda bileşenle verinin kovaryans yapısını açıklamaya çalışan popüler bir çok değişkenli istatistiksel yöntemdir. Genel olarak, değişkenler arasındaki bağımlılık yapısının yok edilmesi ve boyut indirgeme amacıyla kullanılmaktadır (Tatlıdil, 2002). Bu temel bileşenler, orijinal değişkenlerin doğrusal bileşimleridir. Temel bileşenler analizi farklı değişim kaynaklarının daha iyi anlaşılmasına ve yorumlanmasına izin veren bir veri indirgeme yöntemi olarak da tanımlanabilir. Çoğunlukla kemometrik, bilgisayar görüntüsü, mühendislik, genetik ve diğer alanlarda karşılaşılan yüksek boyutlu verinin analizi için yaygın biçimde kullanılır. Temel bileşenler analizi yüksek boyutlu verilere çok değişkenli analiz teknikleri uygulanabilmesi için verinin hazırlanması aşamasıdır, bunu diskriminant analizi, kümeleme analizi veya diğer çok değişkenli teknikler takip eder (Hubert ve Engelen, 2004).

Temel bileşenler analizinde veriye ilişkin korelasyon matrisindeki toplam değişim (varyans) dikkate alınmaktadır. Yani, temel bileşenler analizi toplam değişimi analiz etmektedir. Bir değişkenin kendisi ile korelasyonu 1'e eşit olup bu değer aynı zamanda o değişkenin toplam değişimini verir. Temel bileşenler çözümlemesi sonucunda, p değişkenli uzayı çok iyi tanımlayan k tane yeni ilişkisiz (bağımsız) değişken, yani temel bileşen elde edilmektedir. p tane değişkenin taşıdığı bilginin k tane $(k < p)$ yeni değişken ile açıklanması temel bileşenler analizinin ana amacını oluşturmaktadır. Diğer bir ifade ile,

verideki bilginin çoğunu kapsayan daha az sayıda yeni birbirine dik değişken elde etmek temel bileşenler analizinin en önemli amacıdır (Alpar, 2011).

Elde edilen temel bileşenler değişkenlerin ağırlıklı doğrusal bileşimleri olup veri kümesindeki en fazla değişimin derece derece değişimini temsil etmektedirler. Temel bileşenler analizinde toplam değişim (λ), standartlaştırılmış özvektörlerle ifade edilir. Özdeğerler (λ_i), $\lambda_1 > \lambda_2 > \lambda_3 > \ldots > \lambda_p$ şeklinde sıralanır ve en büyük değişim (özdeğer) birinci standartlaştırılmış özvektöre ait olur. Diğer bir ifade ile, değişkenler kümesindeki toplam değişimin büyük bir bölümü birinci temel bileşen, ondan daha az bir bölümü ikinci temel bileşen tarafından açıklanmaktadır. Genel olarak, toplam varyansın büyük bir bölümünün ilk iki temel bileşen tarafından açıklanması arzu edilmektedir (Alpar, 2011).

2.1. Temel Bileşenler Analizi Aşamaları

Değişkenlerin ölçü birimlerinin aynı olması pratikte sağlanamadığından, temel bileşenler analizinde genellikle ham veri matrisi yerine standartlaştırılmış değerlerden oluşan Z_{p*n} standart veri matrisi kullanılmaktadır. Değişkenlerin ölçü birimlerinin aynı olması durumunda varyans-kovaryans matrisinden, farklı olması durumunda korelasyon matrisinden (R_{p*p}) yararlanılmaktadır. Temel bileşenler analizinde amaç, birbirleriyle ilişkili z_{ij} değerlerinden dönüştürme yolu ile birbiriyle ilişkisiz y_{ij} değerlerini elde etmektir.

Temel bileşenler fonksiyonları (y_i) korelasyon (R_{p*p}) matrisinin özvektörleri (t_i) yardımı ile hesaplanmaktadır. Bunun için öncelikle, $|R - \lambda I| = 0$ açılımı yardımıyla korelasyon (R_{p*p}) matrisinin p tane λ özdeğerleri hesaplanır. Elde edilen p tane özdeğerin her birine karşılık gelen özvektörler yardımı ile temel bileşen fonksiyonları elde edilir. Yani, i. temel bileşen fonksiyonu $y_i = t_i' Z$ ile elde edilir. Temel bileşenler analizi uygulaması aşağıdaki sıralamaya uygun biçimde yapılmaktadır.

İlk olarak, temel bileşenler analizinin gerekli olup olmadığına karar vermek gerekmektedir. Yani, değişkenler arasındaki ilişkinin önemli olup olmadığına karar vermek gerekmektedir, bu amaçla "Bartlett Küresellik Testi" nden faydalanılır.

$H_0 : R = I$ (değişkenler arasındaki ilişkiler önemsizdir.)

$H_1 : R \neq I$ (değişkenler arasındaki ilişkiler önemlidir.)

şeklinde H_0 yokluk ve H_1 alternatif hipotezleri kurulmaktadır. n gözlem sayısı, p değişken sayısı olmak üzere Bartlett test istatistiği Eş. 2.1'deki formül ile hesaplanmaktadır.

$$-[(n-1)-\frac{1}{6}(2p+5)]ln(|R|) \geq \chi^2_{\frac{p(p-1)}{2}} \tag{2.1}$$

Eş. 2.1'de verilen koşul sağlanıyorsa, H_0 hipotezi reddedilmektedir. Yani, değişkenler arasındaki ilişkinin önemli olduğu ve temel bileşenler analizi uygulanabileceği sonucuna ulaşılmaktadır.

Daha sonra, korelasyon (R_{p*p}) matrisinin p tane (değişken sayısı kadar) λ_i özdeğerleri $|R-\lambda I| = 0$ açılımı yardımı ile hesaplanmaktadır.

Korelasyon (R_{p*p}) matrisinin 1'den büyük olan özdeğerleri (λ_i) veya önemli özdeğer sayısı k olmak üzere, $\sum_{i=1}^{k} \lambda_i/p \geq \frac{2}{3} = 0,66$ koşulunu sağlayan özdeğerler (λ_i) belirlenir. Bu özdeğerlerin sayısı (k) önemli temel bileşen sayısını vermektedir. Önemli özdeğer sayısının belirlenmesi Kısım 2.2'de ayrıntılı olarak açıklanmaktadır.

Önemli özdeğerlere karşılık gelen özvektörler (t_i) ile, özvektörler matrisi oluşturulur. p tane özdeğerin her birine karşılık gelen özvektörler (t_i) yardımı ile; i. temel bileşen fonksiyonu $y_i = t_i^{'}Z$ şeklinde elde edilir.

Özdeğer ağırlıklarının katsayılara etki etmesini sağlamak için asıl temel bileşen vektörleri ($V_i = \sqrt{\lambda_i} t_i$) yardımı ile asıl temel bileşenler matrisi ($V = [V_1\ V_2\ \ldots V_k]$) elde edilir. Asıl temel bileşenler matrisinde değeri mutlak değerce %50'nin üzerinde olan değişkenlerin o bileşenin oluşumunda etkili olduğu söylenebilir.

Asıl temel bileşen vektörlerindeki değerlerin kareleri (V_i^2) alınarak yükler matrisi (W) tüm temel bileşenler için elde edilir. Yükler matrisinin sütun toplamları özdeğerleri (λ_i) vermektedir. Her bir özdeğerin toplam değişken sayısına oranı ($\frac{\lambda_i}{p}$) o temel bileşenin toplam varyansı açıklama oranını vermektedir. Yükler matrisinin satır toplamları (h_i^2) ise bütün temel bileşenlerin ilgili satırdaki değişkeni (x_i) açıklama oranını vermektedir. İlgili

satırdaki değişkenin temel bileşenler tarafından açıklanamayan kısmı (b_i^2) ise $b_i^2 = 1 - h_i^2$ ile hesaplanmaktadır.

Beşinci adımda hesaplanan temel bileşenler ile, birbiri arasında yüksek ilişki bulunan değişkenlerden, aralarında ilişki bulunmayan, bağımsızlık özelliğine sahip yeni değişkenler elde edilmiş olur. Mevcut değişkenlerden elde edilen bu yeni bileşenler ile veri analize hazır hale gelmektedir ve çok değişkenli analiz tekniklerine hesaplanan temel bileşenler ile devam edilebilmektedir.

2.2. Temel Bileşen Sayısının Belirlenmesi

Temel bileşenler analizinin amaçlarından bir tanesi değişkenlerin boyutunun indirgenmesidir. Bu nedenle, Kısım 2.1'de bahsedildiği gibi, değişken boyutunun indirgenmesinde önemli özdeğer sayısına karar vermek çok önemlidir. Bu amaçla, çeşitli yaklaşımlar geliştirilmiştir. Burada, bu ölçütler arasından yaygın olarak kullanılan, özdeğer ölçütü, varyans yüzdesi ölçütü ve yamaç grafiği yaklaşımı açıklanmaktadır.

2.2.1. Özdeğer ölçütü

Özdeğer ölçütü, en çok yararlanılan yaklaşımlardan biri olup hem temel bileşenler hem de faktör analizi yöntemlerinde kullanılmaktadır. Bu ölçüt, bir bileşenin açıklayıcılığının en azından bir değişkenin açıklayıcılığı kadar olması mantığına dayanmaktadır. Bu nedenle, bu yaklaşımda öz değeri 1'den büyük olan değişkenler "*önemli temel bileşenler*" olarak dikkate alınmaktadır. 1'e çok yakın özdeğere sahip bir bileşenin dikkate alınmaması bu yaklaşımın olumsuz yönlerinden birisidir. Genellikle, değişken sayısının 20 ile 50 arasında olduğu durumlarda kullanılması önerilmektedir. Bu yaklaşımın dezavantajı, değişken sayısının az olduğu durumlarda veri yapısında gerçekten var olandan daha az bileşen, değişken sayısı çok olduğunda da anlamlı olandan daha fazla temel bileşen belirlenebilmesidir (Alpar, 2011).

2.2.2. Varyans yüzdesi ölçütü

Önemli özdeğer sayısının belirlenmesinde bir diğer önemli ölçüt varyans yüzdesi ölçütüdür. Bu çerçevede, k önemli özdeğer sayısını göstermek üzere, $\sum_{i=1}^{k} \lambda_i / p \geq \frac{2}{3} = 0,66$ koşulunun

sağlandığı en küçük k değeri önemli temel bileşen sayısı olarak belirlenir (Alpar, 2011). Bu ölçütün dezvantajı, eğer birinci özdeğer yeterince büyük ve koşulu sağlıyor ise ikinci özdeğer 1'den büyük olsa da hesaba katılmayabilir.

2.2.3. Yamaç grafiği yaklaşımı

Yamaç grafiği yaklaşımı, temel bileşenlerle elde edilen özdeğerlerin grafiğine dayanmaktadır. Bu yaklaşımda yatay eksende bileşen numaraları, dikey eksende ilgili bileşenin özdeğerleri ya da varyans açıklama yüzdeleri yer almaktadır. Eğimin azaldığı, değişmezleştiği ya da çok azalan değerlere ulaştığı noktadaki özdeğer sayısı kadar bileşenin dikkate alınması önerilmektedir. Yamaç grafiği yaklaşımı, özdeğer ölçütüne göre daha fazla bileşenin dikkate alınması gerektiği şeklinde bir sonuç vermektedir (Alpar, 2011).

3. DİSKRİMİNANT ANALİZİ

Fisher (1936) iki sınıfı (grubu) birbirinden anlamlı bir şekilde ayırabilmek amacı ile, açıklayıcı değişkenlerin doğrusal bir fonksiyonu olan diskriminant fonksiyonunu oluşturmuştur. Diskriminant analizi, gözlemlerin atanacağı sınıfların sayısı ve hangi gözlemin hangi sınıfta olduğunun önceden belli olduğu, gözetimli sınıflandırma için yaygın kullanılan istatistiksel bir yöntemdir, yani, diskriminant analizinin temel amacı gözlemleri sınıflandırmaktır (Filzmoser, Hron ve Temple, 2012). Bu anlamda, bir karar verme olayı olarak da tanımlanabilir. Diskriminant analizinin en önemli özelliklerinden birisi belirlenen gruplardan herhangi birisine yeni birimleri atamak ya da mevcut birimlerin hangi gruptan geldiğini belirlemektir. Çok boyutlu uzayda gruplaşma gösteren birimlerin birbirinden anlamlı bir şekilde ayrılıp ayrılamayacağı, oluşturulacak grup sayısı, her gruba atanacak birimlerin belirlenmesi, sınıflandırma oranı ve grupları ayırmada katkısı olan özelliklerin neler olacağı soruları diskriminant analizi ile cevaplanmaktadır (Albayrak, 2006). Mevcut veya gelecekte elde edilecek birimleri gruplardan birine atamak için bir kriter geliştirme yöntemi olan diskriminant analizi, hatalı sınıflandırma olasılığını en aza indirgeyerek birimleri ait oldukları gruplara atamak, gelmiş oldukları yığınları belirlemektedir (Tatlıdil, 2002).

Diskriminant analizi uygulanacak veri seti l farklı sınıftan n_1, n_2, ..., n_l büyüklüğünde örneklerin p tane özelliğinin (değişken) ölçülmesi ile elde edilir. i grupları ($i = 1, 2, \ldots, l$), j değişkenleri ($j = 1, 2, \ldots, p$) ve n gözlemleri ($n = 1, 2, \ldots, \sum_{i=1}^{l} n_l$) ifade etmek üzere diskriminant analizi uygulanacak veri matrisi Eş. 3.1'deki gibi ifade edilebilir.

$$X = \begin{bmatrix} x_{111} & x_{121} & \cdots & x_{1p1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{11n_1} & x_{12n_1} & \cdots & x_{1pn_1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{l11} & x_{l21} & \cdots & x_{lp1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{l1n_l} & x_{l2n_l} & \cdots & x_{lpn_l} \end{bmatrix} \tag{3.1}$$

$i = 1, 2, \ldots, l$ grupları ve X_i her bir gruba ait veri matrisini (Eş. 3.2) göstermektedir.

$$X_i = \begin{bmatrix} x_{i11} & x_{i21} & \cdots & x_{ip1} \\ x_{i12} & x_{i22} & \cdots & x_{ip2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1n_i} & x_{i2n_i} & \cdots & x_{ipn_i} \end{bmatrix} \tag{3.2}$$

Buradan ana veri matrisi X, Eş. 3.3'teki gibi vektörel biçimde de ifade edilebilir.

$$X = \begin{bmatrix} \underline{X_1} \\ \underline{X_2} \\ \vdots \\ \underline{X_i} \end{bmatrix} \tag{3.3}$$

3.1. Diskriminant Analizinin Varsayımları ve Kısıtlayıcıları

Diskriminant analizinin iki temel varsayımı, çok değişkenli normallik ve varyans-kovaryans matrislerinin homojenliği varsayımlarıdır (Alpar, 2013). Ancak bu varsayımlar sağlanmadığı durumlar için de farklı diskriminant kuralları geliştirilmiştir. Geliştirilmiş olan farklı diskriminant kuralları Kısım 3.3'te detaylı olarak açıklanmaktadır.

3.1.1. Çok değişkenli normallik varsayımı

Çok değişkenli normallik varsayımının sağlanıp sağlanmadığını test etmenin çeşitli yöntemleri mevcuttur. Bu yöntemlerden sıklıkla kullanılan iki tanesi aşağıdaki gibi özetlenmektedir.

Çok değişkenli dağılımlarda, her bir gözlemin merkeze uzaklığı olarak ifade edien m_i^2 uzaklık $(i = 1, 2, \ldots, n)$ değerleri $(m_i^2$=Köş$((x_i - \bar{x})' S^{-1}(x_i - \bar{x})))$ ile p serbestlik dereceli ki-kare $(\chi^2_{p;(i-0,5)/n})$ değerlerine ilişkin saçılım grafiğinin yaklaşık bir doğru üzerinde olması durumunda veri setinin çok değişkenli normal dağılıma uyduğu söylenebilmektedir (Alpar, 2013).

Çok değişkenli bir dağılımın normal dağılım gösterip göstermediğini anlamanın bir diğer

yolu Mardia'nın çok değişkenli çarpıklık ve basıklık katsayılarından yararlanmaktır. $m_{ij} = (x_i - \bar{x})' S^{-1} (x_j - \bar{x})$ olmak üzere Mardia'nın çarpıklık katsayısı Eş. 3.4'te verildiği gibi elde edilmektedir.

$$\hat{\gamma}_{1p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij}^3 \tag{3.4}$$

Mardia'nın basıklık katsayısı ise Eş. 3.5'te verildiği gibi elde edilmektedir.

$$\hat{\gamma}_{2p} = \frac{1}{n} \sum_{i=1}^{n} m_i^4 \tag{3.5}$$

Eğer $\frac{n\hat{\gamma}_{1p}}{6}$ değeri $\frac{p(p+1)(p+2)}{6}$ serbestlik dereceli ki-kare değerine ($\chi^2_{p(p+1)(p+2)/6;(0,05)}$) eşit ya da küçük ise veri setinin çok değişkenli normalliğe ilişkin çarpıklığının sağlandığı söylenebilir. $\hat{\gamma}_{2p}$ basıklık değeri $p(p+2)$ ortalama ve $\frac{8p(p+2)}{n}$ varyansı ile normal dağılım gösterir. $\hat{\gamma}_{2p}$ basıklık değeri için standart normal z değerinin iki yönlü olasılığı $0,05$ değerinden büyük ise veri setinin çok değişkenli normalliğe ilişkin basıklığının sağlandığı söylenebilir (Alpar, 2013).

Eğer, çok değişkenli normallik varsayımı sağlanmazsa doğrusal kombinasyonların kestiriminde sorunlar ortaya çıkabilmektedir. Ancak, çok değişkenli normallik varsayımının sağlanmadığı durum için Kısım 3.3.2'de ve Kısım 3.3.3'te bahsedilen Fisher diskriminant analizi geliştirilmiştir.

3.1.2. Varyans-kovaryans matrislerinin homojenliği varsayımı

Örneklem büyüklükleri yetersiz olduğunda ve varyans-kovaryans matrisleri homojen olmadığında, diskriminant fonksiyonlarının tahmin edilmesi işlemlerinin istatistiksel anlamlılığı olumsuz olarak etkilenebilmektedir. Gruplardaki örneklem büyüklüklerinin yeterli olması ancak varyans-kovaryans matrislerinin homojen olmaması durumunda, gözlemler daha yüksek kovaryansa sahip olan gruplara sınıflandırılabilirler. Varyans-kovaryans matrislerinin homojenliği (eşitliği) Box-M testi ile test edilebilmektedir. Herhangi bir i'inci sınıfın (grubun) geldiği kitleye ilişkin varyans-kovaryans matrisi (Σ_i)

Eş. 3.6'da verilen matris ile ifade edilmektedir.

$$\Sigma_i = \begin{bmatrix} \sigma_{i11} & \sigma_{i12} & \cdots & \sigma_{i1p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{ip1} & \sigma_{ip2} & \cdots & \sigma_{ipp} \end{bmatrix} \tag{3.6}$$

Herhangi bir i'inci sınıfın (grubun) geldiği kitleye ilişkin varyans-kovaryans matrisi (Σ_i) Eş. 3.6'da verildiği gibi olmak üzere, varyans-kovaryans matrislerinin eşitliği yani homojenliği için Box-M testine ilişkin hipotezler aşağıdaki gibi kurulmaktadır.

$H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_l$
H_1 : En az bir varyans-kovaryans matrisi diğerlerinden farklıdır.

Bu hipotezi test etmek için, karşılaştırılacak l tane grubun örnekleme ilişkin varyans-kovaryans matrisleri olan S_i'ler yardımı ile ortak varyans-kovaryans matrisi S_{ortak} Eş. 3.7'deki gibi elde edilmektedir.

$$S_{ortak} = \frac{\sum_{i=1}^{l}(n_i - 1)S_i}{\sum_{i=1}^{l}(n_i - 1)} \tag{3.7}$$

Grupların ortak varyans-kovaryans matrisi S_{ortak} yardımı ile varyans-kovaryans matrislerinin homojenliğini test etmek için Box-M test istatistiği (MC^{-1}) Eş. 3.8 ve Eş. 3.9 yardımı ile elde edilmektedir.

$$M = \sum_{i=1}^{l}(n_i - 1)ln|S_{ortak}| - \sum_{i=1}^{l}(n_i - 1)ln|S_i| \tag{3.8}$$

Gruplardaki gözlem sayıları n_i'ler yeterince büyükken;

$$C^{-1} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(l-1)}\left(\sum_{i-1}^{l}\frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^{l}(n_i - 1)}\right) \tag{3.9}$$

olmak üzere Box-M test istatistiği (MC^{-1}) değeri $\frac{1}{2}(l-1)p(p+1)$ serbestlik derecesi ile yaklaşık ki-kare dağılımı gösterir. Box-M test istatistiği (MC^{-1}) değeri $\frac{1}{2}(l-1)p(p+1)$

serbestlik dereceli ki-kare ($\chi^2_{\frac{1}{2}(l-1)p(p+1)}$) tablo değerinden büyük ($MC^{-1} > \chi^2_{\frac{1}{2}(l-1)p(p+1)}$) ise varyans-kovaryans matrislerinin eşit olmadığı söylenebilir (Alpar, 2013).

Ancak, varyans-kovaryans matrislerinin eşitliği varsayımının sağlanmadığı durumlar için de diskriminant kuralları geliştirilmiştir. Geliştirilen diskriminant kurallarına Kısım 3.3.4'te değinilmektedir.

3.1.3. Kısıtlayıcılar

Bu iki önemli varsayımın dışında diskriminant analizi, açıklayıcı değişkenler arasında çoklu bağlantı sorununun olmaması, açıklayıcı değişkenler arasındaki ilişkilerin doğrusal olması ve veride başka dağılımdan karışan gözlemlerin olmaması durumlarında daha iyi sonuç vermektedir. Ancak tüm bu varsayımlar ve kısıtlayıcılar her zaman sağlanmayabilmektedir. Bu gibi durumlarda farklı diskriminant kuralları ve başka dağılımdan karışan gözlemlerden etkilenmeyen sağlam tahmin ediciler kullanılmaktadır.

3.2. Diskriminant Fonksiyonlarının Belirlenmesi, Önemlilik Testi

Tüm değişkenler dikkate alınarak gözlemlerin dahil olduğu sınıflar (gruplar) arasında önemli bir farklılık olup olmadığı ya da gözlemlerin dahil olduğu sınıfların incelenen açıklayıcı değişkenlerin doğrusal kombinasyonu yardımı ile anlamlı bir şekilde birbirinden ayırt edilip edilemeyeceği belirlenmektedir. Bu amaçla, çok değişkenli F testlerinden yararlanılmaktadır. Test işlemi $Wilks'in\ Lambda$ (Λ) istatistiğinin F dağılımına dönüştürülmesi ile yapılabilmektedir.

l karşılaştırılacak sınıf sayısı, $\bar{x}_i$ $i.$ sınıfa ilişkin ortalama vektörü, $\bar{x}$ genel ortalama vektörü, n_i $i.$ sınıf gözlem sayısı, S_i $i.$ sınıf varyans-kovaryans matrisi olmak üzere, gruplar arası kareler toplamı (B_{p*p}) matrisi Eş. 3.10'daki gibi elde edilmektedir.

$$B_{p*p} = \sum_{i=1}^{l} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \tag{3.10}$$

Ve grup içi kareler toplamı (W_{p*p}) matrisi ise Eş. 3.11'deki gibi elde edilmektedir.

$$W_{p*p} = \sum_{i=1}^{l} (n_i - 1) S_i \tag{3.11}$$

Gruplara ilişkin genel kareler toplamı matrisi (A_{p*p}), gruplar arası kareler toplamı matrisi (B_{p*p}) ve grup içi kareler toplamı matrisinin (W_{p*p}) toplamı olarak $A_{p*p} = B_{p*p} + W_{p*p}$ şeklinde tanımlanabilmektedir.

Bu bilgiler yardımı ile, *Wilks'in Lambda* (Λ) istatistiği, grup içi varyans-kovaryans matrisinin determinantının ($|W|$), grup içi varyans-kovaryans matrisi ile gruplar arası varyans-kovaryans matrisinin toplamlarının determinantına ($|W + B|$) bölünmesi ile Eş. 3.12'deki gibi tanımlanmaktadır.

$$\Lambda = \frac{|W|}{|W + B|} \tag{3.12}$$

Hesaplanan *Wilks'in Lambda* (Λ) istatistiğinin sıfıra yakın olması, sınıflar arasında anlamlı bir fark olduğunun göstergesidir. *Wilks'in Lambda* (Λ) istatistiğinin anlamlılığı F testi yardımı ile test edilir. Ve böylece p tane bağımsız değişkenin oluşturacağı kombinasyon yardımı ile l tane sınıfın ayırt edilip edilemeyeceği ve de yeni bir gözlemin sınıflardan birine doğru olarak atanıp atanamayacağı belirlenmektedir.

3.3. Sınıflama: Yeni Gözlemler İçin Sınıf Üyeliklerinin Belirlenmesi

Diskriminant analizi iki aşamalı bir işlem olarak ele alınabilmektedir. Öncelikle, diskriminant fonksiyonları belirlenmekte, daha sonra, belirlenen diskriminant fonksiyonları kullanılarak gözlemler ait oldukları sınıflara atanmaktadır.

3.3.1. Diskriminant analizi için hipotez testi

Diskriminant analizinde gözlemleri ait oldukları sınıflara atamakta kullanılan yöntemlerden biri olabilirlik oranı kriteri, yani, hipotez testi yaklaşımıdır. Uygulamaların çoğunda yığına ilişkin bilgiler bilinmemektedir ve her bir yığından alınan örneklerden yığın bilgilerine ilişkin çıkarımlar yapılmaktadır. Yani, normal dağılıma sahip iki farklı yığının birisinden gelen bir gözlemin sınıflandırmasında, yığından çekilmiş örnekten elde

edilen bilgi kullanılmaktadır. Ölçek parametresi olan varyans-kovaryans matrisleri eşit, konum parametresi olan ortalama vektörleri farklı çok değişkenli normal dağılıma sahip iki yığın π_1 ve π_2 olmak üzere, π_1 yığını $N(\mu_1,\ \Sigma)$ dağılımına sahipken, π_2 yığını $N(\mu_2,\ \Sigma)$ dağılımına sahiptir. $N(\mu_1,\ \Sigma)$ dağılımından $x_1^{(1)},\ \ldots,\ x_{n_1}^{(1)}$ rassal örneği ve $N(\mu_2,\ \Sigma)$ dağılımından ise $x_1^{(2)},\ \ldots,\ x_{n_2}^{(2)}$ rassal örneği alınmakta ve bu örneğe eğitim örneği denilmektedir. $N(\mu_1,\ \Sigma)$ dağılımına sahip π_1 sınıfına ilişkin yığının konum parametresi μ_1 için en iyi tahmin, yani, en çok olabilirlik tahmin edicisi $\bar{x}_1 = \sum_{i=1}^{n_1} \frac{x_i^{(1)}}{n_1}$ ve $N(\mu_2, \Sigma)$ dağılımına sahip π_2 sınıfına ilişkin yığının konum parametresi μ_2 için en iyi tahmin, yani, en çok olabilirlik tahmin edicisi $\bar{x}_2 = \sum_{i=1}^{n_2} \frac{x_i^{(2)}}{n_2}$ ve π_1 ve π_2 sınıflarına ilişkin ölçek parametresi yani eşit varyans-kovaryans matrisi Σ için en iyi tahmin, yani en çok olabilirlik tahmin edicisi $(n_1+n_2-2)S = \sum_{i=1}^{n_1}(x_i^{(1)}-\bar{x}^{(1)})(x_i^{(1)}-\bar{x}^{(1)})' + \sum_{i=1}^{n_2}(x_i^{(2)}-\bar{x}^{(2)})(x_i^{(2)}-\bar{x}^{(2)})'$ şeklinde tanımlanan S istatistiğidir (Anderson, 1984).

Bu bilgiye dayalı olarak, π_1 veya π_2 sınıfından gelen ve hangi sınıfta olduğu bilinmeyen bir x gözleminin atanacağı sınıfı belirlemek için; $x,\ x_1^{(1)},\ \ldots,\ x_{n_1}^{(1)}$ rassal örneğinin $N(\mu_1,\ \Sigma)$ dağılımından alındığı ve $x_1^{(2)},\ \ldots,\ x_{n_2}^{(2)}$ rassal örneğinin $N(\mu_2,\ \Sigma)$ dağılımından alındığı H_0 yokluk hipotezine karşılık, H_1 alternatif hipotezi, $x_1^{(1)},\ \ldots,\ x_{n_1}^{(1)}$ rassal örneğinin $N(\mu_1,\ \Sigma)$ dağılımından alındığı ve $x,\ x_1^{(2)},\ \ldots,\ x_{n_2}^{(2)}$ rassal örneğinin $N(\mu_2,\ \Sigma)$ dağılımından alındığı şeklinde kurulmaktadır.

$$H_0 : x, x_1^{(1)}, x_2^{(1)}, \ldots x_{n_1}^{(1)} \in \pi_1;\ x_1^{(2)}, x_2^{(2)}, \ldots x_{n_2}^{(2)} \in \pi_2$$
$$H_1 : x_1^{(1)}, x_2^{(1)}, \ldots x_{n_1}^{(1)} \in \pi_1;\ x, x_1^{(2)}, x_2^{(2)}, \ldots x_{n_2}^{(2)} \in \pi_2$$

$N(\mu_1,\ \Sigma)$ dağılımına sahip π_1 sınıfından alınan ve $N(\mu_2,\ \Sigma)$ dağılımına sahip π_2 sınıfından alınan iki eğitim verisinin olabilirliği (L) ise Eş. 3.13'teki gibi yazılabilir:

$$L = \{(2\pi)^p |\Sigma|\}^{-\frac{1}{2}(n_1+n_2)} \prod_{i=1}^{2} [(\prod_{m=1}^{2} p_{im}) exp\{-\frac{1}{2}\sum_{j=1}^{n_i}(x_j^{(i)}-\mu_{ij})'\Sigma^{-1}(x_j^{(i)}-\mu_{ij})\}]. \tag{3.13}$$

Eğer hangi sınıfta olduğu bilinmeyen bir x gözlemi π_1 sınıfından gelen eğitim veri setine dahil ise, Eş. 3.14'te verilen çarpım faktörü (L_{im}) Eş. 3.13'e dahil edilir:

$$L_{im} = (2\pi)^{-\frac{1}{2}p} - |\Sigma|^{-\frac{1}{2}} p_{im} exp\{-\frac{1}{2}(x-\mu_i^{(m)})'\Sigma^{-1}(x-\mu_i^{(m)})\}. \tag{3.14}$$

Böylece, hipotez testi ile atama kuralı, Eş. 3.15'e göre belirlenir ve olabilirlik oranı (LR) 1'den büyük ($LR > 1$) ise hangi sınıfa ait olduğu bilinmeyen bir x gözlemi π_1 sınıfına, diğer durumlarda ise π_2 sınıfına atanmaktadır (Krzanowski, 1982).

$$LR = \frac{sup(L_{1m}L)}{sup(L_{2m}L)} \tag{3.15}$$

θ^i konum parametresi iken, bu yaklaşımın genelleştirilmiş hali için, iki eğitim örneğinin olabilirliği Eş. 3.16'daki gibidir:

$$L(\theta^{(1)},\theta^{(2)}|x_1^{(1)},x_2^{(1)},\ldots x_{n_1}^{(1)},x_1^{(2)},x_2^{(2)},\ldots x_{n_2}^{(2)}) = \prod_{i=1}^{n_1} f_0(x_i^{(1)}|\theta^{(1)}) \prod_{i=1}^{n_2} f_1(x_i^{(2)}|\theta^{(2)}). \tag{3.16}$$

Hangi sınıfta olduğu bilinmeyen bir x gözlemi π_i sınıflarından bir eğitim örneğine dahil ise, Eş. 3.16'da verilen olabilirliğe bir de Eş. 3.17'de verilen çarpım faktörü dahil edilmektedir.

$$L_i(\theta^{(i)}|x) = f_i(x|\theta^{(i)}) \tag{3.17}$$

Böylece, $\hat{\theta}_0^{(i)}$, H_0 hipotezi altında $\theta^{(i)}$'nin en çok olabilirlik tahmin edicisi ve $\hat{\theta}_1^{(i)}$, H_1 hipotezi altında $\theta^{(i)}$'nin en çok olabilirlik tahmin edicisi olmak üzere; genelleştirilmiş olabilirlik oranı Eş. 3.18'deki gibi elde edilir (Gray, Baek, Woodward, Miller ve Fisk, 1996).

$$\begin{aligned} LR &= \frac{sup_{\{\theta^{(1)},\theta^{(2)}|H_0\}}\{L_1(\theta^{(1)}|x)L(\theta^{(1)},\theta^{(2)}|x_1^{(1)},x_2^{(1)},\ldots x_{n_1}^{(1)},x_1^{(2)},x_2^{(2)},\ldots x_{n_2}^{(2)})\}}{sup_{\{\theta^{(1)},\theta^{(2)}|H_1\}}\{L_2(\theta^{(2)}|x)L(\theta^{(1)},\theta^{(2)}|x_1^{(1)},x_2^{(1)},\ldots x_{n_1}^{(1)},x_1^{(2)},x_2^{(2)},\ldots x_{n_2}^{(2)})\}} \\ &= \frac{L_1(\hat{\theta}_0^{(1)}|x)L(\hat{\theta}_0^{(1)},\hat{\theta}_0^{(2)}|x_1^{(1)},x_2^{(1)},\ldots x_{n_1}^{(1)},x_1^{(2)},x_2^{(2)},\ldots x_{n_2}^{(2)})}{L_2(\hat{\theta}_1^{(2)}|x)L(\hat{\theta}_1^{(1)},\hat{\theta}_1^{(2)}|x_1^{(1)},x_2^{(1)},\ldots x_{n_1}^{(1)},x_1^{(2)},x_2^{(2)},\ldots x_{n_2}^{(2)})} \end{aligned} \tag{3.18}$$

Eğer olabilirlik oranı (LR) 1'den büyük ($LR > 1$) ise hangi sınıfa ait olduğu bilinmeyen bir x gözlemi π_1 sınıfına, diğer durumlarda ise π_2 sınıfına atanmaktadır (Krzanowski, 1982).

3.3.2. İki sınıflı veri setlerinde diskriminant analizi

Fisher (1936), π_1 ve π_2 sınıflarından oluşan bir yığından gelen ve hangi sınıfa ait olduğu belli olmayan bir x gözlemini π_1 veya π_2 sınıflarından birine atamak için bir kural önermiştir. Fisher diskriminant kuralı için, gözlemler çok değişkenli normal dağılıma sahipken en iyi sonuç elde edilse bile Fisher diskriminant kuralı dağılımdan bağımsızdır, normallik varsayımını sağlamayan gözlemler için de kullanılmaktadır. Eğer, π_1 ve π_2 sınıflarının yığına ilişkin konum ve ölçek parametreleri biliniyorsa, atama kuralında bu parametre bilgileri kullanılmaktadır. Eğer, yığına ilişkin konum ve ölçek parametre bilgileri bilinmiyorsa, yığına ilişkin konum ve ölçek parametrelerinin tahmini için π_1 grubundan n_1 büyüklüğünde ve π_2 grubundan n_2 büyüklüğünde örnekler kullanılmaktadır. Fisher (1936), atama kuralı için gözlemlerin doğrusal kombinasyonunu kullanmayı önermiştir. Fisher'in yaklaşımında, sınıfları ayırma gücü en iyi olan doğrultu α iken, $d^F = \alpha' x$ doğrusal kombinasyon olmak üzere, d^F diskriminant fonksiyonu ile π_1 sınıfına atanan gözlemlerin ortalaması ($\bar{d_1^F}$) $\alpha'\mu_1$, π_2 sınıfına atanan gözlemlerin ortalaması ($\bar{d_2^F}$) ise $\alpha'\mu_2$ olur. Eğer, $\Sigma_1 = \Sigma_2 = \Sigma$ ise gözlemlerin varyansı her sınıf için $\alpha'\Sigma\alpha$ olur. Grupları ayırma gücü en iyi olan doğrultu α, Eş. 3.19'daki ifadenin maksimum yapılması ile elde edilir (Lachenbruch, 1975).

$$\lambda = \frac{(\alpha'\mu_1 - \alpha'\mu_2)^2}{\alpha'\Sigma\alpha} \tag{3.19}$$

Eş. 3.19'daki ifadenin maksimum yapılması için α'ya göre türevi alınarak sıfıra eşitlenir.

$$\begin{aligned} &\frac{d\lambda}{d\alpha} = \frac{2(\mu_1 - \mu_2)\alpha'\Sigma\alpha - 2\Sigma\alpha(\alpha'\mu_1 - \alpha'\mu_2)}{(\alpha'\Sigma\alpha)^2} = 0 \\ &\mu_1 - \mu_2 = \Sigma\alpha(\frac{\alpha'\mu_1 - \alpha'\mu_2}{\alpha'\Sigma\alpha}) \\ &\Sigma^{-1}(\mu_1 - \mu_2) = \alpha(\frac{\alpha'\mu_1 - \alpha'\mu_2}{\alpha'\Sigma\alpha}) \end{aligned} \tag{3.20}$$

Eş. 3.20'nin sonucuna göre α, $\Sigma^{-1}(\mu_1 - \mu_2)$ ile orantılıdır. Eğer yığına ilişkin konum ve ölçek parametreleri bilinmiyorsa, konum parametresi tahmin edicisi olarak örnek ortalaması vektörü ($\bar{x_1}$, $\bar{x_2}$ tahmin edicileri) ve ölçek parametresi tahmin edicisi olarak örnek varyans-kovaryans matrisi (S tahmin edicisi) kullanılmaktadır. Hangi sınıfta olduğu

bilinmeyen bir x gözleminin diskriminant fonksiyonu değeri olan $d^F = \alpha' x = (\bar{x_1} - \bar{x_2})' S^{-1} x$ değeri, $\bar{d_1^F} = (\bar{x_1} - \bar{x_2})' S^{-1} \bar{x_1}$ değerine daha yakınsa hangi sınıfta olduğu bilinmeyen x gözlemi π_1 sınıfına, $\bar{d_2^F} = (\bar{x_1} - \bar{x_2})' S^{-1} \bar{x_2}$ değerine daha yakınsa hangi sınıfta olduğu bilinmeyen x gözlemi π_2 sınıfına sınıflanmaktadır. Diskriminant fonksiyonu d^F için, birinci sınıftaki gözlemlerin diskriminant fonksiyonu değerlerinin orta noktası olan $\bar{d_1^F}$ ve ikinci sınıftaki gözlemlerin diskriminant fonksiyonu değerlerinin orta noktası olan $\bar{d_2^F}$ aralığının orta noktası Eş. 3.21'deki gibi elde edilmektedir.

$$\frac{(\bar{d_1^F} + \bar{d_2^F})}{2} = \frac{1}{2}(\bar{x_1} - \bar{x_2})' S^{-1} (\bar{x_1} + \bar{x_2}) \tag{3.21}$$

Eş. 3.21 sonucunda elde edilen ortalamalar arası farka göre; $\bar{d_1^F} > \bar{d_2^F}$ iken, $d^F > \frac{1}{2}(\bar{d_1^F} + \bar{d_2^F})$ için, $|d^F - \bar{d_1^F}| < |d^F - \bar{d_2^F}|$ ise, diskriminant fonksiyonu değeri d^F, birinci gruptaki gözlemlerin diskriminant değeri ortalaması $\bar{d_1^F}$'ya yakındır (Lachenbruch, 1975).

3.3.3. Çok sınıflı veri setlerinde diskriminant analizi

Çok sınıflı veri setleri için Fisher'in diskriminant kuralı, Eş. 3.10'da tanımlanan gruplar arası varyans-kovaryas matrisi ile (B_{p*p}) Eş. 3.11'de tanımlanan grup içi varyans-kovaryans matrisine (W_{p*p}) dayanmaktadır. α sınıfları ayırma gücü en iyi olan doğrultuyu göstermek üzere, çok sınıflı Fisher diskriminant fonksiyonunun ($d_i^F(x)$) elde edilmesi Eş. 3.22'de verilen varyans oranlarının (λ) en büyüklenmesi temeline dayanmaktadır (Lachenbruch, 1975).

$$\lambda = \frac{\alpha' B \alpha}{\alpha' W \alpha} \qquad \alpha \in R^p \qquad \alpha \neq 0 \tag{3.22}$$

Diskriminant kriteri olan λ'nın en büyük olması (maksimizasyon proleminin çözümü) için Eş. 3.22'de verilen varyans oranının $\alpha' ya$ göre birinci türevi 0'a eşitlenmektedir.

$$\frac{d\lambda}{d\alpha} = \frac{2[(B\alpha)(\alpha' W \alpha) - (\alpha' B \alpha)(W\alpha)]}{(\alpha' W \alpha)(\alpha' W \alpha)}$$

$$\frac{2B\alpha}{\alpha' W \alpha} - \frac{2\lambda W \alpha}{\alpha' W \alpha} = 0$$

$$\begin{aligned}
&\frac{2(B\alpha - \lambda W\alpha)}{\alpha' W\alpha} = 0 \\
&2(B\alpha - \lambda W\alpha) = 0 \\
&B\alpha - \lambda W\alpha = 0 \\
&(B - \lambda W)\alpha = 0 \\
&(W^{-1}B - \lambda I) = 0
\end{aligned} \tag{3.23}$$

Eş. 3.23'teki problemin çözümü $|W^{-1}B - \lambda I| = 0$ determinantının çözümü ile sağlanır. Ayırma gücü en iyi olan α doğrultusu λ özdeğerlerine karşılık gelen özvektörler olmaktadır. Buna göre i. Fisher diskriminant fonksiyonu Eş. 3.24'teki gibi elde edilmektedir (Lachenbruch, 1975).

$$d_i^F(x) = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p \tag{3.24}$$

Hangi sınıfta olduğu bilinmeyen bir x gözlemi için tüm sınıfların Fisher diskriminant değerleri hesaplanmakta ve hangi sınıfta olduğu bilinmeyen bu x gözlemi hangi sınıfın Fisher diskriminant değeri en küçük ise o sınıfa atanmaktadır (Filzmoser, Hron ve Temple, 2012).

3.3.4. Farklı diskriminant kuralları

Gözlemlerin sınıflara atanması diskriminant kurallarına dayanmaktadır. İki sınıflı ve çok sınıflı veri setleri için Fisher diskriminant kuralları Kısım 3.3.2 ve Kısım 3.3.3'te detaylı olarak anlatılmıştır. Fisher diskriminant kuralı dışında, bir veri setindeki sınıfları birbirinden ayırabilmek ve hangi sınıfa ait olduğu belli olmayan bir gözlemi sınıflardan birine atamak için çeşitli diskriminant kuralları geliştirilmiştir.

Bayesci diskriminant kuralı, Bayes sonsal olasılıklarının en büyük yapılması ile elde edilmektedir. $\pi_1, \pi_2, \ldots, \pi_l$ sınıflarından oluşan yığından alınmış n_1, n_2, $\ldots$, n_l örnek büyüklüğüne sahip n tane gözlemin p tane özelliği ölçülmüş olsun. π_i ile gösterilen sınıfların tüm özellikleri f_i olasılık yoğunluk fonksiyonu ile ifade edilmektedir. Gözlem hakkında herhangi bir bilgi bulunmadığı durumda, ilgili gözlemin i. sınıfa atanma

olasılığı $p_i = n_i/n$ şeklinde elde edilmektedir ve bu olasılık değerine önsel olasılık denilmektedir. Eğer gözlemlerin sınıflara atanmasında kullanılacak ek bir bilgi varsa, gözlemler gözlenen herhangi bir bağımsız x değişken vektörünün, π_i sınıfından gelme olasılığı hesaplanarak, ilgili gözlem için Bayes kuralı ile elde edilen sonsal olasılık $(P(\pi_i/x))$ değerleri arasında en yüksek sonsal olasılığa sahip olduğu sınıfa atanmaktadır (Filzmoser ve diğerleri, 2012).

Sonsal olasılıkları hesaplamak için, açıklayıcı değişken vektörü x, kitleye ilişkin ortalama vektörü μ_i, kitleye ilişkin varyans-kovaryans matrisi Σ_i olmak üzere, Eş. 3.25'te verilen çok değişkenli normal dağılım fonksiyonundan yararlanılır.

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)/2} \qquad i = 1,\ldots,l \tag{3.25}$$

x yeni bir gözlem, $f_i(x)$ çok değişkenli normal dağılım fonksiyonu, p_i ise π_i sınıfının önsel olasılığı olmak üzere, x gözlemi için gruplara ilişkin sonsal olasılıklar Eş. 3.26'daki gibi hesaplanmaktadır.

$$P(\pi_i/x) = \frac{P(\pi_i \cap x)}{P(x)} = \frac{p_i f_i(x)}{\sum_{i=1}^{k} p_i f_i(x)} \qquad i = 1,\ldots,l \tag{3.26}$$

Veri setindeki gözlemlerin normal dağılımdan geldiği, ancak, gözlemlerin dahil oldukları sınıfların varyans-kovaryans matrislerinin farklı olduğu durumlarda karesel diskriminant fonksiyonu kullanılmaktadır. Karesel diskriminant fonksiyonu $(d_i^Q(x))$ Eş. 3.27'deki formül ile elde edilmektedir (Filzmoser ve diğerleri, 2012).

$$d_i^Q(x) = -\frac{1}{2}ln[det(\Sigma_i)] - \frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) + ln(\frac{n_i}{n}) \tag{3.27}$$

Eğer gözlemlerin dahil olduğu sınıfların varyans-kovaryans matrisleri birbirine eşit ise $(\Sigma_1 = \ldots = \Sigma_k = \Sigma)$ karesel diskriminant fonksiyonu $(d_i^Q(x))$, doğrusal diskriminant fonksiyonuna $(d_i^L(x))$ dönüşmektedir.

$$d_i^L(x) = \mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + ln(\frac{n_i}{n}) \tag{3.28}$$

Doğrusal diskriminant fonksiyonu ($d_i^L(x)$) Eş. 3.28'deki gibi elde edilmektedir (Filzmoser ve diğerleri, 2012).

3.3.5. Diskriminant analizinin başarısı

Diskriminant analizi sonuçlarının başarısı farklı yollarla ölçülebilmektedir. Bu yaklaşımlardan birisi, gözlemlerin gerçekte ait oldukları sınıf ile diskriminant analizi sonucu atandıkları sınıfın bir çapraz tablo olan sınıflandırma tablosu ile incelenmesidir. Sınıflandırma tablosunu oluşturmak için iki farklı yaklaşımdan yararlanılabilir. Bu yaklaşımlardan ilkinde, veri seti, eğitim veri seti ve deney veri seti şeklinde tesadüfi olarak iki parçaya ayrılmaktadır. Daha sonra, eğitim veri seti için diskriminant fonksiyonları elde edilmektedir ve deney veri setinde yer alan gözlemlere ilişkin özelliklerin ölçüm değerleri elde edilen diskriminant fonksiyonlarında yerine konularak hangi sınıfa atanacağı tahmin edilmektedir. Deney veri setinde yer alan gözlemlerin gerçekte ait olduğu sınıflar ile diskriminant analizi uygulanması sonucu atandıkları sınıfların uyumu sınıflandırma tablosu aracılığı ile incelenmektedir. İkinci yaklaşımda ise, veri setinde yer alan gözlemlerden birisi dışarıda bırakılarak kalan $n-1$ gözlemden elde edilen diskriminant fonksiyonları yardımı ile hesaplamaya katılmayan gözlemin hangi sınıfa ait olduğu tahmin edilmektedir. Bu işlem tüm gözlemler için yapılmaktadır ve gözlemlerin gerçekte ait olduğu sınıflar ile diskriminant analizi uygulanması sonucu atandıkları sınıfların uyumu sınıflandırma tablosu aracılığı ile incelenmektedir. Çizelge 3.1'de verilen sınıflandırma tablosunun köşegen elemanları doğru sınıfa atanan gözlemlerin sıklığını, köşegen dışı elemanlar ise yanlış sınıfa atanan gözlemlerin sıklığını göstermektedir (Alpar, 2013). Sınıflandırma tablosunun köşegen dışı elemanlarının toplamının genel toplama bölünmesi ile *hatalı sınıflandırma oranı* (*HSO*) (*apparent error rate-AER*) $(f_{12}+f_{1l}+f_{21}+f_{2l}+f_{l1}+f_{l2}+\ldots/n)$ elde edilmektedir. Tüm gözlemlerin köşegen elemanları üzerinde toplanması ve hatalı sınıflandırma oranının 0 olması diskriminant analizi tahmininin çok başarılı olduğunu göstermektedir. Bu oranın 0'a eşit olması uygulamalarda sık rastlanan bir durum değildir. Bu oranın sıfıra yaklaşması istenen bir durumdur ve *hatalı sınıflandırma oranı* sıfıra ne kadar yakınsa diskriminant analizinin kestiriminin o kadar başarılı olduğunu göstermektedir.

Çizelge 3.1. Sınıflandırma tablosu

Gerçekte Olan \ Tahmin Edilen	1	2	...	l
1	f_{11}	f_{12}	...	f_{1l}
2	f_{21}	f_{22}	...	f_{2l}
⋮	⋮	⋮	⋮	⋮
l	f_{l1}	f_{l2}	...	f_{ll}

Diskriminant analizinin başarısı için hatalı sınıflandırma oranının minimize edilmesi gerekmektedir. R_1, π_1 sınıfı için sınıflandırma bölgesi ve R_2, π_2 sınıfı için sınıflandırma bölgesi iken, yığın içerisinde π_1 sınıfının önsel olasılığı p_1, π_2 sınıfının önsel olasılığı p_2 $(= 1 - p_1)$ ve π_1'den gelen gözlemler için yoğunluk fonksiyonu $f_1(x)$, π_2'den gelen gözlemler için yoğunluk fonksiyonu $f_2(x)$ olmak üzere, hangi sınıfta olduğu bilinmeyen x gözlemi R_1'in bazı bölgelerinde ise x gözlemi π_1 sınıfına, R_2'nin bir bölgesindeyse π_2'ye atanmaktadır. R_1 ve R_2 ayrışık kümelerinin R uzayının tamamını oluşturduğu varsayıldığında *hatalı sınıflandırma oranı* Eş. 3.29'daki gibi elde edilmektedir (Lachenbruch, 1975).

$$\begin{aligned} HSO &= p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx \\ &= p_1 + \int_{R_1} [p_2 f_2(x) - p_1 f_1(x)]dx \end{aligned} \tag{3.29}$$

Veri seti çok değişkenli normal dağılıma sahipken, kitleye ilişkin konum ve ölçek parametreleri bilindiğinde elde edilen diskriminant fonksiyonu $d_i^F(x)$ iken, $d_i^F(x)$ kullanılarak yapılan atama sonucu π_1 sınıfındaki gözlemlerin beklenen değeri $E(d_i^F(x)/\pi_1)$, Eş. 3.30'da verildiği gibidir.

$$\begin{aligned} E(d_i^F(x)/\pi_1) &= [\mu_1 - \frac{1}{2}(\mu_1 + \mu_2)]' \Sigma^{-1}(\mu_1 - \mu_2) \\ &= \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) = \frac{1}{2}\delta^2 \end{aligned} \tag{3.30}$$

π_2 sınıfındaki gözlemlerin beklenen değeri $E(d_i^F(x)/\pi_2)$, Eş. 3.31'de verildiği gibidir.

$$E(d_i^F(x)/\pi_2) = [\mu_2 - \frac{1}{2}(\mu_1 + \mu_2)]\Sigma^{-1}(\mu_1 - \mu_2) = -\frac{1}{2}\delta^2 \tag{3.31}$$

Her sınıfın varyansı ise Eş. 3.32'de verildiği gibidir.

$$E[(x-\mu_i)'\Sigma^{-1}(\mu_1-\mu_2)]^2 = E[(\mu_1-\mu_2)'\Sigma^{-1}(x-\mu_i)(x-\mu_i)'\Sigma^{-1}(\mu_1-\mu_2)] \\ = (\mu_1-\mu_2)'\Sigma^{-1}E(x-\mu_i)(x-\mu_i)'\Sigma^{-1}(\mu_1-\mu_2) = (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) = \delta^2 \tag{3.32}$$

Eş. 3.30, Eş. 3.31 ve Eş. 3.32'de verilen beklenen değer ve varyanslar yardımı ile, birinci sınıfa ilişkin hatalı sınıflandırma olasılığı HSO_1 Eş. 3.33'teki gibi hesaplanmaktadır.

$$HSO_1 = Pr\left[\frac{d_i^F(x) - \frac{1}{2}\delta^2}{\delta} < \frac{ln\frac{1-p_1}{p_1} - \frac{\delta^2}{2}}{\delta}\right] = \Phi\left(\frac{ln\frac{1-p_1}{p_1} - \frac{\delta^2}{2}}{\delta}\right) \tag{3.33}$$

Benzer şekilde, ikinci sınıfa ilişkin hatalı sınıflandırma olasılığı HSO_2 Eş. 3.34'teki gibi hesaplanabilmektedir (Lachenbruch, 1975).

$$HSO_2 = \Phi\left(-\frac{ln\frac{1-p_1}{p_1} + \frac{\delta^2}{2}}{\delta}\right) \tag{3.34}$$

Çok sınıflı veri setlerinde hatalı sınıflandırma olasılığı ise Eş. 3.35'te verildiği gibi hesaplanabilmektedir (Lachenbruch, 1975).

$$HSO = 1 - \int_{-\infty}^{\infty}\left[\Phi\left(\frac{\delta^2/2 - (\alpha' x)}{\sqrt{\delta^2/2}}\right)\right]^{l-1}\phi(\alpha' x)d(\alpha' x) \tag{3.35}$$

4. BAŞKA DAĞILIMDAN KARIŞAN GÖZLEMLER

İstatistiksel analizlerde, veri setini oluşturan gözlemlerin birbirinden bağımsız olduğu ve aynı dağılımlardan geldiği varsayılmaktadır. Gerçek veri setlerinde bazı gözlemler, verinin merkezinden çok uzakta yer alabilir veya çalışılan veri setine başka bir dağılımdan gözlemler karışmış olabilir. Başka dağılımdan karışmış gözlemler genel varsayımlardan ve/veya verinin çoğunluğunun geldiği dağılımdan sapan veri noktaları olarak tanımlanabilmektedir. Başka dağılımdan karışmış gözlemlerin çok fazla gözlemli ve/veya çok değişkenli veri setlerinde görülme olasılığı yüksektir ve genellikle görsel inceleme ile ortaya çıkmamaktadır. Klasik çok değişkenli analiz teknikleri (temel bileşenler analizi, diskriminant analizi ve çok değişkenli regresyon gibi) örnek ortalaması, örnek varyansı, örnek kovaryansı, örnek korelasyonu istatistiklerine dayanmaktadır. Tüm bu istatistikler verideki birkaç gözlemin başka bir dağılımdan karışmış olmasından bile güçlü bir şekilde etkilenebilmektedir. Veri ciddi miktarda başka dağılımdan karışan gözlem içerdiğinde, eğer veride böyle gözlemler yokmuş gibi tahmin yapılırsa elde edilen tahminler gerçek tahminlerden çok uzakta gerçekleşmektedir (Hubert ve diğerleri, 2008). Veride yer alan başka dağılımdan karışan gözlemler, görsel inceleme ile tespit edilemeyebilirler. Çok değişkenli dağılımlarda her bir gözlemin merkeze olan uzaklığı m_i^2 olmak üzere, veriye başka dağılımdan karışan gözlem tespiti çeşitli yollarla yapılabilmektedir. Bu yollardan iki tanesi aşağıdaki gibi özetlenmektedir.

İlk olarak, p değişken sayısı, m_i^2 değeri $(x_i - \bar{x})' S^{-1} (x_i - \bar{x})$ eşitliği ile elde edilen matrisin köşegen elemanları ve $0,005$ gibi küçük bir anlamlılık düzeyi için p serbestlik dereceli ki-kare ($\chi^2_{p;(0,005)}$) değerinden büyük m_i^2 değerine sahip gözlemler ($m_i^2 > \chi^2_{p;(0,005)}$) başka dağılımdan karışmış gözlem olarak belirlenmektedir (Alpar, 2013).

İkinci olarak, başka dağılımdan karışan gözlemleri saptayabilmek için Mahalanobis uzaklığından yararlanılabilir. Mahalanobis uzaklığı çok değişkenli veri setinde yer alan her bir gözlemin, verinin merkezinden uzaklığının bir ölçüsüdür. $\hat{\mu}$ konum ve $\hat{\Sigma}$ ölçek parametresi tahmin edicileri olmak üzere, her x_i gözleminin Mahalanobis uzaklığı Eş. 4.1'de verildiği gibi hesaplanmaktadır.

$$MD(x_i) = \sqrt{(x_i - \hat{\mu})' \hat{\Sigma}^{-1} (x_i - \hat{\mu})} \tag{4.1}$$

Mahalanobis uzaklığı (MD) her bir x_i gözleminin yoğunlaşmanın merkezinden ne kadar uzakta olduğunu göstermektedir. Bir gözlemin mahalanobis uzaklığı değeri p serbestlik dereceli ki-kare değerinin karekökünden büyük ($MD(x_i) > \sqrt{\chi^2_{p;0,975}}$) ise o gözlem, başka dağılımdan karışan gözlem olarak belirlenmektedir (Hubert ve diğerleri, 2008).

Daha önce de değinildiği gibi, çok değişkenli istatistiksel analiz yöntemleri veri setine başka dağılımdan karışan gözlemlere karşı çok duyarlıdır. Başka dağılımdan karışan gözlemler içeren veri setleri ile çalışırken güvenilir bir analiz yapılabilmesi için bu gözlemlerin etkisinden kurtulmak gerekmektedir. Tez çalışmasında, çok değişkenli analiz yöntemlerinden diskriminant analizi ve temel bileşenler analizi üzerinde çalışılmıştır.

Diskriminant analizinde, veri setine başka bir dağılımdan karışma olması durumunda, klasik yaklaşımla hesaplanan konum ve ölçek parametrelerini kullanmak yerine, sağlamlık özelliğine sahip sağlam tahmin edicilerle çalışmak hatalı sınıflandırma oranını azaltmaktadır. Örneğin, varyans-kovaryans matrisinin tahmin edicisi olarak yüksek kırılma noktasına sahip sağlam tahmin edici olarak tanımlanan en küçük determinantlı varyans-kovaryans matrisi (minimum covariance determinant (MCD)) tahmin edicisi kullanılmaktadır.

Temel bileşenler analizi de, diskriminant analizinde olduğu gibi varyans-kovaryans matrisine dayalı olarak çalışmaktadır. Daha önce de değinildiği gibi varyans-kovaryans matrisi başka dağılımdan karışan gözlemlere karşı oldukça duyarlıdır. Başka dağılımdan karışan gözlemler temel bileşenleri kendi doğrultularına çektiğinden var olandan daha fazla varyans açıklama oranlarına sebep olabilirler.

5. SAĞLAMLIK

Kimi zaman doğrudan kimi zaman dolaylı olarak kullandığımız örnek ortalaması, örnek varyansı, örnek kovaryansı ya da örnek korelasyonu gibi istatistikler ya da tahmin ediciler, veriye başka dağılımdan gözlem karışmış olması durumundan çok etkilenmektedir. Yani, bahsedilen bu tahmin ediciler başka dağılımdan karışmış gözlemlere karşı çok duyarlıdır. Veri setine başka bir dağılımdan küçük bir karışma olması bile bu tahmin edicilerin değerini gerçek değerlerden çok uzaklaştırabilmektedir. Ve bu gibi durumlar, veri setinden elde edilen tahmin değerleri ile verinin gerçekte ait olduğu yığın değerleri arasında uyum sağlamayan hatalı sonuçlar elde edilmesine yol açabilmektedir (Hubert ve diğerleri, 2008).

Veri setinde başka bir dağılımdan karışma olması durumunda, veri setindeki gözlemlerin geldiği dağılımla iyi uyum gösteren ve veri setinde yer alan karışmış gözlemlerin etkisini azaltarak karışmış gözlem yokmuş gibi tahminler yapmayı sağlayan "*sağlam tahmin ediciler*" kullanılır. Sağlam tahmin ediciler, veri setindeki bu tip gözlemlerin etkisini azaltarak başka dağılımdan karışmış gözlem olmadığında elde edilen tahminlere benzer tahmin sonuçları vermektedir. Sağlam istatistikler ya da sağlam tahmin ediciler kullanmanın amacı, veri setinde başka bir dağılımdan karışma olması ihtimaline karşı bu gözlemlerin etkisinden arındırılmış sağlam tahminler geliştirmektir.

Klasik istatistksel yöntemlerin alternatifi olan sağlam istatistiksel yöntemlerin performansını çalışmak için karışma modelleri kullanılmaktadır. Çok iyi bilinen ve yaygın olarak kullanılan karışma modelleri ilk olarak Tukey (1960) tarafından önerilmiş ve daha sonra Huber (1964) tarafından geliştirilmiştir. $F_0(x)$ ve $F_k(x)$ birbirinden farklı dağılımlar olmak üzere, bir veri setindeki gözlemlerin çoğu ($1-\varepsilon$ kadarı) $F_0(x)$ dağılımından gelirken bu veri setindeki gözlemlerin çok küçük bir oranı (ε kadarı) $F_k(x)$ dağılımından gelebilir. $F_k(x)$ dağılımından gelen bu az sayıdaki gözleme karışan gözlem denilmektedir. ε, küçük bir karışma oranı olmak üzere, klasik karışma modeli Eş. 5.1'de verildiği gibidir (Maronna ve diğerleri, 2006).

$$F_\varepsilon(x) = (1-\varepsilon)F_0(x) + \varepsilon F_k(x) \tag{5.1}$$

5.1. Sağlamlık Ölçüleri

Tahmin edicilerin sağlamlık özelliği kırılma noktası değeri (breakdown point, KN), etki fonksiyonu (influence function, IF) ve duyarlılık eğrisi (sensitivity curve, DE) ölçüleri ile tespit edilmektedir.

5.1.1. Kırılma noktası değeri

Sağlamlık ölçülerinden birisi kırılma noktası değeridir. Tahmin edicideki değişim sınırının aşılmasını sağlayan en yüksek aykırı gözlem oranı kırılma noktası değerini vermektedir. Bir tahmin edicinin kırılma noktası değeri $KN_n^*(T,x)$, tahmin edici üzerinde çok büyük etkiye sahip olabilmektedir. Veri setindeki gözlemlerin herhangi m tanesinin başka bir dağılımdan geldiği karışmalı veri seti x_k olmak üzere, kırılma noktası değeri Eş. 5.2'deki gibi elde edilmektedir.

$$KN_n^*(T,x) = \min_m \{\frac{m}{n}; sup_{x_k} \|T(x_k) - T(x)\| = \infty\} \tag{5.2}$$

Bir tahmin edicinin kırılma noktası değeri aykırı gözlemlerin en küçük $\frac{m}{n}$ oranıdır ve bu kırılma noktası değeri bütün sınırlar üzerinden tahmin edilebilmektedir (Hubert ve diğerleri, 2008).

Sağlamlık özelliğine sahip bir tahmin edicinin kırılma noktası değeri en fazla $0,50$'dir. Yani, veri setindeki gözlemlerin $0,50$'sinden daha azının başka dağılımdan gelen gözlemlerden karışmış gözlemler olması durumunda dahi sağlam istatistikler sınırlı kalmaya, değerini değiştirmeden kalmaya devam eder. Örneğin, elimizde 10, 11, 12, 13 ve 14 ölçüm değerlerine sahip bir örnek olsun. Bu ölçümlerin örnek ortalaması ve örnek medyanı 12'dir. Ancak, son değerin yanlışlıkla 14 değil 50 olarak kaydedildiği varsayılsın. Buna göre, hesaplanan örnek ortalaması 19,20 olarak bulunur. Yani, örnek ortalaması tek bir karışan gözlemin varlığında bile hatalı sonuç vermekte, gerçek değerinden çok uzakta gerçekleşmektedir. Öyleyse, örnek ortalaması istatistiğinin kırılma noktası değeri 0'dır ve veride aykırı gözlemler olduğunda güvenilir olmayan sonuçlar vermektedir. Aynı şekilde, son değerin 14 değil 50 olduğunu varsayarak örnek medyanı hesaplandığında 12 olduğu

görülmektedir. Yani, bir tane karışan gözlem medyanı kıramamıştır. Medyan istatistiği verinin yarısına kadar olan karışmalardan etkilenmemektedir. Medyanın kırılma noktası değeri olabilecek en yüksek kırılma noktası değeri olan $0,50$'dir. Veride karışan gözlem bulunması durumunda medyan istatistiğinin örnek ortalaması istatistiğine göre daha güvenilir sonuçlar verdiği söylenebilmektedir.

5.1.2. Duyarlılık eğrisi ve etki fonksiyonu

Sağlamlığın diğer ölçüleri ise duyarlılık eğrisi (DE) ve etki fonksiyonudur (IF). Karışan gözlem dahil edilerek hesaplanan tahmin edici değeri ile karışan gözlem dahil edilmeden hesaplanan tahmin edicinin değeri arasındaki fark duyarlılık eğrisi ile ifade edilmektedir. Duyarlılık eğrisi, karışan gözlem değerinin bir fonksiyonu olarak çizilebilmektedir. $x_1, \ldots, x_n$ örneğine yeni bir x_0 gözlemi eklendiğinde karışma oranı $\varepsilon = 1/(n+1)$'dir. T bir tahmin edici ve her x_0 için $DE_n(x_0)$ bir rastgele değişken olmak üzere, standartlaştırılmış duyarlılık eğrisi Eş. 5.3'teki gibi tanımlanmaktadır.

$$\begin{aligned} DE_n(x_0) &= \frac{T_{n+1}(x_1,x_2,...,x_n,x_0) - T_n(x_1,...,x_n)}{1/(n+1)} \\ &= (n+1)(T_{n+1}(x_1,x_2,...,x_n,x_0) - T_n(x_1,...,x_n)) \end{aligned} \tag{5.3}$$

Eğer x_i'lerin dağılımı F dağılımı ile benzer ise büyük gözlemler için $\lim_{n\to\infty} DE_n(x_0) \approx IF(x_0,F)$ olur (Maronna ve diğerleri, 2006).

Bir tahmin edicinin etki fonksiyonu onun duyarlılık eğrisinin asimptotik versiyonudur (Hampel, 1974). Etki fonksiyonu, T(F) gibi gerçek değerli fonksiyonellerin ölçülemeyecek kadar küçük davranışlarını araştırmak amacı ile Hampel (1974) tarafından keşfedilmiştir. Etki fonksiyonu, tek bir gözlemin tahmin edici üzerindeki etkisini göstermektedir. Sağlam tahmin ediciler sınırlı etki fonksiyonuna sahiptir. Her tahmin ediciye ilişkin bir etki fonksiyonu vardır.

$$IF(F,T,x_0) = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \frac{\Big[T[F+\varepsilon(\Delta x_0 - F) - T(F)]\Big]}{\varepsilon} \tag{5.4}$$

F dağılımına x_0 aykırı gözleminin T tahmin edicisi üzerinde etkisini gösteren etki fonksiyonu Eş. 5.4'teki gibi ifade edilmektedir (Maronna ve diğerleri, 2006).

5.2. Sağlamlık Özelliğine Sahip Tahmin Ediciler

Kısım 2 ve Kısım 3'de bahsedildiği gibi diskriminant analizi ve temel bileşenler analizinin uygulanması, ilgili değişkenlerin ortalama ve varyans-kovaryans matrislerine dayanmaktadır. Veri setine başka bir dağılımdan gözlem karışması durumunda, klasik tahmin edicilerden çok sağlamlık özelliğine sahip tahmin ediciler tercih edilmektedir (Hubert ve diğerleri, 2008).

Bu kısımda, diskriminant analizi ve temel bileşenler analizinin yapı taşlarını oluşturan tahmin ediciler için sağlamlık özelliğine sahip varyans-kovaryans matrisi ve yine sağlamlık özelliğine sahip hızlı sonuç veren varyans-kovaryans matrisinin tahmin edilmesi ve izdüşüm arama tahmin yöntemleri açıklanmaya çalışılmıştır.

5.2.1. MCD algoritması

Sağlam çok değişkenli tahmin edicilerden biri Rousseeuw (1984) tarafından önerilen en küçük kovaryans determinant (minimum covariance determinant-*MCD*) tahmin edicisidir. En küçük kovaryans determinant tahmin edicisi çok değişkenli konum ve ölçek parametrelerinin oldukça sağlam bir tahmin edicisidir.

Konum ve ölçek parametrelerinin sağlam tahmin edicileri, determinant değeri en küçük olan varyans-kovaryans matrisinin belirlenmesi ile elde edilmektedir.

MCD yöntemi, n tane gözlem içerisinden varyans-kovaryans matrisinin determinantını en küçük yapacak olan h tane gözlemi araştırmaktadır. n tane gözlem içinden seçilecek gözlem sayısı (h), $h = [\frac{n+p+1}{2}] \approx \frac{n}{2}$ ya da $h \approx 0,75n$ ifadeleri ile belirlenmektedir.

MCD yönteminde, tüm mümkün alt kümeler ($\binom{n}{h}$ tane) belirlenmekte ve tüm mümkün alt kümeler için varyans-kovaryans matrisleri ve onların determinantları hesaplanmaktadır. Tüm mümkün alt kümeler içerisinden varyans-kovaryans matrisinin determinantı en küçük

olan altküme belirlenmektedir.

Varyans-kovaryans matrisinin determinantı en küçük olan h çaplı gözlem grubu belirlendikten sonra konum ve ölçek parametrelerinin tahmin edicileri, belirlenen h gözlem yardımı ile hesaplanmaktadır.

En küçük determinant değerine sahip olan varyans-kovaryans matrisine sahip h gözlem yardımı ile başlangıç için bir konum parametresi tahmini ($\hat{\mu}_0$) ve ölçek parametresi tahmini ($\hat{\Sigma}_0$) elde edilmektedir.

Her gözlem için sağlamlık özelliğine sahip olan bir uzaklık tanımlanmaktadır. Bu uzaklık, $\hat{\mu}_0$ merkezine ve $\hat{\Sigma}_0$ ölçeğine göre her nokta için hesaplanmaktadır. Gözlemler μ_0 merkezine uzaklıkları bakımından ikiye ayrılmaktadır. Sağlamlık özelliğine sahip uzaklık *RD* (robust distance) Eş. 5.5'teki gibi hesaplanmaktadır.

$$\begin{aligned} RD_{x_i}(x_i, \hat{\mu}_0, \hat{\Sigma}_0) &= \sqrt{(x_i - \hat{\mu}_0)' \hat{\Sigma}_0^{-1} (x_i - \hat{\mu}_0)} \leq \sqrt{\chi^2_{p,0,975}} \\ RD_{x_i}(x_i, \hat{\mu}_0, \hat{\Sigma}_0) &= \sqrt{(x_i - \hat{\mu}_0)' \hat{\Sigma}_0^{-1} (x_i - \hat{\mu}_0)} > \sqrt{\chi^2_{p,0,975}} \end{aligned} \tag{5.5}$$

Veri merkezine uzaklığı $\sqrt{\chi^2_{p,0,975}}$'den küçük olan gözlemlere ($RD_{x_i} \leq \sqrt{\chi^2_{p,0,975}}$) $w_i = 1$ ağırlığı atanmakta, diğer gözlemlere ($RD_{x_i} > \sqrt{\chi^2_{p,0,975}}$) $w_i = 0$ ağırlığı atanmaktadır.

Ağırlıklı konum ($\hat{\mu}_w$) parametresi tahmini Eş. 5.6'daki gibi hesaplanmaktadır.

$$\hat{\mu}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{5.6}$$

Ağırlıklı ölçek parametresi tahmini ($\hat{\Sigma}_w$) Eş. 5.7'deki gibi hesaplanmaktadır.

$$\hat{\Sigma}_w = \frac{\sum_{i=1}^{n} w_i (x_i - \hat{\mu}_w)(x_i - \hat{\mu}_w)'}{\sum_{i=1}^{n} w_i - 1} \tag{5.7}$$

Sağlamlık özelliğine sahip uzaklıklar $\hat{\mu}_w$ merkezi ve $\hat{\Sigma}_w$ ölçeğine göre tekrar

belirlenmektedir.

Elipsin merkezi sabit kalıncaya kadar aynı işlemler tekarlanmaktadır. Konum ve ölçek parametresi tahminleri aynı kalmaya devam ettiğinde işlem durdurulmaktadır. İterasyon işlemleri sonucu elde edilen tahmin ediciler *MCD* tahmin edicileri olarak adlandırılır (Rousseeuw ve Hubert, 2011).

5.2.2. FAST-MCD algoritması

MCD algoritması, tüm mümkün alt kümelerin varyans-kovaryans matrislerini hesaplamaya dayalı olduğundan, çok büyük gözlemli veri setlerinde hesaplama zorlukları yaşanmaktadır. Bu nedenle Rousseeuw ve Van Driessen (1999) daha kısa sürede güvenilir sonuçlar veren %50 kırılma noktası değerine sahip bir sağlam tahmin edici önermişlerdir.

$X = \{x_1, x_2, ..., x_n\}$ n gözlemli rastgele bir örnek olsun. $H_1 \subset \{1, ...n\}$ X'den alınan $|H_1| = h$ boyutlu alt küme olsun.

Seçilen h boyutlu altkümenin ortalaması $\bar{x}_1$, Eş. 5.8'deki gibi hesaplanmaktadır.

$$\bar{x}_1 = \frac{1}{h} \sum_{i \in H_1} x_i \tag{5.8}$$

Seçilen h boyutlu altkümenin varyansı S_1, Eş. 5.9'daki gibi hesaplanmaktadır.

$$S_1 = \frac{1}{h} \sum_{i \in H_1} (x_i - \bar{x}_1)(x_i - \bar{x}_1)' \tag{5.9}$$

Burada hesaplanan varyans-kovaryans matrisinin determinantı sıfıra eşit değilse $(det(S_1) \neq 0)$ $i = 1, \ldots, n$ için göreli uzaklıklar Eş. 5.10'daki gibi hesaplanmaktadır.

$$d_1(i) = \sqrt{(x_i - \bar{x}_1) S_1^{-1} (x_i - \bar{x}_1)'} \tag{5.10}$$

$(d_1)_{1:n} \leq (d_1)_{2:n} \leq \ldots \leq (d_1)_{n:n}$ sıralı uzaklıkları elde edilmekte ve en küçük uzaklığa

sahip gözlemlerle $H_2 = (d_1)_{1:n}, \ldots, (d_1)_{h:n}$ altkümesi oluşturulmaktadır. Daha sonra H_2 altkümesine ilişkin ortalama matrisi $\bar{x}_2$ ve varyans-kovaryans matrisi S_2 hesaplanmaktadır. Buradan, $det(S_2) \leq det(S_1)$ elde edilmektedir. Bu değer en küçük uzaklık değerli h tane gözleme ve daha küçük determinanta sahip olduğu için S_2'ye S_1'e göre daha fazla yoğunlaşmaktadır. Bu işlemlere *C-adımı* denilmektedir.

Rousseeuw ve Van Driessen (1999) tarafından önerilen *C-adımı* algoritması şöyle çalışmaktadır:

1. $h = \frac{n+p+1}{2}$, H_{eski} rastgele seçilir.
2. $\bar{x}_{eski}$, S_{eski} ve $d_{eski}(i)$ hesaplanır.
3. $d_{eski}(i) = \sqrt{(x_i - \bar{x}_{eski})S_{eski}^{-1}(x_i - \bar{x}_{eski})'}$
4. s permütasyon olmak üzere, hesaplanan uzaklıklar sıralanır.
5. $d_{eski}(s(1)) \leq d_{eski}(s(2)) \leq \ldots \leq d_{eski}(s(n))$
6. $H_{yeni} = \{s(1), s(2), ..., s(h)\}$ en küçük h gözlem için oluşturulur.
7. $\bar{x}_{yeni}$, S_{yeni}, $d_{yeni}(i)$ hesaplanır.
8. $det(S_{yeni}) = 0$ veya $det(S_{yeni}) = det(S_{eski})$ olana kadar iterasyon işlemleri devam eder.

Bu yolla elde edilen $det(S_{yeni})$ sonuç değeri *MCD* tahmin edicisi için en iyi sonuç olmayabilir. Buna rağmen, yaklaşık bir *MCD* çözümü, H_1 in bazı ilk seçenekleri alınarak, her biri için *C-adımı* uygulanarak ve en küçük determinantlı çözüm alınarak elde edilebilmektedir.

Başlangıç H_1 altkümesini oluşturmak için, rastgele $(p+1)$ gözlem alınarak j alt kümesi oluşturulmakta ve $\bar{x}_j$ ve S_j hesaplanmaktadır. Eğer $det(S_j) = 0$ ise $det(S_j) > 0$ olana dek gözlem eklenerek j genişletilebilmektedir. Daha sonra, $i = 1, \ldots, n$ için $d_j^2(i) = (x_i - \bar{x}_j)S_j^{-1}(x_i - \bar{x}_j)'$ hesaplanıp sıralanmakta ve en küçük h tanesi için $H_1 = s(1), \ldots, s(h)$ oluşturulmaktadır.

Bu prosedür gözlem sayısı küçük olan örnekler için çok hızlıdır, ancak gözlem sayısı arttıkça hesaplama zamanı artmaktadır. Çünkü her *C-adımında* tüm gözlemlerin sağlam uzaklıklarının hesaplanması gerekmektedir.

5.2.3. İzdüşüm arama (projection pursuit) algoritması

İzdüşüm Arama (Projeksiyon pursuit-PP) çok boyutlu verideki en dikkat çekici mümkün izdüşümleri bulmak için kullanılan istatistiksel bir tekniktir. İzdüşüm arama yöntemi; yüksek boyutlu veri setini en iyi açıklayan düşük boyutlu (1, 2 veya 3) izdüşümleri bulmak için kullanılmaktadır (Friedman, 1987). *PP* tekniği, Kruskal (1969) tarafından önerilmiş ve denenmiştir. İlk başarılı uygulama Tukey ve Friedman (1974) tarafından yapılmıştır (Huber, 1985).

Klasik temel bileşenler analizi ve diskriminant analizi, sağlam olmayan kovaryans veya korelasyon tahmininin özvektör ve özdeğerlerinden hesaplandığından karışan gözlemlere karşı çok duyarlıdır. Karışan gözlemlerden etkilenmeyen sonuçlar elde etmenin bir yolu, varyans-kovaryans matrisinin sağlam tahminine dayalı özdeğerler ve özvektörler hesaplamaktır. Diğer bir yolu ise, varyans-kovaryans matrisinin sağlam tahminini dikkate almayarak doğrudan özdeğerler ve özvektörlerin sağlam tahminlerini hesaplamaktır. $\hat{\mu}$ konum ve $\hat{\Sigma}$ ölçeğin tek değişkenli tahmin edicileri ve $\alpha' x_i$, α üzerine izdüşürülen gözlemlerin tek değişkenli örneğidir. Eğer $\hat{\mu}$ ve $\hat{\Sigma}$ sırası ile örnek ortalaması istatistiği ve örnek standart sapması istatistiği olarak alınırsa diskriminant analizi ve temel bileşenler analizi için klasik tahmin sonuçları elde edilir. $\hat{\mu}$ ve $\hat{\Sigma}$ için etkin ve sağlam tahmin ediciler kullanılır ise sağlamlık özelliğine sahip sonuçlar elde edilir. Sağlamlık özelliğine sahip konum parametresinin tahmin edicisi olarak spatial medyan gibi farklı çok değişkenli medyanlar kullanılabilir. Bu çalışmada; % 50 kırılma noktası değerine sahip, her bir gözlemin konum parametresinden uzaklıklarının toplamını en küçük yapan konum parametresi olan L_1 medyan istatistiği kullanılabilmektedir. L_1 medyan istatistiği Eş. 5.11'deki gibi elde edilmektedir.

$$L_1(X) = \underset{\mu \varepsilon R^p}{argmin} \sum_{i=1}^{n} ||x_i - \mu|| \tag{5.11}$$

Kullanılabilecek iyi bilinen bir sağlam ölçek parametresi ise medyandan mutlak sapmaların medyanıdır (median absolute deviation, *MAD*). $y_i = x_i - L_1$ iken, merkezileştirilmiş $\{y_1, \ldots, y_n\} \subset R$ örneği için *MAD* istatistiği Eş. 5.12'deki gibi tanımlanır.

$$MAD_n(y_1,...,y_n) = 1,486 \underset{i}{med}|y_i - \underset{j}{med}y_j| \tag{5.12}$$

İzdüşüm arama yöntemi ile temel bileşenler analizi

Temel bileşenler analizinde, izdüşüm arama (projection pursuit, PP) yöntemi verinin izdüşürüldüğü en büyük yayılıma sahip doğrultuları (α) araştırmaktadır. Özdeğer ve özvektörler tamamıyla ölçek parametresine bağlı olarak elde edilmektedir. Bu amaçla, izdüşüm arama uygulamasında sağlam ölçek tahmin edicisi olarak Eş. 5.12'de verilen $S_n = MAD_n(y_1,...,y_n)$ sağlam ölçek tahmin edicisi kullanılmaktadır $x_1, \ldots, x_n \in R^p$ gözlemlerinin bir dizisi için birinci özvektör ($t_{S_n,1}$) Eş. 5.13'teki gibi tanımlanmaktadır.

$$t_{S_n,1} = argmax S_n(\alpha' x_1, \ldots, \alpha' x_n) \tag{5.13}$$

Birinci özvektöre ilişkin özdeğer $\lambda_{S_n,1} = ((t_{S_n,1})' x_1, \ldots, (t_{S_n,1})' x_n)$ şeklinde tanımlanır.

Croux ve Gazen (2005) tarafından önerilen sağlam temel bileşenler analizi algoritmasında, temel bileşenler, sağlam bir ölçek tahminini en büyükleyen doğrultular (α) üzerinde verinin izdüşümleri olarak tanımlanmaktadır. $L_1(X)$, $X = x_1, \ldots, x_n$ örneğinden hesaplanan konum tahmin edicisi L_1 medyan, önemli temel bileşen sayısı k ($1 \leq k \leq p$) ve PP algoritması için izdüşüm endeksi S_n ölçek tahmin edicisi olmak üzere, Croux ve Gazen'in (2005) temel bileşenler analizi için geliştirdiği algoritma adımları aşağıdaki gibi olmaktadır.

$r = 1,\ldots k$ olmak üzere $r = 1$ için; x_i ($i = 1,\ldots, n$) gözlemleri $x_i^1 = x_i - L_1(X)$ dönüşüm matrisi ile merkezileştirilir. Daha sonra, tüm mümkün doğrultuların matrisi $A_{n,1}(X)$, x_i^1 dönüşüm matrisinin satırlarının normalleştirilmesi ile Eş. 5.14'teki gibi elde edilir.

$$A_{n,1}(X) = \{\frac{x_i^1}{||x_i^1||}; 1 \leq i \leq n\} \tag{5.14}$$

Mümkün izdüşüm doğrultuları üzerine izdüşürülen gözlemlerin oluşturduğu örnekler ($\alpha' x_i$) yardımı ile birinci özvektör Eş. 5.15'teki gibi hesaplanır.

$$\hat{t}_{S_{n,1}} = \underset{\alpha \varepsilon A_{n,1}(X)}{\operatorname{argmax}} S_n(\alpha' x_1^1, ..., \alpha' x_1^n) \tag{5.15}$$

Birinci özvektör yardımı ile, ilk temel bileşen skorları $y_i^1 = \hat{t}'_{S_{n,1}} x_i^1$ olarak elde edilir. Her $r = 2, \ldots, k$ için algoritma adımları tekrar tekrar çalışmaktadır.

Öncelikle, her $i = 1, \ldots, n$ için $x_i^r = x_i^{r-1} - y_i^{r-1} \hat{t}_{S_n, r-1}$, $A_{n,r}(X) = \{\frac{x_i^r}{||x_i^r||}; 1 \leq i \leq n\}$ hesaplanır. Daha sonra tahmin edilen özvektör $\hat{t}_{S_n,r} = \underset{\alpha \varepsilon A_{n,r}(X)}{\operatorname{argmax}} S_n(\alpha' x_1^r, \ldots, \alpha' x_n^r)$ elde edilir. Son olarak, her $i = 1, \ldots, n$ için $y_i^r = \hat{t}'_{S_n,r} x_i$ hesaplanır. Böylelikle, r. temel bileşen $(y_1^r, \ldots, y_n^r)'$ üzerindeki skorlar elde edilmektedir.

İzdüşüm arama yöntemi ile diskriminant analizi

Pires ve Branco (2010) izdüşüm arama (*PP*) yönteminin 2 sınıflı veri setlerinde diskriminant analizine uygulanışı üzerinde çalışmışlardır. p tane özelliği (değişken) ölçülmüş, birinci ve ikinci sınıftan (gruptan) gelen n_1 ve n_2 büyüklüğündeki örnekler için, iki sınıf arasındaki diskriminant doğrultusunu karakterize eden p boyutlu vektörün izdüşüm arama tahmini ($\hat{\alpha}$), Eş. 5.16'da verilen diskriminant kriterinin (λ) en büyüklenmesini sağlayan izdüşüm doğrultusu (α) olmaktadır (Pires,2003).

$$\lambda = \frac{[\hat{\mu}(\alpha' x_1) - \hat{\mu}(\alpha' x_2)]^2}{a_1 \hat{\Sigma}^2(\alpha' x_1) + a_2 \hat{\Sigma}^2(\alpha' x_2)} \tag{5.16}$$

Eş. 5.16'da, $\|\alpha{=}1\|$ ve $\hat{\mu}(\alpha' x_1) > \hat{\mu}(\alpha' x_2)$'dir. a_1 ve a_2 sabitleri ise, $a_i = \frac{n_i}{n_1+n_2}$ şeklinde elde edilmektedir. $\hat{\mu}$ ve $\hat{\Sigma}$ sırası ile konum ve ölçeğin tek değişkenli tahmin edicileri ve $\alpha' x_i$, α izdüşüm doğrultusu üzerine izdüşürülmüş i. sınıftaki gözlemlerin tek değişkenli örneğidir ($\alpha' x_{i1}, \ldots, \alpha' x_{in_i}$). Eğer konum tahmin edicisi $\hat{\mu}$ için örnek ortalaması istatistiği ve ölçek tahmin edicisi $\hat{\Sigma}$ için örnek standart sapması istatistiği kullanılırsa klasik diskriminant analizi çözümü ($\alpha \propto S^{-1}(\bar{x_1} - \bar{x_2})$) elde edilir. Ancak, konum tahmin edicisi $\hat{\mu}$ için L_1 medyan istatistiği ve ölçek tahmin edicisi $\hat{\Sigma}$ için *MAD* istatistiği kullanılırsa izdüşüm arama tahmini ($\hat{\alpha}$) için sağlam diskriminant analizi çözümü Eş. 5.17'deki gibi elde edilir.

$$\alpha \propto MAD^{-1}(L_1(x_1) - L_1(x_2)) \tag{5.17}$$

Fisher (1936) sınıflandırma için kesme veya ayırma noktası olarak, izdüşüm doğrultusu (α) üzerine izdüşürülmüş gözlemlerin ($\alpha' x_i$) arasındaki orta noktayı ($\hat{\alpha}_0$) kullanmaktadır. $\hat{\mu}$ konum tahmin edicisi olmak üzere izdüşüm doğrultusu (α) üzerine izdüşürülmüş gözlemlerin ($\alpha' x_i$) arasındaki orta nokta Eş. 5.18'deki gibi elde edilmektedir.

$$\hat{\alpha}_0 = [\hat{\mu}(\hat{\alpha}' x_1) + \hat{\mu}(\hat{\alpha}' x_2)]/2 \tag{5.18}$$

Hangi sınıfa ait olduğu bilinmeyen yeni bir x gözlemi, $\hat{\alpha}' x > \hat{\alpha}_0$ ise birinci sınıfa, diğer durumlarda ikinci sınıfa sınıflanmaktadır.

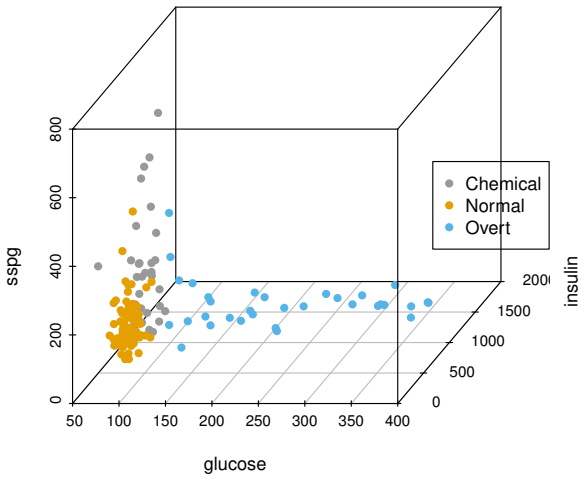
6. UYGULAMA

Sağlam tahmin yöntemleri ile diskriminant analizi ve temel bileşenler analizi uygulamaları ile klasik tahmin yöntemleri ile diskriminant analizi ve temel bileşenler analizi uygulamalarının karşılaştırılması amacı ile, uygulamalarda sıklıkla kullanılan *Diabetes*, *Pottery* ve *Milk* veri setleri ile uygulamalar yapılmıştır. Uygulamalar, R versiyon 3.2.2 programı kullanılarak yapılmıştır.

6.1. Diabetes Veri Seti İle Diskriminant Analizi

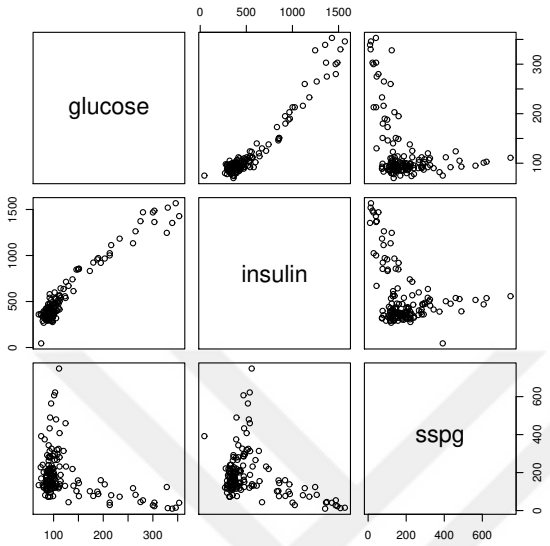
Klasik tahmin ediciler ve MCD tahmin edicileri ile diskriminant analizi sonuçlarının karşılaştırılması için, Reaven ve Miller (1979) tarafından yayınlanan *Diabetes* veri seti kullanılmıştır (R programı mclust paketi). *Diabetes* veri setinde, 145 adet obez olmayan hasta, *Glucose: Oral glikoz tolerans testinden (OGTT) 3 saat sonra plazma glikozu eğrisi altında kalan alan*, *Insulin: Oral glikoz tolerans testinden (OGTT) 3 saat sonra plazma insülin eğrisi altında kalan alan* ve *SSPG: Kararlı durum plazma glikozu* ölçümlerine göre, *Normal*, *Overt (Gizli Olmayan)* ve *Chemical (Kimyasal)* olarak 3 sınıfa (gruba) ayrılmıştır.

Veri setinin görsel incelemesi için Şekil 6.1'de verilen üç boyutlu dağılım grafiği ve Şekil 6.2'de verilen değişkenler arası ilişki grafiği çizilmiştir.



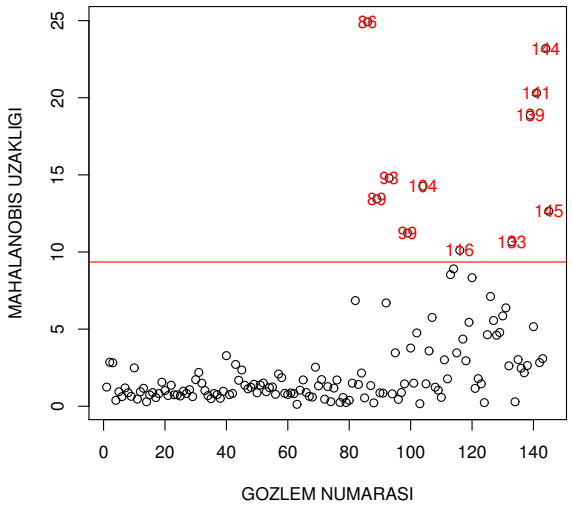
Şekil 6.1. Diabetes veri seti üç boyutlu dağılım grafiği

Şekil 6.1'de verilen üç boyutlu dağılım grafiğine bakıldığında, *Normal* ve *Chemical* sınıfında bulunan gözlemlerin aynı doğrultuda dağıldığı buna karşın *Overt* gözlemlerin onların aksi bir doğrultuda dağıldığı görülmektedir.



Şekil 6.2. Diabetes veri seti değişkenler arası ilişki grafiği

Şekil 6.2'de verilen değişkenler arası ilişki grafiğine göre, glucose ve insulin değişkenleri arasında doğrusal bir ilişki olduğu söylenebilmektedir. Veri setinde başka dağılımdan karışmış gözlem bulunup bulunmadığını tespit etmek için her bir gözleme ait hesaplanan *Mahalanobis Uzaklıkları* Şekil 6.3'te ve Çizelge 6.1'de yer almaktadır.



Şekil 6.3. Diabetes veri seti mahalanobis uzaklıkları grafiği

Çizelge 6.1. Diabetes veri seti mahalanobis uzaklıkları

Gözlem No	1	2	3	4	5	6	7	8	9	10
MD	1,24	2,86	2,82	0,38	0,94	0,62	1,19	0,88	0,64	2,49
Gözlem No	11	12	13	14	15	16	17	18	19	20
MD	0,46	0,94	1,16	0,29	0,74	0,88	0,56	0,82	1,56	1,02
Gözlem No	21	22	23	24	25	26	27	28	29	30
MD	0,70	1,36	0,74	0,74	0,66	0,98	0,83	1,07	0,62	1,72
Gözlem No	31	32	33	34	35	36	37	38	39	40
MD	2,19	1,50	1,01	0,69	0,49	0,83	0,75	0,51	0,97	3,28
Gözlem No	41	42	43	44	45	46	47	48	49	50
MD	0,75	0,83	2,71	1,68	2,35	1,36	1,13	1,22	1,42	0,87
Gözlem No	51	52	53	54	55	56	57	58	59	60
MD	1,37	1,50	0,94	1,20	1,25	0,77	2,09	1,85	0,84	0,76
Gözlem No	61	62	63	64	65	66	67	68	69	70
MD	0,86	0,82	0,12	1,04	1,71	0,89	0,64	0,59	2,53	1,32
Gözlem No	71	72	73	74	75	76	77	78	79	80
MD	1,73	0,46	1,27	0,30	1,18	1,70	0,25	0,56	0,24	0,38
Gözlem No	81	82	83	84	85	86	87	88	89	90
MD	1,50	6,85	1,41	2,15	0,54	24,92	1,34	0,22	13,44	0,87
Gözlem No	91	92	93	94	95	96	97	98	99	100
MD	0,85	6,70	14,79	0,80	3,46	0,45	0,89	1,45	11,23	3,77
Gözlem No	101	102	103	104	105	106	107	108	109	110
MD	1,50	4,75	0,16	14,27	1,45	3,59	5,76	1,24	1,04	0,57
Gözlem No	111	112	113	114	115	116	117	118	119	120
MD	3,01	1,78	8,53	8,91	3,46	10,11	4,35	2,95	5,44	8,34
Gözlem No	121	122	123	124	125	126	127	128	129	130
MD	1,16	1,79	1,46	0,23	4,64	7,12	5,56	4,60	4,78	5,86
Gözlem No	131	132	133	134	135	136	137	138	139	140
MD	6,38	2,62	10,67	0,29	3,02	2,47	2,17	2,64	18,87	5,16
Gözlem No	141	142	143	144	145					
MD	20,31	2,83	3,09	23,19	12,64					

Şekil 6.3'te ve Çizelge 6.1'de yer alan *Mahalanobis Uzaklıkları* değerlerine göre *Mahalanobis Uzaklığı* $\chi^2_{3;0,975} = 9,35$ değerinden büyük olan 86., 89., 93., 99., 104., 116., 133., 139., 141., 144. ve 145. gözlemlerin başka dağılımdan karışan gözlemler olduğu söylenebilmektedir.

6.1.1. Diabetes veri seti klasik tahmin ediciler ile diskriminant analizi

Başka dağılımdan karışan gözlem içeren veri setine klasik tahmin edicilerle diskriminant analizi uygulanmıştır. Klasik tahmin edicilerle hesaplanan diskriminant fonksiyonları aşağıdaki gibi elde edilmiştir.

$d_{Chemical} = -13,15 - 0,01 * glucose + 0,03 * insulin + 0,04 * sspg$

$d_{Normal} = -6,49 + 0,07 * glucose + 0,01 * insulin + 0,02 * sspg$

$d_{Overt} = -28,02 - 0,03 * glucose + 0,05 * insulin + 0,03 * sspg$

Başka dağılımdan karışan gözlemler dahil edilerek klasik tahmin edicilerle hesaplanan diskriminant fonksiyonları ile gözlemlerin hangi sınıfa ait olduğu tespit edilmeye çalışılmış ve sınıflandırma tablosu ile hata matrisi (confusion matrix) sırası ile Çizelge 6.2 ve Çizelge 6.3'te verilmiştir.

Çizelge 6.2. Diabetes veri seti klasik diskriminant analizi sınıflandırma tablosu

Tahmin Edilen / Gerçekte Olan	Chemical	Normal	Overt
Chemical	26	10	0
Normal	2	74	0
Overt	5	2	26

Klasik tahmin edicilerle diskriminant analizinden elde edilen sınıflandırma tablosuna (Çizelge 6.2) göre, gerçekte *chemical* sınıfında yer alan 26 gözlem yine *chemical* olarak tahmin edilmesine karşın gerçekte *chemical* sınıfında yer alan 10 gözlem *normal* olarak tahmin edilmiştir. Gerçekte *normal* sınıfında yer alan 74 gözlem yine *normal* olarak tahmin edilmiş olmasına karşın gerçekte *normal* sınıfında yer alan 2 gözlem *chemical* olarak tahmin edilmiştir. Gerçekte *overt* sınıfında yer alan 26 gözlem yine *overt* olarak tahmin edilmiş olmasına karşın gerçekte *overt* sınıfında yer alan 5 gözlem *chemical* ve 2 gözlem de *normal* olarak tahmin edilmiştir.

Çizelge 6.3. Diabetes veri seti klasik diskriminant analizi hata matrisi

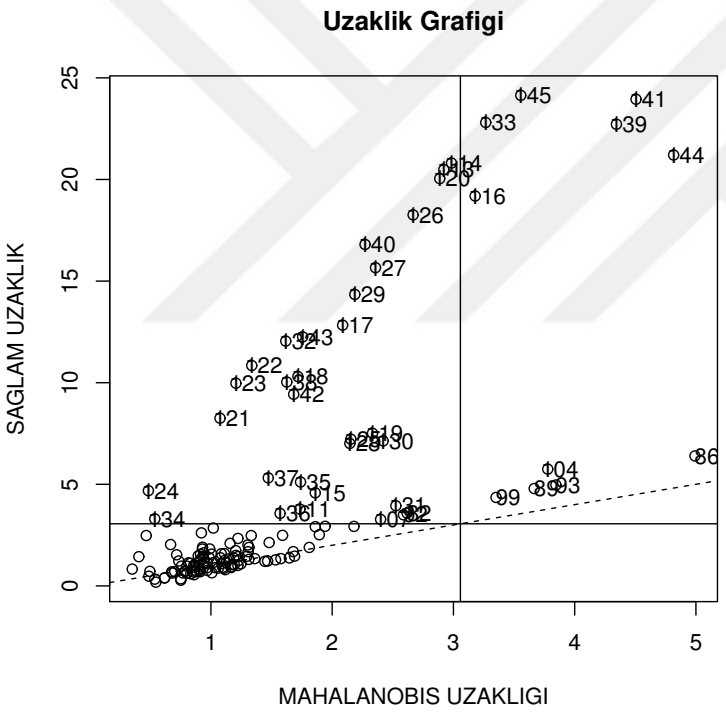
Tahmin Edilen / Gerçekte Olan	Chemical	Normal	Overt
Chemical	0,72	0,28	0,00
Normal	0,03	0,97	0,00
Overt	0,15	0,06	0,79

Çizelge 6.3'te gösterilen hata matrisine göre de *chemical* sınıfında yer alan gözlemlerin $0,72$'si, *normal* sınıfında yer alan gözlemlerin $0,97$'si ve *overt* sınıfında yer alan gözlemlerin $0,79$'u doğru sınıflanmıştır. Ancak hatalı sınıflara atanmış çok fazla gözlem

bulunmaktadır. Buna göre *Hatalı Sınıflandırma Oranı (Apparent error rate)* (Yanlış sınıfa atanan gözlem sayısı/toplam gözlem sayısı) $0,13$ olarak hesaplanmıştır.

6.1.2. Diabetes veri seti *MCD* tahmin edicileri ile diskriminant analizi

Hatalı sınıflama oranını düşürmek için veri setine sağlamlık özelliğine sahip *MCD* tahmin edicileri ile diskriminant analizi uygulanmıştır. Kısım 4.'te tanımlanan klasik tahmin edicilerden elde edilen *Mahalanobis Uzaklıkları* ile Kısım 5.2.1.'de tanımlanan *MCD* tahmin edicilerinden elde edilen *Sağlam (Robust) Uzaklıkların* karşılaştırmalı grafiği Şekil 6.4'te verilmiştir.

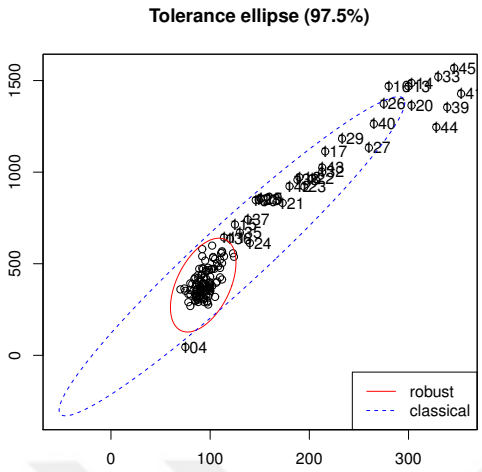


Şekil 6.4. Diabetes veri seti mahalanobis uzaklık-sağlam uzaklık grafiği

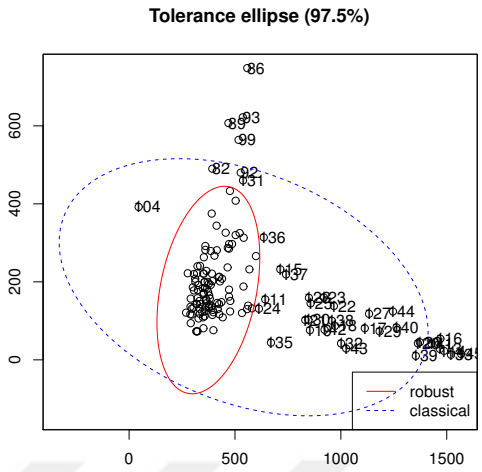
Şekil 6.4'te verilen grafiğe göre, mahalanobis uzaklığı dışında kalan gözlemlerin sayısının sağlam uzaklıklar dışında kalan gözlemlerden daha az olduğu görülebilmektedir.

Kısım 4.'te tanımlanan klasik tahmin edicilerden elde edilen *Mahalanobis Uzaklıkları* ile Kısım 5.2.1.'de tanımlanan *MCD* tahmin edicilerinden elde edilen *Sağlam (Robust) Uzaklıklar* dışında kalan gözlemleri, %97,5 güven düzeyinde gösteren tolerans elipsoidleri çizdirilmiştir. Tolerans elipsoidleri, *glucose* ve *insulin* değişkenleri için Şekil 6.5'te ve *insulin*

ve *SSPG* değişkenleri için Şekil 6.6'da gösterilmiştir.



Şekil 6.5. Glucose-insulin tolerans elipsoid



Şekil 6.6. İnsulin-SSPG tolerans elipsoid

Tolerans elipsoidlerine göre, mahalanobis uzaklıkları dışında kalan gözlemlerin sağlam (robust) uzaklıklar dışında kalan gözlemlerden daha az olduğu görülmektedir. Buradan, sağlam uzaklıkların başka dağılımdan karışmış gözlemleri ayırmada daha başarılı olduğu söylenebilmektedir.

Hatalı sınıflama oranını düşürmek için veri setine sağlamlık özelliğine sahip *MCD* tahmin edicileri ile diskriminant analizi uygulanmıştır. Elde edilen diskriminant fonksiyonları kullanılarak, sınıflandırma tablosu (Çizelge 6.4) ve hata matrisi (confusion matrix) (Çizelge 6.5) yardımı ile hatalı sınıflandırma oranı hesaplanmıştır.

Çizelge 6.4. Diabetes veri seti sağlam (*MCD*) diskriminant analizi sınıflandırma tablosu

Gerçekte Olan \ Tahmin Edilen	Chemical	Normal	Overt
Chemical	32	4	0
Normal	2	74	0
Overt	9	0	24

Sağlamlık özelliğine sahip *MCD* tahmin edicileri ile diskriminant analizinden elde edilen sınıflandırma tablosuna (Çizelge 6.4) göre, gerçekte *chemical* sınıfında yer alan 32 gözlem yine *chemical* olarak tahmin edilmesine karşın gerçekte *chemical* sınıfında yer alan 4 gözlem *normal* olarak tahmin edilmiştir. Gerçekte *normal* sınıfında yer alan 74 gözlem yine *normal*

olarak tahmin edilmiş olmasına karşın gerçekte *normal* sınıfında yer alan 2 gözlem *chemical* olarak tahmin edilmiştir. Gerçekte *overt* sınıfında yer alan 24 gözlem yine *overt* olarak tahmin edilmiş olmasına karşın gerçekte *overt* sınıfında yer alan 9 gözlem *chemical* olarak tahmin edilmiştir.

Çizelge 6.5. Diabetes veri seti sağlam (*MCD*) diskriminant analizi hata matrisi

Gerçekte Olan \ Tahmin Edilen	Chemical	Normal	Overt
Chemical	0,89	0,11	0,00
Normal	0,00	0,97	0,00
Overt	0,27	0,00	0,73

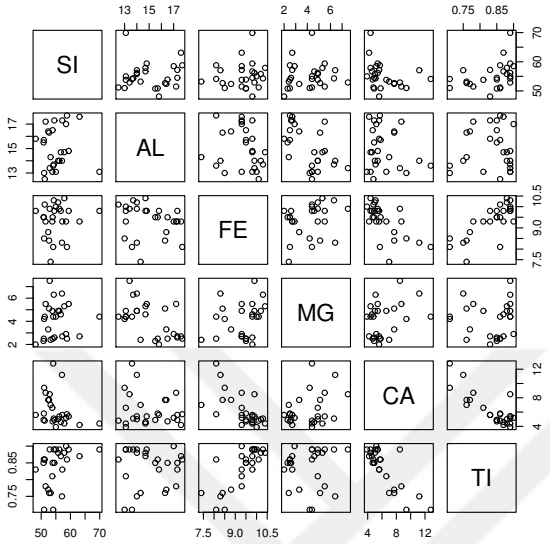
Hata matrisine göre (Çizelge 6.5) *chemical* sınıfında yer alan gözlemlerin $0,89$'u, *normal* sınıfında yer alan gözlemlerin $0,97$'si, *overt* sınıfında yer alan gözlemlerin $0,73$'ü doğru sınıflanmıştır. Buna göre *Hatalı Sınıflandırma Oranı (Apparent error rate)* $0,10$ olarak hesaplanmıştır.

İçerisinde başka dağılımdan karışmış gözlem bulunan *Diabetes* veri setine klasik tahmin edicilerle ve *MCD* tahmin edicileri ile diskriminant analizi uygulanmıştır. Klasik tahmin edicilerle diskriminant analizi ile yapılan sınıflamada hatalı sınıflandırma oranı $0,13$, *MCD* tahmin edicileri ile diskriminant analizi ile yapılan sınıflamada hatalı sınıflandırma oranı $0,10$ olarak bulunmuştur. Bu bulgulara göre, böyle gözlemler varlığında sağlam tahmin edicilerle diskriminant analizi uygulamalarının daha doğru sonuçlar verdiği söylenebilir.

6.2. Pottery Veri Seti İle Diskriminant Analizi

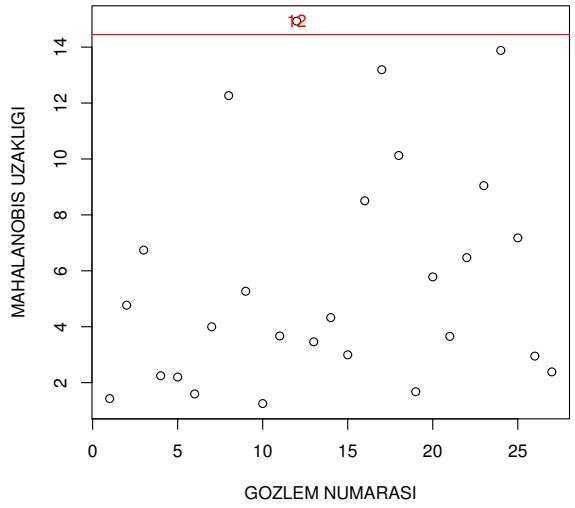
İlk kez Stern ve Descoeudres (1977) tarafından yayınlanan ve Pires ve Branco (2010) tarafından da kullanılan *Pottery (Ceramic)* veri seti için klasik tahmin ediciler ile , MCD tahmin edicileri ile, izdüşüm arama (*PP*) yöntemi ile ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi (PP_{MAD}) ile diskriminant analizi uygulanmış ve sonuçlar karşılaştırılmıştır. *Pottery (Ceramic)* veri setinde, 27 adet Yunan çömleği parçası incelenmiştir. 6 adet metalik oksit bileşen (Si, Al, Fe, Mg, Ca, Ti) değerleri ölçülmüş ve ölçüm sonuçlarına göre gözlemler "Attic" ve "Eritrean" köken olarak iki sınıfa ayrılmıştır. Veri setinin görsel incelemesi için değişkenler arası ilişki grafiği çizdirilmiştir.

Şekil 6.7'de yer alan değişkenler arası ilişki grafiğine bakıldığında CA ve TI değişkenleri arasında doğrusal bir ilişki olduğu, FE ve TI değişkenleri arasında da doğrusala yakın bir ilişki olduğu söylenebilmektedir.



Şekil 6.7. Pottery veri seti değişkenler arası ilişki grafiği

Veride başka dağılımdan karışmış gözlem olup olmadığının tespiti için *Mahalanobis Uzaklıklarından* yararlanılmıştır. Hesaplanan *Mahalanobis Uzaklıkları* Şekil 6.8'de ve Çizelge 6.6'da yer almaktadır.



Şekil 6.8. Pottery veri seti mahalanobis uzaklıkları grafiği

Çizelge 6.6. Pottery veri seti mahalanobis uzaklıkları

Gözlem No	1	2	3	4	5	6	7
MD	1,43	4,77	6,74	2,25	2,20	1,60	4,00
Gözlem No	8	9	10	11	12	13	14
MD	12,27	5,27	1,25	3,67	14,93	3,46	4,33
Gözlem No	15	16	17	18	19	20	21
MD	3,00	8,50	13,19	10,12	1,67	5,78	3,66
Gözlem No	22	23	24	25	26	27	
MD	6,47	9,05	13,88	7,18	2,95	2,39	

Şekil 6.8'de ve Çizelge 6.6'da verilen *Mahalanobis Uzaklığı* değerlerine göre, *Mahalanobis Uzaklığı* $\chi^2_{6;0,975} = 14,45$ değerinden büyük olan 12. gözlemin başka dağılımdan karışmış gözlem olduğu söylenebilmektedir.

6.2.1. Pottery veri seti klasik tahmin ediciler ile diskriminant analizi

Pottery veri setine klasik tahmin edicilerle diskriminant analizi uygulanmıştır. *Pottery* veri seti için klasik tahmin edicilerle hesaplanan diskriminant fonksiyonları aşağıdaki gibi elde edilmiştir.

$$d_{Attic} = -853,33 + (4,55 * \mathbf{SI}) - (2,60 * \mathbf{AL}) + (40,67 * \mathbf{FE}) - (9,83 * \mathbf{MG})$$
$$+(34,47 * \mathbf{CA}) + (1079,99 * \mathbf{TI})$$
$$d_{Eritrean} = -796,97 + (4,22 * \mathbf{SI}) + (0,23 * \mathbf{AL}) + (41,72 * \mathbf{FE}) - (13,16 * \mathbf{MG})$$
$$+(34,81 * \mathbf{CA}) + (986,31 * \mathbf{TI})$$

Başka dağılımdan karışmış gözlem dahil edilerek klasik tahmin edicilerle hesaplanan diskriminant fonksiyonları ile gözlemlerin hangi sınıfa ait olduğu tespit edilmeye çalışılmış ve sınıflandırma tablosu ile hata matrisi (confusion matrix) sırası ile Çizelge 6.7 ve Çizelge 6.8'de verilmiştir.

Çizelge 6.7. Pottery veri seti klasik diskriminant analizi sınıflandırma tablosu

Tahmin Edilen / Gerçekte Olan	Attic	Eritrean
Attic	12	1
Eritrean	0	14

Klasik tahmin edicilerle elde edilen sınıflandırma tablosuna (Çielge 6.7) göre, gerçekte *Attic* sınıfında yer alan 12 gözlem yine *Attic* sınıfında tahmin edilmesine karşın gerçekte *Attic* sınıfında yer alan 1 gözlem *Eritrean* sınıfında tahmin edilmiştir. Gerçekte *Eritrean* sınıfında yer alan 14 gözlem yine *Eritrean* sınıfında tahmin edilmiştir.

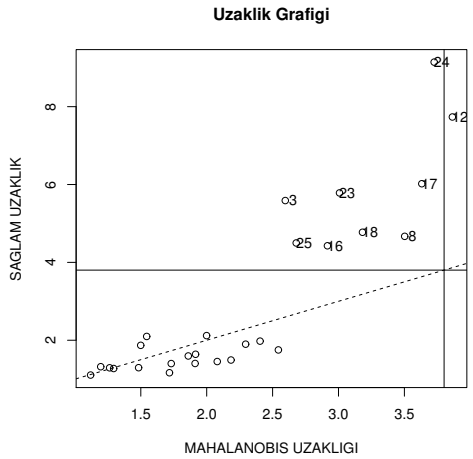
Çizelge 6.8. Pottery veri seti klasik diskriminant analizi hata matrisi

Tahmin Edilen / Gerçekte Olan	Attic	Eritrean
Attic	0,92	0,08
Eritrean	0,00	1,00

Çizelge 6.8'de verilen hata matrisine göre *Attic* sınıfında yer alan gözlemlerin %92'si ve *Eritrean* sınıfında yer alan gözlemlerin tamamı doğru sınıflanmıştır. Buna göre, *Hatalı Sınıflandırma Oranı (Apparent error rate)* $0,04$ olarak hesaplanmıştır. Hatalı sınıflandırma oranını düşürmek için, sağlamlık özelliğine sahip *MCD* tahmin edicileri, izdüşüm arama yöntemi ve *MAD* istatistiği kullanılarak izdüşüm arama yöntemi kullanılabilmektedir.

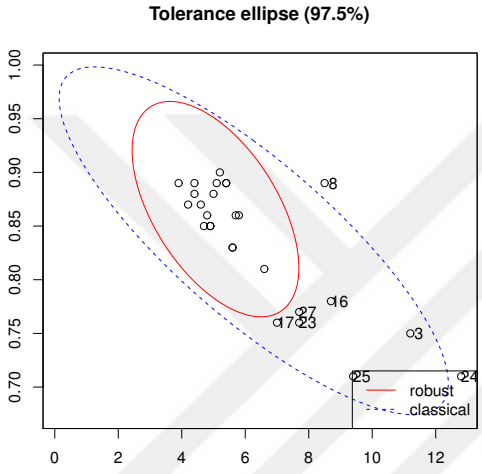
6.2.2. Pottery veri seti için MCD tahmin edicileri ile diskriminant analizi

Hatalı sınıflandırma oranını düşürmek amacı ile, veri setine sağlamlık özelliğine sahip *MCD* tahmin edicileri ile diskriminant analizi uygulanmıştır.

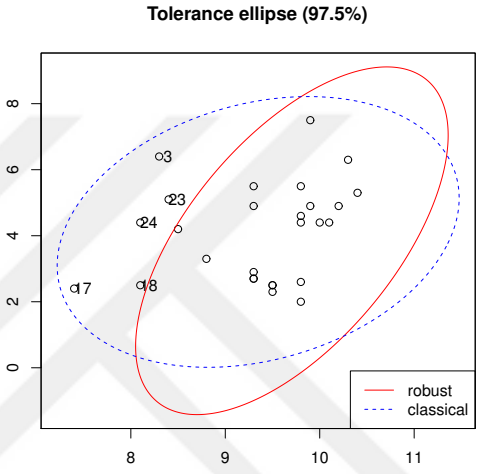


Şekil 6.9. Pottery veri seti mahalanobis uzaklık-sağlam uzaklık grafiği

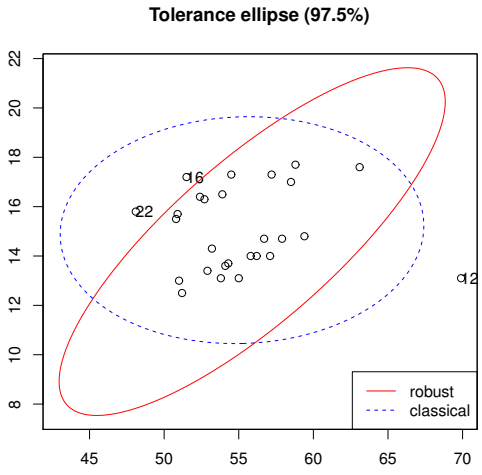
Şekil 6.9'da verilen uzaklık grafiğine göre, mahalanobis uzaklığı kesme değerinden büyük olan tek gözlemin 12 numaralı gözlem olduğu, buna karşın sağlam uzaklık kesme değerinin dışında kalan çok sayıda gözlem olduğu görülmektedir. Mahalanobis Uzaklıkları ile *MCD* tahmin edicilerinden elde edilen Sağlam (Robust) Uzaklıklar dışında kalan gözlemleri, %97,5 güven düzeyinde gösteren tolerans elipsoidleri çizdirilmiştir. Tolerans elipsoidleri, *CA* ve *TI* değişkenleri için Şekil 6.10'da, *FE* ve *MG* değişkenleri için Şekil 6.11'de ve *SI* ve *AL* değişkenleri için Şekil 6.12'de gösterilmiştir.



Şekil 6.10. CA ve TI tolerans elipsoidi



Şekil 6.11. FE ve MG tolerans elipsoidi



Şekil 6.12. SI ve AL tolerans elipsoidi

CA ve *TI* değişkenleri için Şekil 6.10'da, *FE* ve *MG* değişkenleri için Şekil 6.11'de ve *SI* ve *AL* değişkenleri için Şekil 6.12'de verilen tolerans elipsoidlerine göre, mahalanobis

uzaklıkları dışında kalan gözlemlerin sağlam (robust) uzaklıklar dışında kalan gözlemlerden daha az olduğu görülmektedir. Buradan, sağlam uzaklıkların başka dağılımdan karışmış gözlemleri ayırmada daha titiz olduğu söylenebilmektedir.

MCD tahmin edicileri ile hesaplanan diskriminant fonksiyonları aşağıdaki gibi elde edilmiştir.

$$d_{Attic} = -1655,69 + (65,71 * \mathbf{SI}) - (198,10 * \mathbf{AL}) + (325,63 * \mathbf{FE}) - (3,25 * \mathbf{MG})$$
$$+(68,70 * \mathbf{CA}) - (1388,74 * \mathbf{TI})$$
$$d_{Eritrean} = -994,27 + (43,21 * \mathbf{SI}) - (122,26 * \mathbf{AL}) + (218,78 * \mathbf{FE}) - (9,83 * \mathbf{MG})$$
$$+(53,73 * \mathbf{CA}) - (736,13 * \mathbf{TI})$$

Elde edilen diskrminant fonksiyonları yardımı ile sınıflandırma tablosu ve hata matrisi (confusion matrix) sırası ile Çizelge 6.9 ve Çizelge 6.10'daki gibi elde edilmiştir.

Çizelge 6.9. Pottery veri seti sağlam (*MCD*) diskriminant analizi sınıflandırma tablosu

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	12	1
Eritrean	2	12

MCD tahmin edicileri ile elde edilen sınıflandırma tablosuna (Çizelge 6.9) göre, gerçekte *Attic* sınıfında yer alan 12 gözlem yine *Attic* olarak sınıflanmasına karşın gerçekte *Attic* sınıfında yer alan 1 gözlem *Eritrean* sınıfında tahmin edilmiştir. Gerçekte *Eritrean* sınıfında yer alan 12 gözlem yine *Eritrean* sınıfında tahmin edilmiş olmasına karşın gerçekte *Eritrean* sınıfında yer alan 2 gözlem *Attic* sınıfında tahmin edilmiştir.

Çizelge 6.10. Pottery veri seti sağlam (*MCD*) diskriminant analizi hata matrisi

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	0,92	0,08
Eritrean	0,14	0,86

Çizelge 6.10'da verilen hata matrisine göre, *Attic* sınıfında yer alan gözlemlerin %92'si ve *Eritrean* sınıfında yer alan gözlemlerin %86'sı doğru sınıflanmıştır. Buna göre, *Hatalı*

Sınıflandırma Oranı (Apparent error rate) $0,11$ olarak hesaplanmıştır.

6.2.3. Pottery veri seti izdüşüm arama yöntemi ile diskriminant analizi

Pottery veri setine izdüşüm arama yöntemi ile diskriminant analizi uygulanmasından elde edilen sınıflandırma tablosu ve hata matrisi sırası ile Çizelge 6.11 ve Çizelge 6.12'de verilmektedir.

Çizelge 6.11. Pottery veri seti PP yöntemi ile diskriminant analizi sınıflandırma tablosu

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	12	1
Eritrean	0	14

Sağlam tahmin yöntemlerinden biri olan izdüşüm arama yöntemine göre elde edilen sınıflandırma tablosuna (Çizelge 6.11) göre, gerçekte *Attic* sınıfında yer alan 12 gözlem yine *Attic* sınıfında tahmin edilmesine karşın gerçekte *Attic* sınıfında yer alan 1 gözlem *Eritrean* sınıfında tahmin edilmiştir. Gerçekte *Eritrean* sınıfında yer alan 14 gözlem yine *Eritrean* sınıfında tahmin edilmiştir.

Çizelge 6.12. Pottery veri seti PP yöntemi ile diskriminant analizi hata matrisi

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	0,92	0,08
Eritrean	0,00	1,00

Çizelge 6.12'de verilen hata matrisine göre, *Attic* sınıfında yer alan gözlemlerin %92'si ve *Eritrean* sınıfında yer alan gözlemlerin tamamı doğru sınıflanmıştır. Buna göre *Hatalı Sınıflandırma Oranı (Apparent error rate)* $0,04$ olarak hesaplanmıştır.

6.2.4. Pottery verisi *MAD* istatistiği kullanılarak *PP* yöntemi ile diskriminant analizi

Pottery veri setine *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizinden elde edilen sınıflandırma tablosu ve hata matrisi sırası ile Çizelge 6.13 ve Çizelge 6.14'te verilmektedir.

Çizelge 6.13. Pottery veri seti PP_{MAD} yöntemi ile diskriminant analizi sınıflandırma tablosu

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	12	1
Eritrean	2	12

Sağlam tahmin yöntemlerinden biri olan (MAD) istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile elde edilen sınıflandırma tablosuna (Çizelge 6.13) göre, gerçekte *Attic* sınıfında yer alan 12 gözlem yine *Attic* olarak sınıflanmasına karşın gerçekte *Attic* sınıfında yer alan 1 gözlem *Eritrean* sınıfında tahmin edilmiştir. Gerçekte *Eritrean* sınıfında yer alan 12 gözlem yine *Eritrean* sınıfında tahmin edilmiş olmasına karşın gerçekte *Eritrean* sınıfında yer alan 2 gözlem *Attic* sınıfında tahmin edilmiştir.

Çizelge 6.14. Pottery veri seti PP_{MAD} yöntemi ile diskriminant analizi hata matrisi

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	0,92	0,08
Eritrean	0,14	0,86

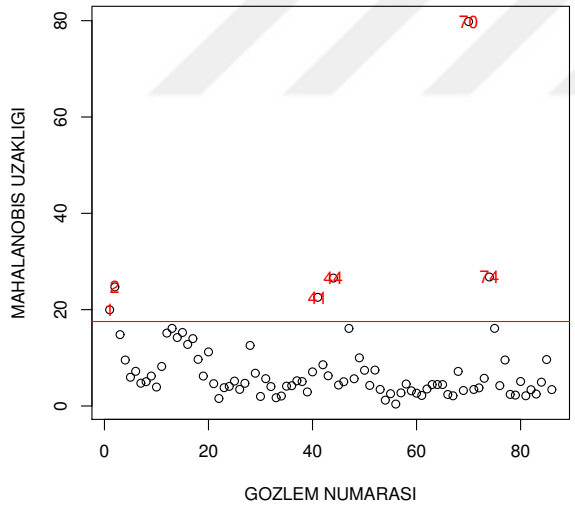
Çizelge 6.14'te verilen hata matrisine göre, *Attic* sınıfında yer alan gözlemlerin %92'si ve *Eritrean* sınıfında yer alan gözlemlerin %86'sı doğru sınıflanmıştır. Buna göre, *Hatalı Sınıflandırma Oranı (Apparent error rate)* $0,11$ olarak hesaplanmıştır.

Pottery veri setine klasik ve sağlam diskriminant analizi yöntemleri uygulanmıştır. Klasik tahmin edicilerle diskriminant analizinde yapılan sınıflamada hatalı sınıflandırma oranı $0,04$, *MCD* tahmin edicileri ile diskriminant analizinde yapılan sınıflamada hatalı sıflandırma oranı $0,11$, izdüşüm arama yöntemi ile diskriminant analizinde yapılan sınıflamada hatalı sınıflandırma oranı $0,04$ ve (MAD) istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizinde yapılan sınıflamada hatalı sınıflandırma oranı $0,11$ olarak bulunmuştur. Bu bulgulara göre *Pottery* veri setinde sağlam tahmin yöntemlerinden biri olan izdüşüm arama yöntemi ile diskriminant analizi ve klasik tahmin edicilerle diskriminant analizi uygulamalarının sonuçlarının benzer olduğu ve gözlemleri sınıflara atamada daha başarılı oldukları söylenebilir.

6.3. Milk Veri Seti İle Temel Bileşenler Analizi

Daudin, Duby ve Trecourt (1988) tarafından yayınlanan *Milk* veri seti için klasik tahmin ediciler ile , MCD tahmin edicileri ile ve izdüşüm arama (*PP*) yöntemi ile temel bileşenler analizi uygulanmış ve sonuçlar karşılaştırılmıştır. *Milk* veri seti 86 konteyner sütün 8 bileşiminin ölçülmesi ile elde edilmiştir. *Milk* veri setinde 8 açıklayıcı değişken, x_1:*yoğunluk (density) (gram/litre)*, x_2:*yağ miktarı (fat content)*, x_3:*protein miktarı (protein content)*, x_4:*kazein miktarı (casein content)*, x_5:*fabrikada ölçülen peynir kurutma maddesi (cheese dry substance measured in the factory)*, x_6:*laboratuvarda ölçülen peynir kurutma maddesi (cheese dry substance measured in the laboratory)*, x_7:*süt kurutma maddesi (milk dry substance)*, x_8:*peynir ürünü (cheese product)* ölçüm değerleri olarak belirlenmiştir.

Veri setinde başka dağılımdan karışmış gözlem olup olmadığını araştırmak amacıyla hesaplanan *Mahalanobis Uzaklıkları* Şekil 6.13'te ve Çizelge 6.15'te verilmiştir.



Şekil 6.13. Milk veri seti mahalanobis uzaklıkları grafiği

Çizelge 6.15. Milk veri seti mahalanobis uzaklıkları

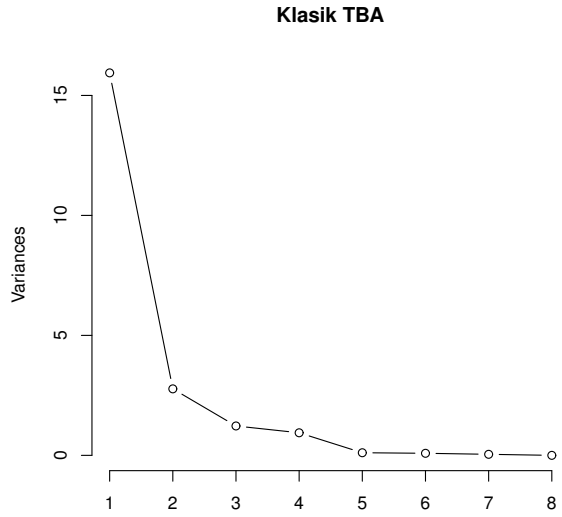
Gözlem No	1	2	3	4	5	6	7	8	9	10
MD	19,98	24,72	14,82	9,54	5,97	7,20	4,73	5,05	6,22	3,94
Gözlem No	11	12	13	14	15	16	17	18	19	20
MD	8,21	15,14	16,12	14,21	15,26	12,78	13,99	9,68	6,21	11,23
Gözlem No	21	22	23	24	25	26	27	28	29	30
MD	4,64	1,56	3,80	4,07	5,17	3,45	4,71	12,57	6,81	1,98
Gözlem No	31	32	33	34	35	36	37	38	39	40
MD	5,67	4,04	1,72	2,07	4,12	4,22	5,24	5,07	2,92	7,09
Gözlem No	41	42	43	44	45	46	47	48	49	50
MD	22,56	8,56	6,27	26,56	4,35	5,04	16,10	5,66	10,00	7,42
Gözlem No	51	52	53	54	55	56	57	58	59	60
MD	4,27	7,46	3,44	1,21	2,55	0,40	2,73	4,57	3,15	2,66
Gözlem No	61	62	63	64	65	66	67	68	69	70
MD	2,19	3,53	4,45	4,45	4,45	2,38	2,12	7,18	3,22	79,81
Gözlem No	71	72	73	74	75	76	77	78	79	80
MD	3,43	3,76	5,75	26,79	16,12	4,23	9,55	2,41	2,26	5,09
Gözlem No	81	82	83	84	85	86				
MD	2,12	3,38	2,48	4,95	9,64	3,40				

Şekil 6.13'te ve Çizelge 6.15'te yer alan *Mahalanobis Uzaklıkları* değerlerine göre *Mahalanobis Uzaklığı* $\chi^2_{8;0,975} = 17,54$ değerinden büyük olan 1., 2., 41., 44., 70. ve 74. gözlemlerin başka dağılımdan karışmış gözlem olduğu söylenebilmektedir.

Başka dağılımdan karışmış gözlem yer alan veride klasik ve sağlam temel bileşenler analizi yöntemlerinin karşılaştırılması için, klasik tahmin edicilerle, *MCD* tahmin edicileriyle, izdüşüm arama yöntemi ile ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile temel bileşenler analizi uygulanmıştır.

6.3.1. Milk veri seti klasik tahmin ediciler ile temel bileşenler analizi

Milk veri setine klasik tahmin edicilerle temel bileşenler analizi uygulanması sonucu önemli temel bileşen sayısını gösteren yamaç grafiği Şekil 6.14'teki gibi elde edilmiştir.



Şekil 6.14. Milk veri seti klasik tahmin ediciler ile TBA yamaç grafiği

Şekil 6.14'te verilen yamaç grafiğine göre, 8 açıklayıcı değişkenle tanımlanan *Milk* veri seti 4 temel bileşenle açıklanabilmektedir.

4 temel bileşenle açıklanabilecek *Milk* veri setine klasik tahmin edicilerle temel bileşenler analizi uygulanması sonucu elde edilen temel bileşenlerin standart sapmaları, toplam varyansı açıklama oranları ve kümülatif toplam varyansı açıklama oranları Çizelge 6.16'da verilmiştir.

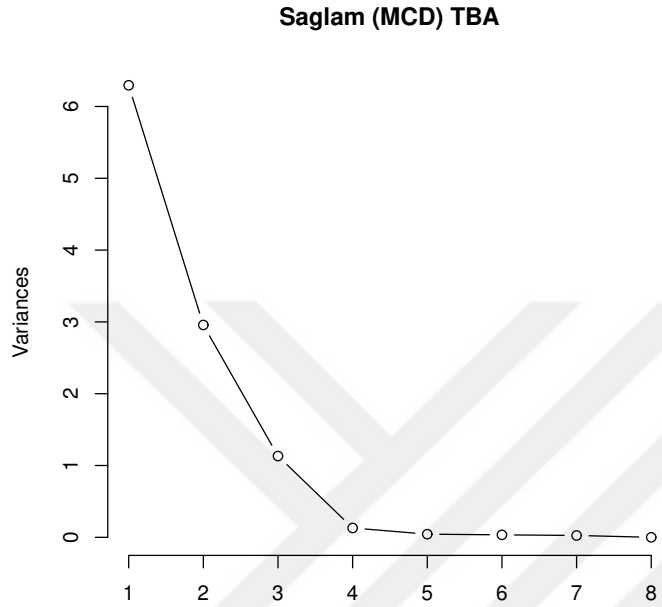
Çizelge 6.16. Milk veri seti klasik tahmin ediciler ile temel bileşenler analizi

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
Standart Sapma	3,993	1,665	1,107	0,970	0,330	0,293	0,208	0,001
Varyans Açıklama Oranı	0,755	0,131	0,058	0,045	0,005	0,004	0,002	0,000
Kümülatif Oranlar	0,755	0,886	0,944	0,989	0,994	0,998	1,000	1,000

Çizelge 6.16'da yer alan klasik tahmin edicilerle temel bileşenler analizi sonuçlarına göre, birinci temel bileşen (y_1) toplam varyansın %75'ini, ikinci temel bileşen (y_2) toplam varyansın %13'ünü, üçüncü temel bileşen (y_3) toplam varyansın %6'sını ve dördüncü temel bileşen toplam varyansın %4'ünü açıklamaktadır. 4 önemli temel bileşen toplam varyansın %99'unu açıklamaktadır.

6.3.2. Milk veri seti *MCD* tahmin edicileri ile temel bileşenler analizi

Milk veri setine *MCD* tahmin edicileri ile temel bileşenler analizi uygulanması sonucu önemli temel bileşen sayısını gösteren yamaç grafiği Şekil 6.15'teki gibi elde edilmiştir.



Şekil 6.15. Milk veri seti *MCD* tahmin edicileriyle *TBA* yamaç grafiği

Şekil 6.15'te verilen yamaç grafiğine göre, 8 açıklayıcı değişkenle tanımlanan *Milk* veri seti 3 temel bileşenle açıklanabilmektedir. 3 temel bileşenle açıklanabilecek *Milk* veri setine *MCD* tahmin edicileri ile temel bileşenler analizi uygulanması sonucu elde edilen temel bileşenlerin standart sapmaları, toplam varyansı açıklama oranları ve kümülatif toplam varyansı açıklama oranları Çizelge 6.17'de verilmiştir.

Çizelge 6.17. Milk veri seti *MCD* tahmin edicileri ile temel bileşenler analizi

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
Standart Sapma	2,396	1,741	1,071	0,365	0,210	0,187	0,156	0,001
Varyans Açıklama Oranı	0,565	0,299	0,113	0,013	0,004	0,003	0,002	0,001
Kümülatif Oranlar	0,565	0,864	0,977	0,990	0,994	0,997	0,999	1,000

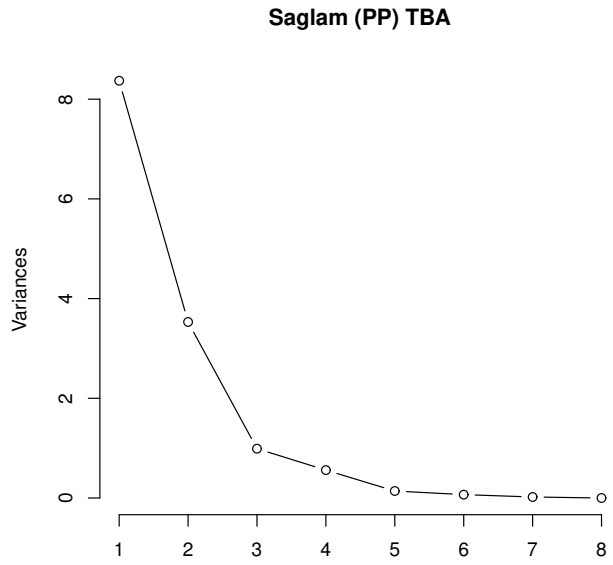
Çizelge 6.17'de yer alan *MCD* tahmin edicileri ile temel bileşenler analizi sonuçlarına göre, birinci temel bileşen (y_1) toplam varyansın %57'sini, ikinci temel bileşen (y_2) toplam

varyansın %30'unu ve üçüncü temel bileşen (y_3) toplam varyansın %11'ini açıklamaktadır. 3 önemli temel bileşen toplam varyansın %98'ini açıklamaktadır.

Klasik temel bileşenler analizi uygulanması ile 4 temel bileşene indirgenen *Milk* veri seti *MCD* tahmin edicileri ile temel bileşenler analizi uygulanması sonucu 3 temel bileşene indirgenebilmiştir. Ayrıca, Çizelge 6.16'da verilen klasik temel bileşenler analizi ile elde edilen temel bileşenlerin standart sapmaları ve Çizelge 6.17'de verilen *MCD* tahmin edicileri ile temel bileşenler analizi uygulaması sonucu elde edilen temel bileşenlerin standart sapmaları karşılaştırıldığında, Çizelge 6.17'de verilen MCD tahmin edicileri ile temel bileşenler analizi uygulaması sonucu elde edilen temel bileşenlerin standart sapmalarının daha düşük olduğu görülmektedir.

6.3.3. Milk veri seti izdüşüm arama yöntemi ile temel bileşenler analizi

Milk veri setine izdüşüm arama yöntemi ile temel bileşenler analizi uygulanması sonucu önemli temel bileşen sayısını gösteren yamaç grafiği Şekil 6.16'daki gibi elde edilmiştir.



Şekil 6.16. Milk veri seti izdüşüm arama yöntemi ile *TBA* yamaç grafiği

Şekil 6.16'da verilen yamaç grafiğine göre, 8 açıklayıcı değişkenle tanımlanan *Milk* veri seti 4 temel bileşenle açıklanabilmektedir. 4 temel bileşenle açıklanabilecek *Milk* veri setine izdüşüm arama yöntemi ile temel bileşenler analizi uygulanması sonucu elde edilen temel

bileşenlerin standart sapmaları, toplam varyansı açıklama oranları ve kümülatif toplam varyansı açıklama oranları Çizelge 6.18'de verilmiştir.

Çizelge 6.18. Milk veri seti izdüşüm arama yöntemi ile temel bileşenler analizi

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
Standart Sapma	2,893	1,879	0,995	0,748	0,374	0,260	0,142	0,003
Varyans Açıklama Oranı	0,612	0,258	0,072	0,041	0,010	0,005	0,001	0,001
Kümülatif Oranlar	0,612	0,870	0,942	0,983	0,993	0,998	0,999	1,000

Çizelge 6.18'de yer alan izdüşüm arama yöntemi ile temel bileşenler analizi sonuçlarına göre, birinci temel bileşen (y_1) toplam varyansın %61'ini, ikinci temel bileşen (y_2) toplam varyansın %26'sını, üçüncü temel bileşen (y_3) toplam varyansın %7'sini ve dördüncü temel bileşen (y_4) toplam varyansın %4'ünü açıklamaktadır. 4 önemli temel bileşen toplam varyansın %98'ini açıklamaktadır.

Milk veri seti, klasik tahmin edicilerle temel bileşenler analizi uygulaması ile 4 temel bileşene, *MCD* tahmin edicileri ile temel bileşenler analizi uygulaması sonucu 3 temel bileşene ve izdüşüm arama yöntemi ile temel bileşenler analizi uygulaması sonucu 4 temel bileşene indirgenebilmiştir. Ayrıca Çizelge 6.16'da verilen klasik tahmin edicilerle temel bileşenler analizi ile elde edilen temel bileşenlerin standart sapmaları ve Çizelge 6.17'de verilen *MCD* tahmin edicileri ile temel bileşenler analizi uygulaması sonucu elde edilen temel bileşenlerin standart sapmaları karşılaştırıldığında, Çizelge 6.17'de verilen *MCD* tahmin edicileri ile temel bileşenler analizi uygulaması sonucu elde edilen temel bileşenlerin standart sapmalarının daha düşük olduğu görülmektedir. Benzer şekilde, Çizelge 6.16'da verilen klasik tahmin edicilerle temel bileşenler analizi ile elde edilen temel bileşenlerin standart sapmaları ve Çizelge 6.18'de verilen izdüşüm arama yöntemi ile temel bileşenler analizi uygulaması sonucu elde edilen temel bileşenlerin standart sapmaları karşılaştırıldığında, Çizelge 6.18'de verilen izdüşüm arama yöntemi ile temel bileşenler analizi uygulaması sonucu elde edilen temel bileşenlerin standart sapmalarının daha düşük olduğu görülmektedir. Tüm sonuçlar kıyaslanmak istendiğinde, *Milk* veri seti için, veri setini daha az temel bileşene indirgeyen ve temel bileşenlerin standart sapmaları en düşük olan *MCD* tahmin edicileri ile temel bileşenler analizinin en iyi sonucu verdiği söylenebilmektedir. Ayrıca, klasik tahmin edicilerle temel bileşenler analizi ve izdüşüm

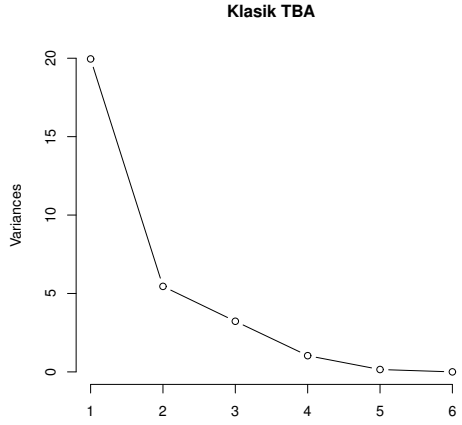
arama yöntemi ile temel bileşenler analizi uygulamaları veri setini 4 temel bileşene indirgemiş olmasına rağmen, izdüşüm arama yöntemi ile temel bileşenler analizi sonucu elde edilen temel bileşenlerin standart sapmalarının daha düşük olduğu görülmektedir. Tüm bu sonuçlara göre, *Milk* veri setine temel bileşenler analizi uygulamasında en iyi sonucu veren yöntemin *MCD* tahmin edicileri ile temel bileşenler analizi, ikinci en iyi yöntemin izdüşüm arama yöntemi ile temel bileşenler analizi olduğu söylenebilmektedir.

6.4. Pottery Veri Seti Temel Bileşenler Analizi ve Diskriminant Analizi

İçerisinde başka dağılımdan karışmış gözlem yer alan *Pottery* verisinde klasik ve sağlam temel bileşenler analizi yöntemlerinin karşılaştırılması için, klasik tahmin edicilerle temel bileşenler analizi, *MCD* tahmin edicileriyle temel bileşenler analizi ve izdüşüm arama yöntemi ile temel bileşenler analizi uygulanmıştır. Temel bileşenler analizi sonucu temel bileşen skorları ile elde edilen veri setlerine klasik tahmin edicilerle, *MCD* tahmin edicileriyle, izdüşüm arama yöntemi ile ve (MAD) istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizi uygulanarak sonuçlar karşılaştırılmıştır.

6.4.1. Pottery veri seti klasik tahmin ediciler ile temel bileşenler analizi

Pottery veri setine klasik tahmin edicilerle temel bileşenler analizi uygulanması sonucu önemli temel bileşen sayısını gösteren yamaç grafiği Şekil 6.17'deki gibi elde edilmiştir.



Şekil 6.17. Pottery veri seti klasik tahmin ediciler ile *TBA* yamaç grafiği

Şekil 6.17'de verilen yamaç grafiğine göre, 6 açıklayıcı değişkenle tanımlanan *Pottery* veri seti 3 temel bileşenle açıklanabilmektedir.

3 temel bileşenle açıklanabilecek *Pottery* veri setine klasik tahmin edicilerle temel bileşenler analizi uygulanması sonucu elde edilen temel bileşenlerin standart sapmaları, toplam varyansı açıklama oranları ve kümülatif toplam varyansı açıklama oranları Çizelge 6.19'da verilmiştir.

Çizelge 6.19. Pottery veri seti klasik tahmin ediciler ile temel bileşenler analizi

	y_1	y_2	y_3	y_4	y_5	y_6
Standart Sapma	4,467	2,335	1,797	1,017	0,389	0,024
Varyans Açıklama Oranı	0,669	0,183	0,108	0,035	0,005	0,000
Kümülatif Oranlar	0,669	0,852	0,960	0,995	1,000	1,000

Çizelge 6.19'da yer alan klasik tahmin edicilerle temel bileşenler analizi sonuçlarına göre, birinci temel bileşen (y_1) toplam varyansın %67'sini, ikinci temel bileşen (y_2) toplam varyansın %18'ini ve üçüncü temel bileşen (y_3) toplam varyansın %11'ini açıklamaktadır. 3 önemli temel bileşen toplam varyansın %96'sını açıklamaktadır.

3 önemli temel bileşen ile hesaplanan temel bileşen skorlarına göre elde edilen veriye klasik tahmin edicilerle, *MCD* tahmin edicileriyle, izdüşüm arama yöntemi ile ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizi uygulanmıştır.

Pottery veri setine klasik TBA uygulanmasının ardından klasik ve sağlam DA uygulanması

Pottery veri setine klasik tahmin edicilerle temel bileşenler analizi uygulanması sonucu verinin 3 temel bileşenle açıklanabildiği görülmüştür. 3 temel bileşenle açıklanan veri setine klasik tahmin edicilerle, *MCD* tahmin edicileri ile, izdüşüm arama yöntemi ile ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizi uygulanması ile elde edilen sınıflandırma tablosu ve hata matrisi sırası ile Çizelge 6.20 ve Çizelge 6.21'deki gibi elde edilmiştir.

Çizelge 6.20. Pottery veri seti klasik *TBA* sonrası *DA* sınıflandırma tablosu

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	12	1
Eritrean	0	14

Elde edilen sınıflandırma tablosuna (Çizelge 6.20) göre, gerçekte *Attic* sınıfında yer alan 12 gözlem yine *Attic* sınıfında tahmin edilmesine karşın gerçekte *Attic* sınıfında yer alan 1 gözlem *Eritrean* sınıfında tahmin edilmiştir. Gerçekte *Eritrean* sınıfında yer alan 14 gözlem yine *Eritrean* sınıfında tahmin edilmiştir.

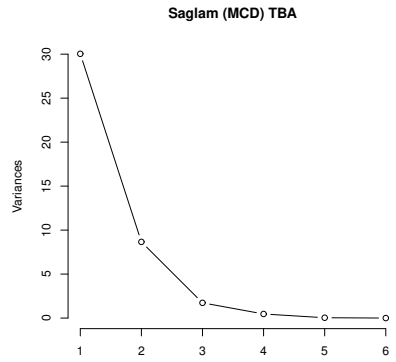
Çizelge 6.21. Pottery veri seti klasik *TBA* sonrası *DA* hata matrisi

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	0,92	0,08
Eritrean	0,00	1,00

Çizelge 6.21'de verilen hata matrisine göre, *Attic* sınıfında yer alan gözlemlerin %92'si ve *Eritrean* sınıfında yer alan gözlemlerin tamamı doğru sınıflanmıştır. Buna göre *Hatalı Sınıflandırma Oranı (Apparent error rate)* $0,04$ olarak hesaplanmıştır.

6.4.2. Pottery veri seti *MCD* tahmin edicileri ile temel bileşenler analizi

Pottery veri setine *MCD* tahmin edicileri ile temel bileşenler analizi uygulanması sonucu önemli temel bileşen sayısını gösteren yamaç grafiği Şekil 6.18'deki gibi elde edilmiştir.



Şekil 6.18. Pottery veri seti *MCD* tahmin edicileri ile *TBA* yamaç grafiği

Şekil 6.18'de verilen yamaç grafiğine göre, 6 açıklayıcı değişkenle tanımlanan *Pottery* veri seti 2 temel bileşenle açıklanabilmektedir.

2 temel bileşenle açıklanabilecek *Pottery* veri setine *MCD* tahmin edicileri ile temel bileşenler analizi uygulanması sonucu elde edilen temel bileşenlerin standart sapmaları, toplam varyansı açıklama oranları ve kümülatif toplam varyansı açıklama oranları Çizelge 6.22'de verilmiştir.

Çizelge 6.22. Pottery veri seti *MCD* tahmin edicileri ile temel bileşenler analizi

	y_1	y_2	y_3	y_4	y_5	y_6
Standart Sapma	5,481	2,943	1,316	0,682	0,195	0,027
Varyans Açıklama Oranı	0,734	0,212	0,042	0,011	0,001	0,000
Kümülatif Oranlar	0,734	0,945	0,988	0,999	1,000	1,000

Çizelge 6.22'de yer alan *MCD* tahmin edicileri ile temel bileşenler analizi sonuçlarına göre, birinci temel bileşen (y_1) toplam varyansın %73'ünü ve ikinci temel bileşen (y_2) toplam varyansın %21'ini açıklamaktadır. 2 önemli temel bileşen toplam varyansın %95'ini açıklamaktadır.

2 önemli temel bileşen ile hesaplanan temel bileşen skorlarına göre elde edilen veriye klasik tahmin edicilerle, *MCD* tahmin edicileriyle, izdüşüm arama yöntemi ile ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizi uygulanmıştır.

Pottery veri setine MCD ile TBA uygulanmasından sonra klasik ve sağlam DA uygulanması

Pottery veri setine *MCD* tahmin edicileri ile temel bileşenler analizi uygulanması sonucu verinin 2 temel bileşenle açıklanabildiği görülmüştür. 2 temel bileşenle açıklanan veri setine klasik tahmin ediciler ile, *MCD* tahmin edicileri ile, izdüşüm arama yöntemi ile ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizi uygulanması ile elde edilen sınıflandırma tablosu ve hata matrisi sırası ile Çizelge 6.23 ve Çizelge 6.24'teki gibi elde edilmiştir.

Çizelge 6.23. Pottery veri seti *MCD TBA* sonrası *DA* sınıflandırma tablosu

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	12	1
Eritrean	2	12

Elde edilen sınıflandırma tablosuna (Çizelge 6.23) göre, gerçekte *Attic* sınıfında yer alan 12 gözlem yine *Attic* sınıfında tahmin edilmesine karşın gerçekte *Attic* sınıfında yer alan 1 gözlem *Eritrean* sınıfında tahmin edilmiştir. Gerçekte *Eritrean* sınıfında yer alan 12 gözlem yine *Eritrean* sınıfında tahmin edilmesine karşın gerçekte *Eritrean* sınıfında yer alan 2 gözlem *Attic* sınıfında tahmin edilmiştir.

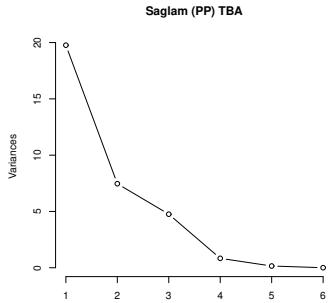
Çizelge 6.24. Pottery veri seti *MCD TBA* sonrası *DA* hata matrisi

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	0,92	0,08
Eritrean	0,14	0,86

Çizelge 6.24'te verilen hata matrisine göre, *Attic* sınıfında yer alan gözlemlerin %92'si ve *Eritrean* sınıfında yer alan gözlemlerin %86'sı doğru sınıflanmıştır. Buna göre *Hatalı Sınıflandırma Oranı (Apparent error rate)* $0,11$ olarak hesaplanmıştır.

6.4.3. Pottery veri seti izdüşüm arama yöntemi ile temel bileşenler analizi

Pottery veri setine izdüşüm arama yöntemi ile temel bileşenler analizi uygulanması sonucu önemli temel bileşen sayısını gösteren yamaç grafiği Şekil 6.19'daki gibi elde edilmiştir.



Şekil 6.19. Pottery veri seti izdüşüm arama yöntemi ile *TBA* yamaç grafiği

Şekil 6.19'da verilen yamaç grafiğine göre, 6 açıklayıcı değişkenle tanımlanan *pottery* veri seti 3 temel bileşenle açıklanabilmektedir.

3 temel bileşenle açıklanabilecek *Pottery* veri setine izdüşüm arama yöntemi ile temel bileşenler analizi uygulanması sonucu elde edilen temel bileşenlerin standart sapmaları, toplam varyansı açıklama oranları ve kümülatif toplam varyansı açıklama oranları Çizelge 6.25'te verilmiştir.

Çizelge 6.25. Pottery veri seti izdüşüm arama yöntemi ile temel bileşenler analizi

	y_1	y_2	y_3	y_4	y_5	y_6
Standart Sapma	4,446	2,732	2,181	0,914	0,393	0,058
Varyans Açıklama Oranı	0,599	0,226	0,144	0,025	0,005	0,000
Kümülatif Oranlar	0,599	0,826	0,970	0,995	1,000	1,000

Çizelge 6.25'te yer alan izdüşüm arama yöntemi ile temel bileşenler analizi sonuçlarına göre, birinci temel bileşen (y_1) toplam varyansın %60'ını, ikinci temel bileşen (y_2) toplam varyansın %23'ünü ve üçüncü temel bileşen (y_3) toplam varyansın %14'ünü açıklamaktadır. 3 önemli temel bileşen toplam varyansın %97'sini açıklamaktadır.

3 önemli temel bileşen ile hesaplanan temel bileşen skorlarına göre elde edilen veriye klasik tahmin edicilerle, *MCD* tahmin edicileriyle, izdüşüm arama yöntemi ile ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizi uygulanmıştır.

Pottery veri setine PP ile TBA uygulamasından sonra klasik ve sağlam DA uygulanması

Pottery veri setine izdüşüm arama yöntemi ile temel bileşenler analizi uygulanması sonucu verinin 3 temel bileşenle açıklanabildiği görülmüştür. 3 temel bileşenle açıklanan veri setine klasik tahmin ediciler, *MCD* tahmin edicileri, izdüşüm arama yöntemi ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizi uygulamasından elde edilen sınıflandırma tablosu ve hata matrisi sırası ile Çizelge 6.26 ve Çizelge 6.27'deki gibi elde edilmiştir.

Çizelge 6.26. Pottery veri seti *PP* yöntemi ile *TBA* sonrası *DA* sınıflandırma tablosu

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	12	1
Eritrean	0	14

Elde edilen sınıflandırma tablosuna (Çizelge 6.26) göre, gerçekte *Attic* sınıfında yer alan 12 gözlem yine *Attic* sınıfında tahmin edilmesine karşın gerçekte *Attic* sınıfında yer alan 1 gözlem *Eritrean* sınıfında tahmin edilmiştir. Gerçekte *Eritrean* sınıfında yer alan 14 gözlem yine *Eritrean* sınıfında tahmin edilmiştir.

Çizelge 6.27. Pottery veri seti *PP* yöntemi ile *TBA* sonrası *DA* hata matrisi

Gerçekte Olan \ Tahmin Edilen	Attic	Eritrean
Attic	0,92	0,08
Eritrean	0,00	1,00

Çizelge 6.27'de verilen hata matrisine göre, *Attic* sınıfında yer alan gözlemlerin %92'si ve *Eritrean* sınıfında yer alan gözlemlerin tamamı doğru sınıflanmıştır. Buna göre *Hatalı Sınıflandırma Oranı (Apparent error rate)* $0,04$ olarak hesaplanmıştır.

7. SİMÜLASYON ÇALIŞMASI

Tahmin edicilerin karşılaştırılması için farklı özellikte iki değişkenli, 2 ve 3 sınıflı tesadüfi veri setleri üretilerek simülasyon çalışması yapılmıştır. İncelenen tüm durumlar için gözlemlerin ait oldukları sınıflar Normal dağılıma sahiptir. Sınıfların varyans-kovaryans matrisleri birbirine eşit ve $p * p$ boyutlu birim matristir (I_{p*p}). Birinci sınıfın ortalaması merkezden geçer, ikinci sınıfın ortalaması merkeze 2 birim uzaklıktadır ve üçüncü sınıfın ortalaması da aynı şekilde merkeze 2 birim uzaklıktadır fakat ikinci grubun doğrultusuna dik bir doğrultudadır. Grupların varyans kovaryans matrisleri ve ortalama vektörleri Eş. 7.1'deki gibidir.

$$\mu_1 = \begin{bmatrix} 0 & 0 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 2 & 0 \end{bmatrix} \qquad \mu_3 = \begin{bmatrix} 0 & 2 \end{bmatrix} \qquad \Sigma_1 = \Sigma_2 = \Sigma_3 = I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{7.1}$$

Tesadüfi olarak üretilen veri setine aşağıda sıralanan 4 durum için $\varepsilon = \{\%0, \%5, \%10, \%15\}$ oranlarında karışma uygulanmıştır.

Durum 1: Veri setine herhangi bir karışmanın olmadığı ($\varepsilon = 0$) durum incelenmiştir.

Durum 2: n_i, i. sınıfın gözlem sayısı, μ_i, i. grubun ortalama vektörü, ε karışma oranı, κ ölçek büyütme faktörü ($\kappa = \{9,\ 100\}$) olmak üzere, ölçek karışması verileri Eş. 7.2'deki gibi oluşturulmuştur (Todorov ve Pires, 2007). Ölçek karışması için 48 farklı durum sonuçları incelenmiştir.

$$F_\varepsilon(x) = (1-\varepsilon)F_0(x) + \varepsilon F_k(x) = (1-\varepsilon)N_i(\mu_i, I_{p*p}) + \varepsilon N_i(\mu_i, \kappa I_{p*p}) \tag{7.2}$$

Durum 3: n_i, i. sınıfın gözlem sayısı, μ_i, i. grubun ortalama vektörü, ε karışma oranı, ν konum değiştirme parametresi ($\nu = \{4,\ 8\}$), Q_p, konum kaydırma ölçüsü olmak üzere, konum karışması verileri Eş. 7.3'teki gibi oluşturulmuştur (Todorov ve Pires, 2007). Konum karışması için 48 farklı durum sonuçları incelenmiştir.

$$F_\varepsilon(x) = (1-\varepsilon)F_0(x) + \varepsilon F_k(x) = (1-\varepsilon)N_i(\mu_i, I_{p*p}) + \varepsilon N_i(\hat{\mu}_i, 0,25^2 I_{p*p})$$

$$Q_p = \sqrt{\chi^2_{p;0,001}/p}$$
$$\hat{\mu}_i = \mu_i + (\nu Q_p, ..., \nu Q_p) \tag{7.3}$$

Durum 4: μ_i, i. grubun ortalama vektörü, ε karışma oranı, ρ, ilişki (korelasyon) katsayısı ($\rho = \{0,10;\ 0,25;\ 0,50;\ 0,75\}$) olmak üzere $\mu = [0\ 0]$ ve $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ karışma durumu incelenmiştir (Gündüz ve Fokoue, 2017). İlişki katsayısına göre karışma durumu için 96 farklı durum incelenmiştir.

Tesadüfi olarak üretilen veri setlerine klasik ve sağlam diskriminant analizi yöntemleri uygulanmıştır. Simülasyon çalışması, R versiyon 3.2.2 programı ile 500 tekrarlı yapılmıştır.

7.1. Durum 1: Veride Karışma Olmaması Durumu

Eş. 7.1'de verilen tesadüfi olarak üretilmiş iki ve üç sınıflı normal dağılımlı veri setlerine başka bir dağılımdan karışma olmaması durumu incelenmiştir. Klasik diskriminant analizi, *MCD* tahmin edicileri ile diskriminant analizi, izdüşüm arama yöntemi ile diskriminant analizi, *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizi sonuçları karşılaştırılmış ve elde edilen sonuçlar Çizelge 7.1'de özetlenmiştir.

Çizelge 7.1. Durum 1- karışmasız veri seti hatalı sınıflandırma oranları

Sınıf Sayısı	Klasik	MCD	PP	PP_{MAD}
2	0,1562	0,1549	0,1553	0,1637
3	0,2168	0,2173	0,4303	0,4398

Karışmanın olmadığı durumda tesadüfi olarak üretilen iki sınıflı veri seti için yapılan simülasyon çalışmasında, klasik tahmin edicilerle diskriminant analizi ile sağlam tahmin yöntemlerinden olan *MCD* tahmin edicileri ile diskriminant analizinin ve izdüşüm arama (*PP*) yöntemi ile diskriminant analizinin hatalı sınıflandırma oranlarının birbirine çok benzer elde edildiği Çizelge 7.1'de görülmektedir. Sadece, *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi (PP_{MAD}) ile diskriminant analizi sonucunda elde edilen hatalı sınıflandırma olasılığının diğer yöntemlere göre biraz daha yüksek olduğu gözlemlenmektedir (Çizelge 7.1). Karışmanın olmadığı durumda tesadüfi olarak üretilen üç

sınıflı veri seti için yapılan simülasyon çalışmasında ise, klasik tahmin edicilerle dikriminant analizi ve *MCD* tahmin edicileri ile diskriminant analizinden elde edilen hatalı sınıflandırma oranlarının benzer olduğu gözlemlenmektedir. İzdüşüm arama yöntemi (*PP*) ile diskriminant analizi ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi (PP_{MAD}) ile diskriminant analizi sonucunda elde edilen hatalı sınıflandırma olasılıklarının ikiden fazla sınıflı veri setlerinde başarılı sonuçlar vermediği gözlemlenmektedir (Çizelge 7.1).

7.2. Durum 2: Ölçek Karışması Durumu

Tesadüfi olarak üretilen iki ve üç sınıflı veri setlerine, ölçek büyütme faktörü $\kappa = \{9,\ 100\}$ ve $\varepsilon = \{\%0, \%5,\ \%10,\ \%15\}$ oranlarında karışma olan veri setleri için yapılan simülasyon çalışmasında, klasik tahmin edicilerle diskriminant analizi, *MCD* tahmin edicileri ile diskriminant analizi, izdüşüm arama yöntemi (*PP*) ile diskriminant analizi, *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi (PP_{MAD}) ile diskriminant analizi uygulanması sonucu hesaplanan hatalı sınıflandırma oranları Çizelge 7.2'de verilmiş ve Şekil 7.1'de ve Şekil 7.2'de grafikler ile gösterilmiştir.

Çizelge 7.2. Durum 2 - ölçek karışması durumları için hatalı sınıflandırma oranları

Sınıf Sayısı	κ	ε	Klasik	MCD	PP	PP_{MAD}
2	9	0,00	0,1562	0,1549	0,1553	0,1637
		0,05	0,1700	0,1662	0,1669	0,1743
		0,10	0,1795	0,1755	0,1746	0,1831
		0,15	0,1893	0,1866	0,1841	0,1909
	100	0,00	0,1562	0,1549	0,1553	0,1637
		0,05	0,1938	0,1700	0,1708	0,1782
		0,10	0,2105	0,1839	0,1821	0,1906
		0,15	0,2326	0,1985	0,1977	0,2027
3	9	0,00	0,2168	0,2173	0,4303	0,4398
		0,05	0,2340	0,2323	0,4382	0,4460
		0,10	0,2479	0,2459	0,4446	0,4521
		0,15	0,2648	0,2587	0,4535	0,4585
	100	0,00	0,2168	0,2173	0,4303	0,4398
		0,05	0,2719	0,2379	0,4406	0,4477
		0,10	0,3066	0,2576	0,4490	0,4565
		0,15	0,3466	0,2755	0,4613	0,4649

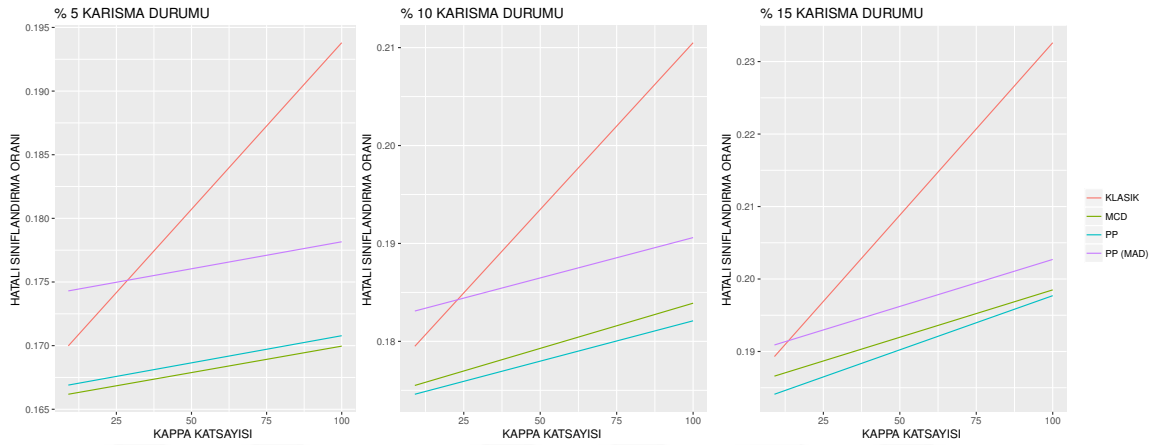
Tesadüfi olarak üretilen iki sınıflı, $\kappa = \{9, 100\}$ ölçek büyütme faktörü ve $\varepsilon = \{\%0, \%5, \%10, \%15\}$ oranlarında karışma olan veri setileri için yapılan simülasyon

çalışmasında elde edilen hatalı sınıflandırma oranlarına göre, *MCD* tahmin edicileri ile diskriminant analizinin ve izdüşüm arama yöntemi (*PP*) ile diskriminant analizinin hatalı sınıflandırma oranlarının birbirine çok benzer elde edildiği ve klasik tahmin edicilerle diskriminant analizi ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizinden elde edilen hatalı sınıflandırma oranlarından daha düşük olduğu Çizelge 7.2'de görülmektedir. Ayrıca, tesadüfi olarak üretilen iki sınıflı, $\kappa = \{9, 100\}$ ölçek büyütme faktörü ve $\varepsilon = \%5$ oranında karışma olan veri seti için yapılan simülasyon çalışmasında *MCD* tahmin edicileri ile diskriminant analizinden elde edilen hatalı sınıflandırma oranının en düşük olduğu, yani, *MCD* tahmin edicileri ile diskriminant analizinin daha başarılı sonuç verdiği söylenebilir. Benzer şekilde, tesadüfi olarak üretilen 2 sınıflı, $\kappa = \{9, 100\}$ ölçek büyütme faktörü ve $\varepsilon = \{\%10, \%15\}$ oranlarında karışma olan veri setleri için yapılan simülasyon çalışmasında izdüşüm arama yöntemi (*PP*) ile diskriminant analizinden elde edilen hatalı sınıflandırma oranının en düşük olduğu, yani, izdüşüm arama yöntemi (*PP*) ile diskriminant analizinin daha başarılı sonuç verdiği söylenebilir. Ayrıca, ölçek büyütme faktörü (κ) ve karışma oranı (ε) arttıkça klasik tahmin edicilerle diskrimnant analizinden elde edilen hatalı sınıflandırma oranlarının arttığı da görülmektedir.

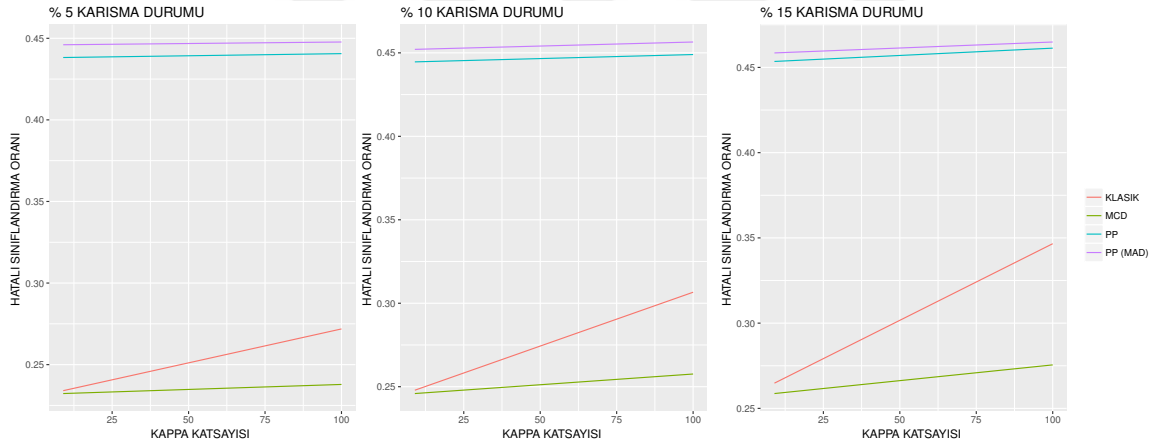
Tesadüfi olarak üretilen üç sınıflı, $\kappa = \{9, 100\}$ ölçek büyütme faktörü ve $\varepsilon = \{\%0, \%5, \%10, \%15\}$ oranlarında karışma olan veri setileri için yapılan simülasyon çalışmasında elde edilen hatalı sınıflandırma oranlarına göre, *MCD* tahmin edicileri ile diskriminant analizinin ve klasik tahmin ediciler ile diskriminant analizinin hatalı sınıflandırma oranlarının birbirine benzer elde edildiği ve izdüşüm arama yöntemi (*PP*) ile diskriminant analizi ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizinden elde edilen hatalı sınıflandırma oranlarından çok daha düşük olduğu Çizelge 7.2'de görülmektedir. Ancak, tesadüfi olarak üretilen 3 sınıflı, $\kappa = \{9, 100\}$ ölçek büyütme faktörü ve $\varepsilon = \{\%5, \%10, \%15\}$ oranında karışma olan veri setleri için yapılan simülasyon çalışmasında *MCD* tahmin edicileri ile diskriminant analizinden elde edilen hatalı sınıflandırma oranının en düşük olduğu, yani, *MCD* tahmin edicileri ile diskriminant analizinin daha başarılı sonuç verdiği söylenebilir. Ayrıca ölçek büyütme faktörü (κ) ve karışma oranı (ε) arttıkça klasik tahmin edicilerle diskrimnant analizinden elde edilen hatalı sınıflandırma oranlarının arttığı da görülmektedir.

Tesadüfi olarak üretilen 2 ve 3 sınıflı veri setleri için farklı karışma oranlarında hatalı

sınıflandırma oranlarına ilişkin grafikler sırası ile Şekil 7.1'de ve Şekil 7.2'de gösterilmiştir.



Şekil 7.1. İki sınıflı veri seti %5, % 10 ve %15 oranlarında ölçek karışması



Şekil 7.2. Üç sınıflı veri seti %5, % 10 ve %15 oranlarında ölçek karışması

Şekil 7.1'e göre; tesadüfi olarak üretilen 2 sınıflı, $\kappa = \{9, 100\}$ ölçek büyütme faktörü ve $\varepsilon = \{\%5, \%10, \%15\}$ oranlarında karışma olan veri setileri için yapılan simülasyon çalışmasında, veri setine %5 oranında ölçek karışması olması durumunda en düşük hatalı sınıflandırma oranlarının *MCD* tahmin edicileri ile diskriminant analizi uygulaması sonucu elde edildiği gözlemlenmektedir. 2 sınıflı veri setine %10 ve %15 oranlarında ölçek karışmaları olması durumlarında en düşük hatalı sınıflandırma oranlarının izdüşüm arama (*PP*) yöntemi ile diskriminant analizi uygulaması sonucu elde edildiği gözlemlenmektedir. Ayrıca klasik tahmin ediciler ile diskriminant analizi uygulanması sonucu elde edilen hatalı

sınıflandırma oranlarının ölçek büyütme faktörü (κ) ve karışma oranı (ε) arttıkça arttığı görülmektedir.

Şekil 7.2'ye göre; tesadüfi olarak üretilen 3 sınıflı, $\kappa = \{9, 100\}$ ölçek büyütme faktörü ve $\varepsilon = \{\%5, \%10, \%15\}$ oranlarında karışma olan veri setileri için yapılan simülasyon çalışmasında, en iyi sonuçların *MCD* tahmin edicileri ile diskriminant analizi uygulaması sonucu elde edildiği gözlemlenmektedir. İzdüşüm arama yöntemi (*PP*) ile diskriminant analizi ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizinden elde edilen hatalı sınıflandırma oranlarının klasik tahmin ediciler ile diskriminant analizi ve *MCD* tahmin edicileri ile diskriminant analizi uygulaması sonucu elde edilen hatalı sınıflandırma oranlarından daha yüksek olduğu görülmektedir. Ayrıca, klasik tahmin ediciler ile diskriminant analizi uygulanması sonucu elde edilen hatalı sınıflandırma oranlarının ölçek büyütme faktörü (κ) ve karışma oranı (ε) arttıkça arttığı görülmektedir.

7.3. Durum 3: Konum Karışması Durumu

Tesadüfi olarak üretilen iki ve üç sınıflı veri setlerine, konum değiştirme parametresi $\nu = \{4, 8\}$ ve $\varepsilon = \{\%0, \%5, \%10, \%15\}$ oranlarında karışma olan veri setleri için yapılan simülasyon çalışmasında, klasik tahmin edicilerle diskriminant analizi, *MCD* tahmin edicileri ile diskriminant analizi, izdüşüm arama yöntemi (*PP*) ile diskriminant analizi ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi (PP_{MAD}) ile diskriminant analizi uygulanması sonucu hesaplanan hatalı sınıflandırma oranları Çizelge 7.3'te verilmiş ve Şekil 7.3'te ve Şekil 7.4'te grafikler ile gösterilmiştir.

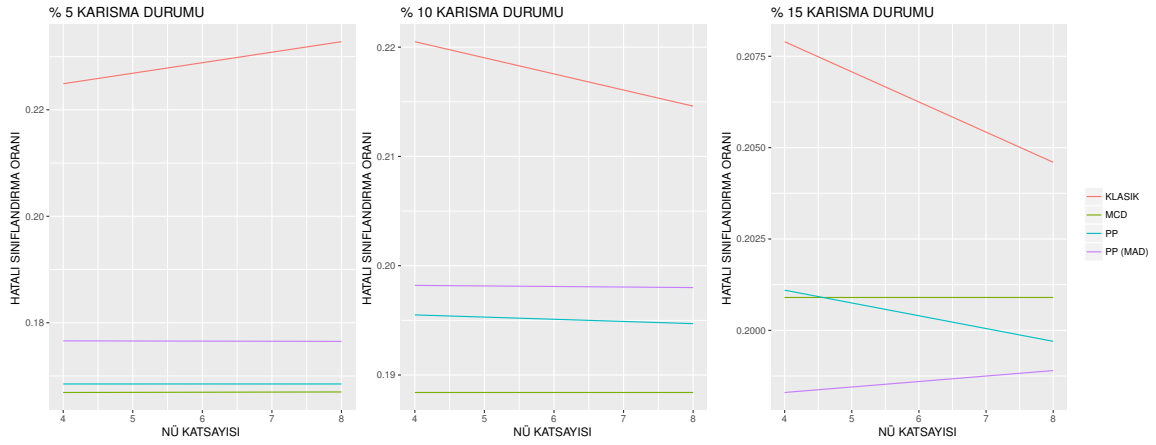
Çizelge 7.3. Durum 3 - konum karışması durumları için hatalı sınıflandırma oranları

Sınıf Sayısı	ν	ε	Klasik	MCD	PP	PP_{MAD}
2	4	0,00	0,1562	0,1549	0,1553	0,1637
		0,05	0,2249	0,1669	0,1685	0,1766
		0,10	0,2205	0,1884	0,1955	0,1982
		0,15	0,2079	0,2009	0,2011	0,1983
	8	0,00	0,1562	0,1549	0,1553	0,1637
		0,05	0,2328	0,1670	0,1685	0,1765
		0,10	0,2146	0,1884	0,1947	0,1980
		0,15	0,2046	0,2009	0,1997	0,1989
3	4	0,00	0,2168	0,2173	0,4303	0,4398
		0,05	0,2976	0,2347	0,4468	0,4537
		0,10	0,2863	0,2497	0,4606	0,4638
		0,15	0,2758	0,2574	0,4635	0,4646
	8	0,00	0,2168	0,2173	0,4303	0,4398
		0,05	0,3080	0,2378	0,4469	0,4537
		0,10	0,2839	0,2561	0,4607	0,4640
		0,15	0,2725	0,2675	0,4631	0,4643

Tesadüfi olarak üretilen 2 sınıflı, $\nu = \{4,8\}$ konum değiştirme parametresi ve $\varepsilon = \{\%0,\%5,\%10,\%15\}$ oranlarında karışma olan veri setileri için yapılan simülasyon çalışmasında elde edilen hatalı sınıflandırma oranlarına göre, *MCD* tahmin edicileri ile diskriminant analizinin ve izdüşüm arama yöntemi (*PP*) ile diskriminant analizinin hatalı sınıflandırma oranlarının birbirine benzer elde edildiği ve klasik tahmin edicilerle diskriminant analizi ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizinden elde edilen hatalı sınıflandırma oranlarından daha düşük olduğu Çizelge 7.3'te görülmektedir. Ancak, tesadüfi olarak üretilen 2 sınıflı, $\nu = \{4,8\}$ konum değiştirme parametresi ve $\varepsilon = \{\%5,\%10\}$ oranında karışma olan veri setleri için yapılan simülasyon çalışmasında *MCD* tahmin edicileri ile diskriminant analizinden elde edilen hatalı sınıflandırma oranının en düşük olduğu, yani, *MCD* tahmin edicileri ile diskriminant analizinin daha başarılı sonuç verdiği söylenebilir. Benzer şekilde, tesadüfi olarak üretilen 2 sınıflı, $\nu = \{4,8\}$ konum değiştirme parametresi ve $\varepsilon = \%15$ oranlarında karışma olan veri setleri için yapılan simülasyon çalışmasında *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi (PP_{MAD}) ile diskriminant analizinden elde edilen hatalı sınıflandırma oranının en düşük olduğu, yani, *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi (PP_{MAD}) ile diskriminant analizinin daha başarılı sonuç verdiği söylenebilir. Ayrıca, konum karışması durumlarında en yüksek hatalı sınıflandırma oranlarının klasik tahmin ediciler ile diskriminant analizinden elde edildiği görülmektedir.

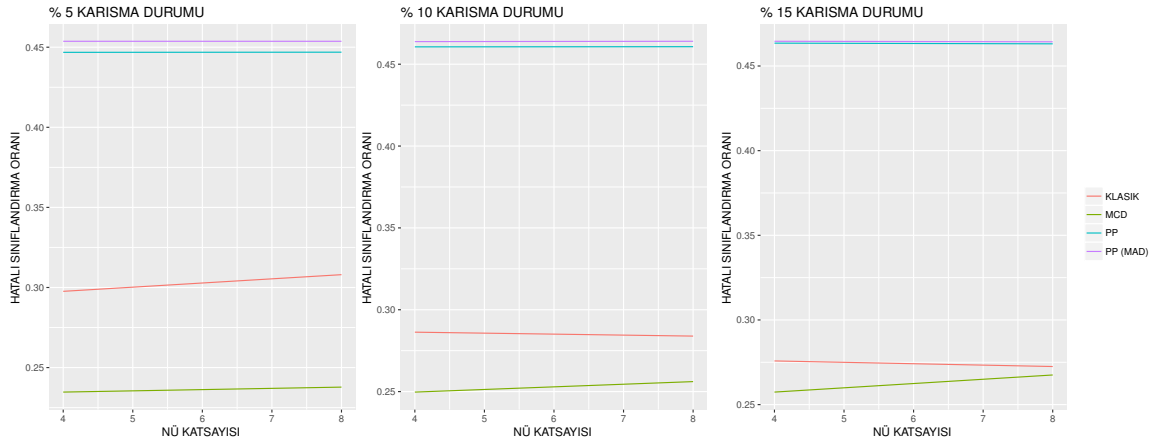
Tesadüfi olarak üretilen 3 sınıflı, $\nu = \{4,8\}$ konum değiştirme parametresi ve $\varepsilon = \{\%5,\%10,\%15\}$ oranlarında karışma olan veri setileri için yapılan simülasyon çalışmasında elde edilen hatalı sınıflandırma oranlarına göre, *MCD* tahmin edicileri ile diskriminant analizinin hatalı sınıflandırma oranlarının en düşük olduğu, yani, *MCD* tahmin edicileri ile diskriminant analizinin daha başarılı sonuç verdiği söylenebilir.

Tesadüfi olarak üretilen 2 ve 3 sınıflı veri setleri için farklı karışma oranlarında hatalı sınıflandırma oranlarına ilişkin grafikler sırası ile Şekil 7.3'te ve Şekil 7.4'te gösterilmiştir.



Şekil 7.3. İki sınıflı veri seti %5, % 10 ve %15 oranlarında konum karışması

Şekil 7.3'e göre; tesadüfi olarak üretilen 2 sınıflı, $\nu = \{4,8\}$ konum değiştirme parametresi ve $\varepsilon = \{\%5,\%10,\%15\}$ oranlarında karışma olan veri setileri için yapılan simülasyon çalışmasında, veri setine $\varepsilon = \{\%5,\%10\}$ oranında konum karışmaları olması durumlarında en düşük hatalı sınıflandırma oranlarının *MCD* tahmin edicileri ile diskriminant analizi uygulaması sonucu elde edildiği gözlemlenmektedir. 2 sınıflı veri setine %15 oranında konum karışması olması durumunda en düşük hatalı sınıflandırma oranlarının *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizi uygulaması sonucu elde edildiği gözlemlenmektedir.

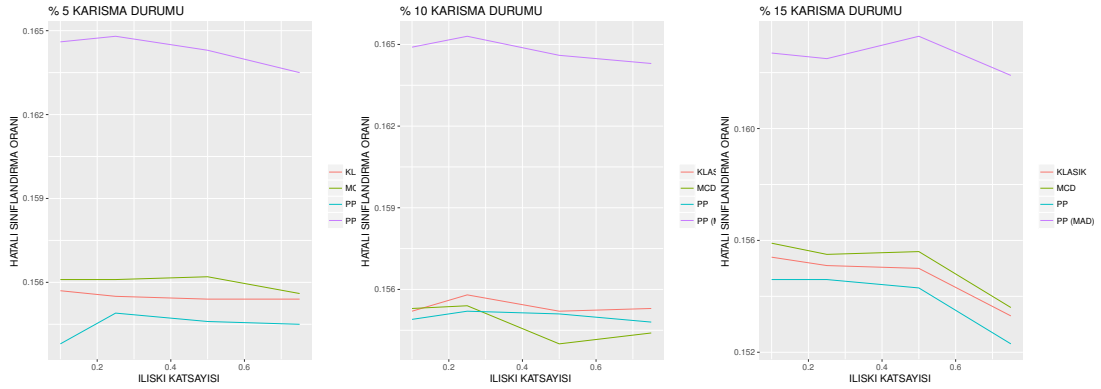


Şekil 7.4. Üç sınıflı veri seti %5, % 10 ve %15 oranlarında konum karışması

Şekil 7.4'e göre; tesadüfi olarak üretilen 3 sınıflı, $\nu = \{4, 8\}$ konum değiştirme parametresi ve $\varepsilon = \{\%5, \%10, \%15\}$ oranlarında karışma olan veri setileri için yapılan simülasyon çalışmasında, en iyi sonuçların *MCD* tahmin edicileri ile diskriminant analizi uygulaması sonucu elde edildiği gözlemlenmektedir. İzdüşüm arama yöntemi (*PP*) ile diskriminant analizi ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizinden elde edilen hatalı sınıflandırma oranlarının klasik tahmin ediciler ile diskriminant analizi ve *MCD* tahmin edicileri ile diskriminant analizi uygulaması sonucu elde edilen hatalı sınıflandırma oranlarından daha yüksek olduğu görülmektedir.

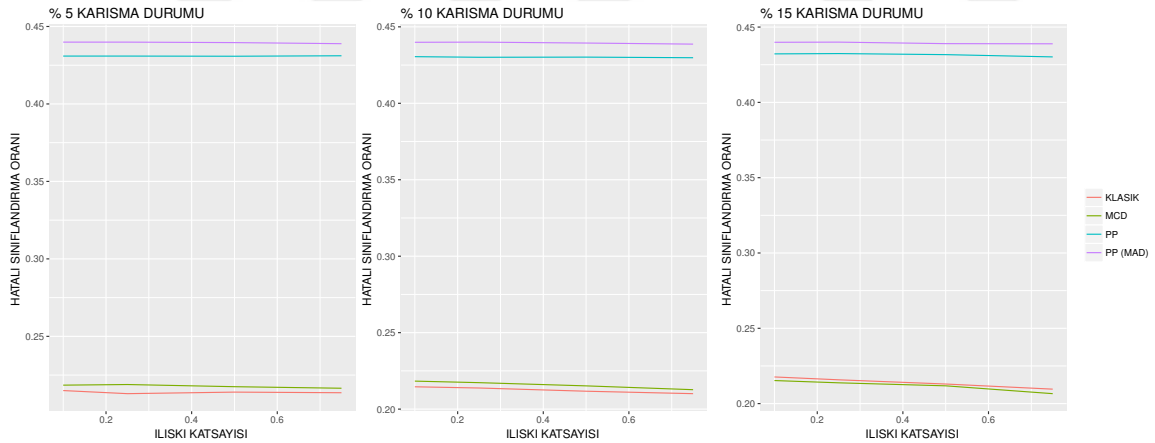
7.4. Durum 4: İlişki Katsayılarına Göre Karışma Durumu

Tesadüfi olarak üretilen iki ve üç sınıflı veri setlerine, ilişki katsayısı $\rho = \{0,10; 0,25; 0,50; 0,75\}$ ve $\varepsilon = \{\%0, \%5, \%10, \%15\}$ oranlarında karışma olan veri setleri için yapılan simülasyon çalışmasında, klasik tahmin edicilerle, *MCD* tahmin edicileri ile , izdüşüm arama yöntemi ile ve *MAD* istatistiği kullanılarak uygulanan izdüşüm arama yöntemi ile diskriminant analizi uygulanması sonucu hesaplanan hatalı sınıflandırma oranları Çizelge 7.4'te verilmiş ve Şekil 7.5'te ve Şekil 7.6'da grafikler ile gösterilmiştir.



Şekil 7.5. İki sınıflı veri seti %5, % 10 ve %15 oranlarında ilişki katsayısı karışması

Şekil 7.5'e göre; tesadüfi olarak üretilen 2 sınıflı, ilişki katsayısı $\rho = \{0,10; 0,25; 0,50; 0,75\}$ ve $\varepsilon = \{\%5, \%10, \%15\}$ oranlarında karışma olan veri setileri için yapılan simülasyon çalışmasında, veri setine $\varepsilon = \{\%5, \%15\}$ oranında konum karışmaları olması durumlarında en düşük hatalı sınıflandırma oranlarının izdüşüm arama yöntemi (PP) ile diskriminant analizi uygulaması sonucu elde edildiği gözlemlenmektedir.



Şekil 7.6. Üç sınıflı veri seti %5, % 10 ve %15 oranlarında ilişki katsayısı karışması

Şekil 7.6'ya göre; tesadüfi olarak üretilen 3 sınıflı, ilişki katsayıı $\rho = \{0,10; 0,25; 0,50; 0,75\}$ ve $\varepsilon = \{\%5, \%10\}$ oranlarında karışma olan veri setileri için yapılan simülasyon çalışmasında, en iyi sonuçların klasik tahmin ediciler ile diskriminant analizi uygulaması sonucu elde edildiği gözlemlenmektedir. %15 karışma olması durumunda ise MCD tahmin edicileri ile diskriminant analizi uygulaması sonucu elde edilen hatalı sınıflandırma oranlarının en düşük olduğu görülmektedir.

Tesadüfi olarak üretilen 2 sınıflı, ilişki katsayısı $\rho = \{0,10;0,25;0,50;0,75\}$ ve $\varepsilon = \{\%0,\%5,\%10,\%15\}$ oranlarında karışma olan veri setileri için yapılan simülasyon çalışmasında elde edilen hatalı sınıflandırma oranlarına göre, tahmin ediciler arasında ciddi bir farklılık olmadığı söylenebilir. Ancak, izdüşüm arama yöntemi ile diskriminant analizinden elde edilen hatalı sınıflandırma oranlarının en düşük olduğu Çizelge 7.4'de görülmektedir. Fakat, ilişki katsayısı $\rho = \{0,50;0,75\}$ ve karışma oranı $\%10$ olduğu durumlar için *MCD* tahmin edicileri ile diskriminant analizi sonucu elde edilen hatalı sınıflandırma oranlarının izdüşüm arama yöntemi ile diskriminant analizinden elde edilen hatalı sınıflandırma oranlarından düşük olduğu görülmektedir.

Tesadüfi olarak üretilen 3 sınıflı, ilişki katsayısı $\rho = \{0,10;0,25;0,50;0,75\}$ ve $\varepsilon = \{\%0,\%5,\%10,\%15\}$ oranlarında karışma olan veri setileri için yapılan simülasyon çalışmasında elde edilen hatalı sınıflandırma oranlarına göre, ilişki katsayısı $\rho = \{0,10;0,25;0,50;0,75\}$ ve karışma oranı $\%15$ iken *MCD* tahmin edicileri ile diskriminant analizi sonucu elde edilen hatalı sınıflandırma oranlarının en düşük olduğu, yani, *MCD* tahmin edicileri ile diskriminant analizinin daha başarılı sonuç verdiği söylenebilir. İlişki katsayısı $\rho = \{0,10;0,25;0,50;0,75\}$ ve $\varepsilon = \{\%5,\%10\}$ karışma oranı için klasik tahmin edicier ile diskriminant analizi uygulamasından elde edilen hatalı sınıflandırma oranlarının daha düşük olduğu görülmektedir.

Çizelge 7.4. Durum 4 - ilişki katsayıları için hatalı sınıflandırma oranları

Sınıf Sayısı	İlişki Katsayısı	Karışma Oranı	Klasik	MCD	PP	$PP_{(MAD)}$
2	0,10	0,00	0,1562	0,1549	0,1553	0,1637
		0,05	0,1557	0,1561	0,1538	0,1646
		0,10	0,1552	0,1553	0,1549	0,1649
		0,15	0,1554	0,1559	0,1546	0,1627
	0,25	0,00	0,1562	0,1549	0,1553	0,1637
		0,05	0,1555	0,1561	0,1549	0,1648
		0,10	0,1558	0,1554	0,1552	0,1653
		0,15	0,1551	0,1555	0,1546	0,1625
	0,50	0,00	0,1562	0,1549	0,1553	0,1637
		0,05	0,1554	0,1562	0,1546	0,1643
		0,10	0,1552	0,1540	0,1551	0,1646
		0,15	0,1550	0,1556	0,1543	0,1633
	0,75	0,00	0,1562	0,1549	0,1553	0,1637
		0,05	0,1554	0,1556	0,1545	0,1635
		0,10	0,1553	0,1544	0,1548	0,1643
		0,15	0,1533	0,1536	0,1523	0,1619
3	0,10	0,00	0,2168	0,2173	0,4303	0,4398
		0,05	0,2150	0,2185	0,4309	0,4399
		0,10	0,2146	0,2183	0,4305	0,4399
		0,15	0,2177	0,2153	0,4322	0,4399
	0,25	0,00	0,2168	0,2173	0,4303	0,4398
		0,05	0,2130	0,2189	0,4309	0,4399
		0,10	0,2138	0,2173	0,4301	0,4400
		0,15	0,2158	0,2138	0,4324	0,4400
	0,50	0,00	0,2168	0,2173	0,4303	0,4398
		0,05	0,2140	0,2175	0,4308	0,4396
		0,10	0,2117	0,2152	0,4302	0,4394
		0,15	0,2130	0,2118	0,4317	0,4390
	0,75	0,00	0,2168	0,2173	0,4303	0,4398
		0,05	0,2136	0,2165	0,4311	0,4389
		0,10	0,2101	0,2127	0,4298	0,4387
		0,15	0,2096	0,2066	0,4302	0,4389

8. SONUÇ VE ÖNERİLER

Bilgisayar teknolojisindeki hızlı gelişmelerle çok yüksek boyutlu bilgiler saklanabilir hale gelmiştir. Ancak, bu yüksek boyutlu bilgilerle faydalı analizler yapılmadığı durumda bilgilerin yalnızca saklanması anlamlı olmamaktadır. Böylesine büyük veri setleri içerisinde başka dağılımlardan karışmış gözlemler yer alabilir. Bu durumlarda, veri seti için yapılan analizler gerçeği yansıtmayabilirler. Bu tip gerçeği yansıtmayan sonuçlardan kaçınmak için, çok değişkenli büyük veri setlerinin analizinde kullanılan, diskriminant analizi ya da temel bileşenler analizi gibi analizlerde, başka dağılımlardan karışmış gözlemlerden etkilenmeyen sağlam tahmin ediciler kullanılabilir.

Bu tez çalışmasında, veri setinde başka dağılımdan karışan gözlemler olması ve olmaması durumları için, klasik tahmin ediciler ile, *MCD* tahmin edicileri ile ve izdüşüm arama yöntemleri ile diskriminant analizi ve temel bileşenler analizi karşılaştırılmıştır. Gerçek veri setleri ile yapılan uygulamalarda, sağlam tahmin yöntemlerinden olan *MCD* tahmin edicileri ile analiz yönteminin ve izdüşüm arama yönteminin diskriminant analizinde ve temel bileşenler analizinde daha güvenilir sonuçlar verdiği görülmüştür. Çeşitli senaryolar için rastgele üretilmiş veri setleri ile simüasyon çalışması yapılmıştır. Rastgele üretilmiş veri setleri ile diskriminant analizi için yapılan simülasyon çalışmasında, iki sınıflı veri seti için, izdüşüm arama yöntemi ile diskriminant analizinin daha iyi sonuçlar verdiği, ikinci en iyi sonucu *MCD* tahmin edicileri ile diskriminant analizinin verdiği tespit edilmiştir. Klasik tahmin ediciler ile diskriminant analizi sonuçlarının ise sağlam tahmin yöntemleri ile diskriminant analizine göre oldukça yüksek olduğu, gözlemlerin dahil olmadıkları sınıflara daha fazla atandığı görülmüştür. Rastgele üretilmiş veri seti ile diskriminant analizi için yapılan simülasyon çalışmasında, üç sınıflı veri seti için, *MCD* tahmin edicilerinin en iyi sonucu verdiği gözlemlenmiştir.

Gerçek veri setleri ile yapılan uygulamalarda ve çeşitli senayolar için rastgele üretilmiş veri setleri ile yapılan simülasyon çalışmalarında, sağlamlık özelliğine sahip tahmin ediciler ile yapılan analizlerin daha başarılı sonuçlar verdiği gözlemlenmiştir.

KAYNAKLAR

1. Albayrak, A. S. (2006). *Uygulamalı Çok Değişkenli İstatistik Teknikleri.* Ankara: Asil Yayın Dağıtım, 309-322.

2. Alpar, C. R. (2011). *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*. (Üçüncü Baskı) Detay Anatolia Akademik Yayıncılık Ltd. Şti.

3. Alpar, C. R. (2013). *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*. (Dördüncü Baskı) Detay Anatolia Akademik Yayıncılık Ltd. Şti., 119-790.

4. Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. (Second Edition). New York: John Wiley Sons, Inc., 6.

5. Croux, C. ve Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis,* 95 (1), 206-226.

6. Daudin, J. J., Duby, C. ve Trecourt, P. (1988). Stability of Principal Component Analysis Studied by the Bootstrap Method. *Statistics*, 19, 241–258.

7. Filzmoser, P., Hron, K. ve Temple, M. (2012). Discriminant Analysis For Compositional Data And Robust Parameter Estimation. *Computational Statistics*, 27, 585–604.

8. Fisher, R. A. (1936). The Use Of Multiple Measurements In Taxonomic Problems. *Annals of Eugenics,* 7, 179-188.

9. Friedman, J. H. (1987). Exploratory Projection Pursuit. *Journal of the American Statistical Association*, 82 (397), 249-266.

10. Gray, H. L., Baek, J., Woodward, W. A., Miller, J. ve Fisk, M. (1996). A bootstrap generalized likelihood ratio test in discriminant analysis. *Computational Statistics & Data Analysis,* 22 (2), 137-158.

11. Gündüz, N. ve Fokoué, E. (2017). Predictive Performances of implicitly and explicitly robust classifiers on high dimensional data. *Communications,* 66 (2), 14-36.

12. Hampel, F. R. (1974). The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, 69 (346), 383-393.

13. Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35 (1), 73-101.

14. Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematics and Statistics Probability*, 1, 221-233.

15. Huber, P. J. (1985). Projection Pursuit. *The Annals of Statistics,* 13 (2), 435-475.

16. Hubert, M. ve Engelen, S. (2004). Fast cross-validation in robust PCA. *COMPSTAT 2004: Proceedings in Computational Statistics*, 989-996.

17. Hubert, M., Rousseeuw, P. J. ve Branden, V. K. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, 47 (1), 64-79.

18. Hubert, M., Rousseeuw, P. J. ve Van Aelst, S. (2008). High-Breakdown Robust Multivariate Methods. *Statistical Science*, 23 (1), 92-119.

19. Johnson, R. A. ve Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* (Sixth edition). New York: Prentice Hall, 9.

20. Krzanowski, W. J. (1982). Mixtures of Continuous and Categorical Variables in Discriminant Analysis: A Hypothesis-Testing Approach. *Biometrics*, 38 (4), 991-1002.

21. Kruskal, J. B. (1969). *Toward a practical method which helps uncover the structure of a set of observations by finding the line transformation which optimizes a new index of condensation.* Milton, RC, Nelder, JA (eds), *Statistical computation;* New York: Academic Press, 427-440.

22. Lachenbruch, P. A. (1975). *Discriminant Analysis*. Hafner Press: A Division of Macmillan Publishing Co., 1-68.

23. Maronna, R. A., Martin, R. D. ve Yohaai, V. J. (2006). *Robust Statistics: Theory and Methods.* (Sixth edition). New York: John Wiley Sons, Inc., 3.

24. Newcomb, S. (1886). A Generalized Theory of the Combination of Observations so as to Obtain the Best Result. *American Journal of Mathematics*, 8(4), 343-366.

25. Pires, A. M. (2003). Robust Linear Discriminant Analysis and the Projection Pursuit Approach. *Developments in Robust Statistics*, 317-329.

26. Pires, A. M. ve Branco, J. A. (2010). Projection-pursuit approach to robust linear discriminant analysis. *Journal of Multivariate Analysis*, 101 (10), 2464–2485.

27. Reaven, G. M. ve Miller, R. G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 16, 17-24.

28. Rousseeuw, P. J. (1984). Least Median Of Squares Regression. *Journal of the American Statistical Association,* 79 (388), 871-880.

29. Rousseeuw, P. J. ve Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223.

30. Rousseeuw, P. J. ve Hubert, M. (2011). Robust Statistics For Outlier Detection. *John Wiley Sons, Inc.*, 1 (1), 73-79.

31. Stern, W. B., Descoeudres, J. P. (1977). X-Ray Fluorescence Analysis of Archaic Greek Pottery. *Archaeometry*, 19 (1), 73–86.

32. Tatlıdil, H. (2002). *Uygulamalı Çok Değişkenli İstatistiksel Analiz*. Ankara: Ziraat Matbaacılık A.Ş., 138-288.

33. Todorov, V. ve Pires, A. M. (2007). Comparative Performance of Several Robust Linear Discriminant Analysis Methods. *REVSTAT-Statistical Journal,* 5 (1), 63-83.

34. Todorov, V. ve Filzmoser, P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, 32 (3), 585–604.

35. Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In Olkin, I.(Ed.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling.* 448–485. Stanford, CA: Stanford University Press.

36. Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33 (1), 1-67.

37. Tukey, J. W. ve Friedman, J. H. (1974). A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*, 23 (9), 881-890.

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : ŞAHİN, Gamze
Uyruğu : T.C.
Doğum tarihi ve yeri : 31.08.1989, Kartal
Medeni hali : Bekar
Telefon : 0 (312) 410 31 15
e-mail : gamze.sahin@dmo.gov.tr

Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Yüksek lisans	Gazi Üniversitesi / İstatistik	Devam Ediyor
Lisans	Hacettepe Üniversitesi / İstatistik	2012
Lise	Kaya Bayazıtoğlu Lisesi	2007

İş Deneyimi

Yıl	Yer	Görev
2013-Halen	DMO Genel Müdürlüğü Bilgi İşlem Daire Başkanlığı	Merkezi Satınalma Uzman Yardımcısı

Yabancı Dil

İngilizce

Yayınlar

-

Hobiler

Halk oyunları, Treaking

GAZİ GELECEKTİR..