



T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

KARMA AYRIŞTIRMA ANALİZİNDE KAYIP GÖZLEM TAHMİN YÖNTEMLERİNİN DEĞERLENDİRİLMESİ

İsmet Kürşat SARI

YÜKSEK LİSANS

İstatistik Anabilim Dalını

Ağustos -2012

KONYA

Her Hakkı Saklıdır

ÖZET

YÜKSEK LİSANS

KARMA AYRIŞTIRMA ANALİZİNDE KAYIP GÖZLEM TAHMİN YÖNTEMLERİNİN DEĞERLENDİRİLMESİ

İsmet Kürşat SARI

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Dr. Murat ERİŞOĞLU

2012, 103 Sayfa

Jüri
Prof. Dr. Hamza EROL
Prof. Dr. Aşır GENÇ
Dr. Murat ERİŞOĞLU

Kayıp veri günümüzde birçok araştırmacının karşılaştığı önemli bir problemdir. Literatürde kayıp verilerin tahmini için birçok yöntem önerilmiştir. Bu çalışmada farklı tip ve hacimlerdeki veri setleri kullanılarak karma ayrıştırma analizinde mevcut kayıp gözlem tahmin yöntemlerinin etkinliği incelenecektir.

Anahtar Kelimeler: Çoklu Atama, Eksik Gözlem, EM Atama, Karma Ayrıştırma Analizi, Ortalama Atama, Regresyon Atama.

ABSTRACT

MS THESIS

EVALUTION OF MISSING VALUE ESTIMATION METHODS FOR MIXTURE DISCRIMINANT ANALYSIS

İsmet Kürşat SARI

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE IN STATISTICS

Advisor: Dr. Murat ERİŞOĞLU

2012, 103 Pages

Jury
Prof. Dr. Hamza EROL
Prof. Dr. Aşır GENÇ
Dr. Murat ERİŞOĞLU

Missing data is a problem that permeates much of resarch biring done today. In the literature many methods have been proposed to estimate missing value. In this study, we present an evaluation of the performance of current imputation methods across a variety of types and sizes of data sets in mixture discriminant analysis (MDA).

Keywords: Multiple Imputation, Missing Value, EM Imputation, Mixture Discriminant Analysis, Mean Imputation, Regression Imputation

ÖNSÖZ

Tezin hazırlanmasında büyük bir emeği olan danışmanın Dr. Murat ERİŞOĞLU'na teşekkür ederim. Öğrencilik yaşamımın her aşamasında bana destek olan Prof. Dr. Aşır GENÇ'e teşekkür ederim. Ayrıca bu tezim hazırlanma aşamasında bana sabır ve sevgisiyle destek olan sevgili eşime ve hayatıma anlam katan çocuklarıma en içten dileklerimle teşekkür ederim.

İsmet Kürşat SARI

KONYA-2012

İÇİNDEKİLER

ÖNSÖZ vi
İÇİNDEKİLER vii
SİMGELER VE KISALTMALAR viii
1. GİRİŞ 1
2. LİTERATÜR ÖZETİ 5
3. EKSİK VERİLERİN İNCELENMESİ VE ATAMA YÖNTEMLERİ 11
3.1. Eksik Veri Kavramı 12
3.2. Eksik Veri Mekanizmaları 20
3.2.1. Tamamıyla rassal olarak kayıp mekanizması (TROK) 20
3.2.2. Rassal olarak kayıp mekanizması (ROK) 21
3.2.3. Rassal olarak kayıp değil Mekanizması (ROKD) 21
3.3. Rassallığın İncelenmesi 23
3.3.1. Bağımsız iki örneklem t testi 23
3.3.2. Korelasyon testi 26
3.3.3. Little'nin MCAR testi 27
3.4. Eksik Veri Sorununun Giderilmesine Yönelik Yöntemler Ve Atama Yöntemleri 29
3.4.1. Liste bazında silme yöntemi(Listwise Method) 30
3.4.2. Çiftler bazında silme yöntemi (Pairwise Method) 33
3.4.3. Atama yöntemleri (Imputation Methods) 36
3.4.3.1. Ortalama değerin atanması (Mean Imputation) 37
3.4.3.2. En benzer birime ya da birimlere benzetme (Hot Deck Imputation). 39
3.4.3.3. Dış kaynaklı atama (Cold Deck Imputation) 41
3.4.3.4. Regresyon atama (Regression Imputation) 41
3.4.3.5. EM (Expactation-Maximization) yöntemi 46
3.4.3.6 Çoklu atama yöntemi (Multiple Imputation) 52
4. KARMA DAĞILIM MODELİNE DAYALI AYRIŞTIRMA ANALİZİ 53
4.1. Karma Dağılım Modelleri 53
4.1.1. Karma dağılım modelleri için temel gösterimler 55
4.1.2. Karma dağılım modelleri için parametrik gösterimler 56
4.1.3. Karma dağılım modelinde parametre tahmini 57
4.2. Karma Dağılım Modellerine Dayalı Ayrıştırma Analizi 62
4.3. Uygulama 64
5. EKSİK GÖZLEM ÇÖZÜM YÖNTEMLERİNİN KARMA AYRIŞTIRMA ANALİZİNDEKİ ETKİNLİĞİN İNCELENMESİ 66
5.1. Cam (Glass) Verisi 66
5.2. İris Verisi 71
5.3. Şarap (Wine) Verisi 75
5.4. Genel Değerlendirme 80
6. SONUÇLAR VE ÖNERİLER 82
6.1. Sonuçlar 82
6.2. Öneriler 84
7. KAYNAKLAR 86
8. EKLER 90
8. ÖZGEÇMİŞ 103

SİMGELER VE KISALTMALAR

Simgeler

$\mathbf{Y}$: $n \times p$ boyutlu ham veri matrisi

n : Örneklem hacmi

p : Değişken sayısı

y_{ij} : $i.$ birim $j.$ değişken için gözlem değeri

$\mathbf{y}_{i.}$: $i.$ birime ait $1 \times p$ boyutlu gözlem vektörü

$\mathbf{M}$: Eksik veri gösterge matrisi

Q_1 : Birinci çeyreklik

Q_3 : Üçüncü çeyreklik

IQR : Çeyrek aralık

$f(\mathbf{M} / \mathbf{Y}, \phi)$: $\mathbf{Y}$ verilmişken $\mathbf{M}$'nin koşullu dağılımı

t_{hesap} : Bağımsız iki örneklem t test istatistiği

$\overline{x}_j$: $j.$ değişken için örneklem ortalaması

s_j^2 : $j.$ değişken için örneklem varyansı

sd : Serbestlik derecesi

α : Anlam düzeyi

r_{xy} : X ve Y değişkenleri arasındaki örneklem Pearson korelasyon katsayısı

χ^2_{MCAR} : Little'nin MCAR testi

$\hat{\mu}_j$: $j.$ alt gruba ait ortalama vektörü

n_j : $j.$ alt grubun gözlem sayısı

$\hat{\Sigma}_j$: $j.$ alt gruptaki ilgili değişkenler için kovaryans matrisinin tahmini

d_{ik} : $i.$ ve $k.$ birimler arasındaki Öklid uzaklığı

$f(x_i; \mathbf{\Psi})$: Karma olasılık fonksiyonu

π_k : $k.$ için bileşen için karma oranı

$f_k(x_i)$: $k.$ alt grupta x_i rassal vektörün olasılık yoğunluk fonksiyonu

$\mathbf{\Psi}$: Karma dağılım modelinde tüm bilinmeyen parametreler vektörü

$\Phi_k(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: Çok değişkenli normal dağılım olasılık yoğunluk fonksiyonu

$L(\boldsymbol{\Psi})$: Olabilirlik fonksiyonu

$\mathbf{z}$: Etiket vektörü

$p_{i,k}$: $\mathbf{x}_i$ gözlem vektörünün $k.$ sınıfa ait olma olasılığı

Kısaltmalar

TROK : Tamamıyla Rassal Olarak Kayıp

ROK : Rassal Olarak Kayıp

ROKD : Rassal Olarak Kayıp Değil

HKO : Hata Kareler Ortalaması

EM : Expactation-Maximization

AIC : Akaike Bilgi Kriteri

BIC : Bayesçi Bilgi Kriteri

MDA : Karma Ayrıştırma Analizi

LDA : Doğrusal Ayrıştırma Analizi

FIML : Tam Bilgili En Çok Olabilirlik

CML : En Çok Olabilirlik Kriteri

CMML : Karma En Çok Olabilirlik Kriteri

1. GİRİŞ

Eksik veri problemi yıllardır araştırmacıları meşgul etmiştir. Çok iyi planlanmış bir araştırmada bile eksik verilere sık sık rastlanmaktadır. Eksik veriler gerek saha araştırmalarında gerekse laboratuar koşullarında yapılan çalışmalarda araştırmacının karşısına çıkabilir. Eksik veriler yapılacak istatistiksel analizlerde önemli problemler yaratmaktadır. Çünkü istatistiksel analizler ve bu analizlerin yapılmasına olanak veren ilgili paket programlar verilerin tam olduğu durumlar için geliştirilmiştir. Eksik veri ya da veriler bir bilgi yokluğunu temsil ederler dolayısıyla bir veri kaybına neden olurlar. Standart istatistiksel yöntemler ve bunların çözümlerini sağlayan paket programlar tam bilgi durumu için düzenlendiklerinden eksik verileri işlem dışı bırakırlar ve dolayısıyla yanlı tahminlere neden olurlar. Ayrıca veri setinde eksik gözlemlerin çıkartılması hesaplanan istatistiklerin kararlılığını düşürmekte ve çalışmanın geçerliliğini ve genellenebilirliğini etkilemektedirler. Araştırmalardaki eksik veri problemi için çeşitli çözüm ve atama yöntemleri geliştirilmiştir.

Birçok araştırmacı eksik veri durumunda onlarca yıldır eksik veri yerine sabit bir değeri atama mantığına dayanan ad-hoc yöntemlerini kullanmışlardır. Ne yazık ki, kayıp veri durumunda kullanılan bu yöntemlerin kullanılabilmesi için birçok ön koşul bulunmaktadır. Ayrıca bu yöntemlerin tahminlerin yanlı olmasında önemli bir etkisi vardır. Bu yöntemler gerektirdiği önkoşullar ve dezavantajlarından dolayı literatürde kullanımı giderek azalmıştır.

Eksik gözlem tahmin yöntemleri üzerine yapılan araştırmalar ile 1970'li yıllarda geliştirilen en çok olabilirlik tahminine dayanan EM atama ve çoklu atama yöntemlerinin, ad-hoc yöntemlerine göre olan avantajlarından dolayı kullanımı yaygınlaşmıştır. Ayrıca bugünde yaygın bir kullanıma sahip olan kayıp veri probleminin teorik çerçevesi Rubin(1976) tarafından tanımlanmıştır. Geleneksel yöntemleri tercih eden araştırmacılar en çok olabilirlik yöntemi ve çoklu atama yönteminin geleneksel yaklaşımlara göre daha az varsayım gerektirmesine karşın daha güçlü parametre tahminler üretmesine itiraz etmektedirler.

Sorulara yanıt vermeme, bulunamama gibi gözlem birimi kaynaklı ya da veri toplama, kayıt ve veri giriş hataları gibi gözlem birimi dışındaki nedenlerden kaynaklanan ve eksik veriye yol açan sürece eksik veri süreci denir. Bazen eksik veri süreciyle karşılaşılabileceği önceden tahmin edilebilir ve araştırma planında gerekli önlemler alınabilir. Ancak özellikle gözlem birimi kaynaklı olarak ortaya çıkan eksik

veri yapıları önceden öngörülemeyebilir. Bu durumda veride eksik değerleri ortaya çıkaran bir yapı olup olmadığının araştırmacı tarafından incelenmesi gerekir.

Eksik gözlem problemi daha öncede ifade edildiği her araştırmacının karşısına çıkabilecek bir problemdir. Tam veri durumu için geliştirilen analizlerin uygulanabilmesi için eksik gözlem probleminin çözümünde kullanılacak yöntemin doğru olarak tespit edilip, uygulanması da araştırmalardan elde edilecek sonuçların güvenirliliği ve genellenebilirliği açısından son derece önemlidir. Bu tez çalışmasında sınıflama ve ayrıştırma problemlerinin çözümünde son dönemde giderek kullanımı yaygınlaşan ve bu alandaki uygulamalardaki en önemli analizlerden biri olan karma ayrıştırma analizinde eksik gözlem tahmin yöntemlerinin etkinliği değerlendirilecektir.

Ayrıştırma analizi, birimlerin grup üyeliklerinin analiz öncesi bilindiği çok boyutlu veri setlerinde birimleri ait oldukları gruplara en az hata ile atamak üzere ayırma fonksiyonlarının oluşturulduğu ve ayırma fonksiyonları yardımı ile sonradan gözlemlenen veya grup üyeliği bilinmeyen birimlerin gruplarının belirlenmeye çalışıldığı çok değişkenli istatistiksel bir analizdir. Ayrıştırma analizinde, önceden tanımlanmış ve yapısı hakkında bilgi sahibi olunan gruplar eğitim verisi olarak adlandırılır. Yeni elde edilen ve hangi gruba ait olduğu test edilmek istenen nesnelerin ya da gözlemlerin oluşturduğu gruplar ise test verisi olarak adlandırılmaktadır. Ayrıştırma analizinde test verisinin hangi gruba ait olduğunu belirlemek için eğitim verileri kullanılarak ayrıştırma fonksiyonu ve karar kuralı belirlenir.

Ayrıştırma analizi uygulamalarından biri olan doğrusal ayrıştırma analizinde ayırma fonksiyonları, çok değişkenli normal dağılım varsayımı ve grupların kovaryans matrislerinin eşitliği varsayımı altında, tahmin değişkenlerinin doğrusal bileşenleri şeklinde elde edilir. Doğrusal ayrıştırma analizinin taşıdığı varsayımlar nedeni ile kullanılabilirliği oldukça zordur. Doğrusal ayrıştırma analizinin taşıdığı bu dezavantajlar nedeni ile farklı yaklaşım geliştirilmiştir. Bu yaklaşımlardan biride karma ayrıştırma analizidir. Karma ayrıştırma analizi özellikle çok değişkenli normal dağılımlarının karmasının kullanıldığı, grupların çok değişkenli normal dağılım varsayımı ve grupların kovaryans matrislerinin eşitliği varsayımı gerektirmeyen, sınıflama ve ayrıştırma problemlerinde oldukça etkili bir istatistiksel bir yöntemdir. Karma ayrıştırma analizinin temel mantığı; grupların modellenmesinde çok değişkenli normal dağılımın yetersiz olması durumunda grupların modellenmesinde çok değişkenli normal dağılımların karmasının kullanılmasıdır.

Karma ayrıştırma analizi de tam veri durumunda uygulanabilir bir analizdir. Çok boyutlu veri setlerinde eksik gözlem bulunması durumunda karma ayrıştırma analizinin uygulanabilmesi için ya eksik gözlem içeren değişken veya birimler veri setinden çıkartılacak ya da eksik gözlemler uygun atama yöntemleri ile doldurularak veri tamamlanmış veri durumuna getirilecektir. Bu noktada en etkili eksik gözlem çözüm yönteminin doğru olarak belirlenmesi oldukça önem arz etmektedir. Bu çalışmada karma ayrıştırma analizi uygulamaları için eksik gözlem atama yöntemlerinin farklı eksik gözlem yüzdesine sahip, farklı birim ve değişken içeren veri setlerindeki performansları değerlendirilecektir. Eksik veri setleri atama yöntemleri kullanılarak tamamlanmış veri durumuna getirilecek ve sonrasında uygulanan karma ayrıştırma analizi ile elde edilen hatalı sınıflandırma yüzdeleri karşılaştırılacaktır. Ayrıca eksik gözlem taşıyan birim ya da değişkenlerin veri setinden çıkartılması durumunda hatalı sınıflandırma yüzdesindeki değişimde incelenerek, veri silme yönteminin karma ayrıştırma analizindeki etkinliği belirlenecektir.

Çalışmanın ikinci bölümünde karma ayrıştırma analizi ve eksik gözlem tahmin yöntemleri ile ilgili literatürde yer alan bazı önemli çalışmalar incelenecektir.

Çalışmanın üçüncü bölümünde öncelikle eksik veri kavramı ve eksik veri süreci üzerinde durulacaktır. Eksik veri sürecini oluşturan eksik veri mekanizmaları tanıtılarak eksik veri probleminin çözümünde göz önünde bulundurulması gereken adımlar anlatıldıktan sonra rassallığın incelenmesinde kullanılan testler anlatılacaktır. Bu bölümde son olarak eksik gözlem probleminin çözümünde liste bazında silme, çiftler bazında silme ve atama yöntemleri uygulamaları ile birlikte detaylı olarak açıklanacaktır.

Çalışmanın dördüncü bölümünde karma dağılım modellerinin temel ve parametrik gösterimleri verilerek çok değişkenli normal dağılımların karmasındaki parametrelerin en çok olabilirlik tahminlerinin elde edilmesinde kullanılan EM algoritması tanımlanacaktır. Daha sonra karma ayrıştırma analizinde karma dağılım modellerinin kullanımı anlatılacak ve örnek bir uygulama gerçekleştirilecektir.

Eksik gözlem tahmin yöntemlerinin karma ayrıştırma analizindeki etkinliğinin inceleneceği beşinci bölümde, literatürde yaygın bir kullanıma sahip olan farklı örneklem hacimlerinde ve farklı veri özellikleri taşıyan cam (glass), iris ve şarap (wine) verileri kullanılacaktır. Bu bölümde öncelikle tam veri durumunda karma ayrıştırma analizi ile hatalı sınıflandırma yüzdeleri belirlenecektir. Tam veri durumunda karma

ayrıştırma analizi gerçekleştirildikten sonra veri setlerinde %5, %10 ve %15 oranında rassal olarak eksik gözlemler oluşturulacaktır.

Eksik gözlemli veri setleri oluşturulduktan sonra, eksik veri problemi liste bazında silme ile ortalama atama, grup ortalamalarının atanması, regresyon atama ve EM atama yöntemleriyle çözümlenerek tamamlanmış veri durumuna getirilecektir. Tamamlanmış veri durumuna getirilen veri setlerinde gerçekleştirilen karma ayrıştırma analizi sonucu elde edilen hatalı sınıflandırma yüzdeleri göz önünde bulundurularak eksik veri probleminin çözümünde farklı eksik gözlem yapılarında hangi yöntemlerin daha iyi sonuçlar verdiği belirlenecektir.

Çalışmanın sonuç ve öneriler bölümünde elde edilen bulgular değerlendirilerek, karma ayrıştırma analizi uygulamalarında eksik veri probleminin çözümünde hangi yöntemlerin kullanılması gerektiği belirtilecektir.

2. LİTERATÜR ÖZETİ

Çalışmanın bu bölümünde kayıp veri probleminde kullanılan yöntemlerle, karma ayrıştırma analizi ile ilgili bazı çalışmalar incelenecektir.

Jackson (1968), değişken sayısının ve birim sayısının çok büyük olduğu bir ayrıştırma analizi için kayıp veri probleminin ele alındığı ampirik çalışma sonucunda kayıp veri probleminde, kayıp gözlem içeren birimlerin veriden silinmesi yönteminin kullanılmaması gerektiğini vurgulamıştır. Çalışmada kayıp veri oranının %10 ve %20 olduğu durum için gerçekleştirilen ayrıştırma analizinde ortalama atama yöntemi ile iteratif regresyon atama yöntemini hatalı sınıflandırma yüzdesi bakımından karşılaştırmıştır. Karşılaştırma sonucunda daha düşük hatalı sınıflandırma yüzdesinin elde edildiği iteratif regresyon atama yönteminin kayıp veri problemi içeren ayrıştırma analizi problemlerinde kullanılmasını önermiştir.

Buck (1996), çok boyutlu verilerin istatistiksel analizinde kayıp gözlemlerin veri içersinden silinmesinin bilgi kaybına yol açacağını ifade ettiği çalışmada kayıp gözlemlerin veriden dışlanması yerine kayıp gözlemlerin regresyon yöntemleri ile tahmin edilmesinin daha faydalı olacağını belirtmiştir. Değişkenler arasında yüksek korelasyonlu değişkenler arasında regresyon yöntemi ile belirlenecek lineer fonksiyonlar yardımıyla kayıp gözlemlerin tahmin edilmesinde varyans-kovaryans matrisinin yansız ve daha tutarlı bir şekilde tahmin edilebileceğini bir örnek veri seti üzerinde göstermiştir.

Singh ve Horn (2000) çalışmalarında, ortalama ve oransal yöntemlerinden yararlanarak yeni bir veri türetme yöntemi ve buna bağlı yeni bir kitle ortalaması tahmin edicisi önermişlerdir. Bu veri türetme yönteminde sadece ilgilenilen değişkenin kayıp verileri için değil aynı zamanda gözlenen verileri için de yardımcı değişkenden faydalanılarak tahmin değerleri verilmiştir. Burada amaç, en küçük hata kareler ortalamasına (HKO) sahip kitle ortalaması tahmin edicisine ulaşmaktır.

Olga ve ark. (2001), gen dizi veri setlerinde kayıp gözlem tahminlerinden tekil değere dönüştürme yöntemine dayalı yerine koyma yöntemi, ağırlıklandırılmış en yakın k komşuya dayalı yerine koyma yöntemi ile sıra ortalama yöntemlerini karşılaştırmışlardır. Veri içersindeki kayıp gözlem oranının %1 ile %20 olması durumunda karşılaştırdıkları üç yöntemden en yakın k komşuya dayalı yerine koyma yönteminin daha etkili sonuçlar verdiğini ortaya koymuşlardır.

Enders. ve Bandalos (2001), Monte Carlo simülasyon yöntemiyle yapısal eşitlik modellerinde, tam bilgili en çok olabilirlik (Full Infotmation Maximum Likelihood, FIML) yöntemi, liste bazında silme, çiftler bazında silme ve benzer kayıp gözlem tahmin yöntemlerinin performansını incelemişlerdir. Üç bağımsız değişkenin etkilerini yakınsama hataları, yanlı parametre tahmini, parametre tahmin etkinliği ve uyum iyiliği kriterlerine göre incelemişlerdir. Sonuç olarak FIML tahmininin, tasarlanan tüm koşullar arasında üstün olduğunu göstermişlerdir. Kayıp gözlem koşulları (rassal olarak kayıp ve tamamen rassal olarak kayıp) göz ardı edildiğinde FIML tahminlerinin yansız ve diğer yöntemlerden daha verimli olduğunu göstermişlerdir. İlaveten, FIML tahmininde yakınsama oranlarının en düşük olduğunu belirtmişlerdir.

Bal (2003), kayıp gözlem tahmin yöntemlerinin performansını simülasyon çalışması ile karşılaştırmıştır. Çalışmada örneklem hacminin düşük olması ($n < 200$) durumunda sıralama, eşleştirme, ortalama, regresyon ve EM atama yöntemlerinin tutarsız olduğunu belirtmiştir. Büyük örneklem hacimlerinde ($n > 200$) özellikle EM atama yönteminin çok iyi sonuçlar verdiğini göstermiştir.

Acuna ve Rodriguez (2004), kayıp veri probleminin çözümü ile ilgili silme yöntemi, ortalama atama, medyan atama ve en benzer birime benzetme yöntemlerinin doğrusal ayrıştırma analizinde (Lineer Discriminant Analysis, LDA) yanlış sınıflandırma oranı üzerindeki etkisini değerlendirmek için on iki veri seti üzerine uygulama yapmışlardır.

Acock (2005) makalesinde, geleneksel yaklaşımları (liste bazında silme, çiftler bazında silme ve ortalama atama) incelendikten sonra tekli atma, çoklu atama ve tam bilgili en çok olabilirlik tahminlerini geleneksel yaklaşımlara alternatif olarak önermiştir. Makalesinde kayıp gözlemlerin etkilerini bir lineer model için göstermiş ve kayıp gözlem probleminin çözümünde, çoklu atama ve tam bilgili en çok olabilirlik tahminlerinin geleneksel yaklaşımlardan daha üstün performans gösterdiğini belirtmiştir. Makalede ayrıca örnekler ile kayıp veri analizinde SPSS, NORM, Stata ve Mplus programlarının kullanımını göstermiştir.

Baygül (2007), kayıp veri analizi yöntemlerini incelediği çalışmasında, üretilen rassal veri setlerinde oluşturulan kayıplar üzerinde bu yöntemleri uygulayarak elde edilen sonuçları gerçek sonuçlar çerçevesinde değerlendirmiştir. Karşılaştırma sonucunda bu yöntemlerin anlamlı bir farklılık göstermediğini, ancak farklı veri setleri, farklı oranda kayıplar, önceki ve sonraki değerler arasındaki anlamlılığın sınıra yakın olması gibi çeşitli durumlarda yine aynı kayıp veri analizi yaklaşımlarının farklılık

gösterebileceğini ifade etmiştir. Çalışma sonucunda, kayıp veri analizi yöntemlerinin veri setinde uygun bir şekilde kullanıldığında anlamlı sonuçlar verdiğini ve özellikle tıp gibi veri toplamanın zahmetli olduğu alanlarda araştırmacılar için çok büyük bir sorun olan kayıp değer sorununa bilimsel çözümler sunduğunu belirtmiştir.

Aik ve Zainuddin (2008), özellikle gen açılım verilerinde kayıp veri problemi ile sıklıkla karşılaşıldığını belirtikleri çalışmalarında kayıp gözlem tahmininde sinir ağı yöntemi ile yerel en küçük kareler yerine koyma yöntemi ve Bayesci temel bileşenler analizine dayalı kayıp gözlem tahmin yöntemini karşılaştırmışlardır. Karşılaştırma kriteri olarak radyal taban fonksiyonunun kullanıldığı çalışmada, yerel en küçük kareler yerine koyma yöntemi ile Bayesci temel bileşenler analizine dayanan tahmin yönteminin daha etkili olduğunu vurgulamışlardır.

Satıcı (2009), kayıp gözlem tahmin yöntemlerinin incelendiği çalışmasında ard arda örnekleme yönteminde kayıp gözlem olması durumunda son araştırma kitle ortalaması tahmini için yeni bir tahmin edici önermiştir. Bu tahmin edicinin minimum hata kareler ortalaması ve optimal yenileme ilkesi kapsamında minimum eşleştirme oranı elde edilmiştir. Önerilen tahmin edici, Singh ve Priyanka (2007) tahmin edicisi ile karşılaştırılmış ve etkin olma koşulu bulunmuştur.

Belen (2009), gizli kayıp veri sorununu çözmek için gömülü yansız örneklem sezgisel yöntemi incelenmiş ve yöntemin eksikliklerini göstermiştir. Çalışmada ki-kare iki örneklem testi üzerine kurulu yeni bir yöntem önerilmiştir. Önerilen yöntemin hiç bir alan bilgisine ihtiyaç duymadığı ve gömülü yansız örneklem sezgisel yönteminden daha iyi performans gösterdiği belirtilmiştir.

Hastie ve Tibshirani (1996) karma ayrıştırma analizinde karma yoğunluk tahminini sınıflandırmada kullanmışlardır. Bu çalışmada ayrıştırma yapmadan önce k-ortalamalar (k-means) algoritması veya benzer algoritmalar yardımı ile veri için başlangıç sınıfları oluşturulmuştur. Daha sonra Expactation-Maximization (EM) algoritması M adımında her bir bileşen için parametrelerin ağırlıklandırılmış maksimum likelihood tahminleri kullanılarak veriye uygulanmıştır. Bu şekilde veri için parametre tahminleri yapıldıktan sonra ayrıştırma fonksiyonu oluşturulmuş ve gözlemleri aitlik olasılıklarının maksimum olduğu kümelere sınıflandırılarak ayrıştırma analizi yapılmıştır.

Ashikaga ve Chang (1981) tarafından iki bileşenli karma normal dağılım modeli altında Fisher' in lineer ayrıştırma fonksiyonunun tutarlılığı çalışılmıştır. Ashikaga ve Chang, parametreleri bilinen iki bileşenli karma normal dağılıma sahip iki kitle için

Fisher' in lineer ayrıştırma fonksiyonunun tutarlılığını belirlemeye çalışmışlardır. Bu çalışmanın sonucunda kitlelerin dağılımı normallikten önemli ölçüde sapma göstermiyorsa ve kitle merkezleri çok fazla ayrık değilse lineer ayrıştırma fonksiyonunun daha tutarlı sonuç verdiği gözlenmiştir. Burada iki kitlenin benzer şekillere sahip olduğu kabul edilmektedir.

Krzanowski (1982) ayrıştırma analizinde sürekli ve kategorik değişkenlerin karmaları için hipotez testi yaklaşımını incelemiştir. Krzanowski, birden fazla sürekli değişken içeren karma dağılım modellerine dayalı ayrıştırma analizi yöntemlerini tüm tipteki değişkenleri içeren karma dağılım modellerine genelleştirmiştir. Çok değişkenli normal dağılıma sahip iki kitlenin oluşturduğu karma dağılım modeli için ayrıştırma kuralı John' un Z istatistiğini içeren bir hipotez testi yaklaşımına dayalıdır. Bu yaklaşım Bayes kuralına (Anderson W istatistiği) ve basit parametre tahmin yöntemine dayalı olduğu için hesaplamalarda çok fazla işlem gerektirmektedir. Son çalışmalar Z istatistiğinin belirgin koşullarda kullanımının daha uygun olacağını göstermektedir. Krzanowski, karma dağılım modeline dayalı yeni bir ayrıştırma kuralı elde etmiştir. Elde ettiği ayrıştırma kuralını kullanılarak yapılan hata oranlarının tahminleri için yeni bir yöntem de ortaya atmıştır.

Erol ve Akdeniz (1996) tarımsal bir bölgedeki farklı alanları sınıflandırmak için tek değişkenli normal dağılımların karmalarına dayalı multispectral (çok bantlı) sınıflandırma algoritması önermişlerdir. Önerilen bu algoritma tek değişkenli durumda farklı iki karma normal dağılım modeline ait olasılık yoğunluk fonksiyonlarının karşılaştırılmasına dayalı bir algoritmadır. Erol ve Akdeniz (1996) uzaktan algılanmış uydu görüntü verisin 3, 4 ve 5. bantlarını kullanarak tarımsal bölgedeki her bir kontrol (eğitim) ve test alanı için tek değişkenli ve üç bileşenden oluşan karma normal dağılım modeli oluşturmuşlardır. Erol ve Akdeniz (1996) önerdikleri çok bantlı sınıflandırma algoritması için ayrıştırma fonksiyonu olarak Hellinger uzaklığını ve karar kuralı olarak ta Hellinger uzaklığının minimum değerini kullanmışlardır.

Biernacki ve Govaert (1999), modele dayalı kümeleme ve ayrıştırma analizinde model seçimi üzerine çalışmışlardır. Celeux ve Govaert (1993) varyans-kovaryans matrisinin özdeğer ayrışımını kullanarak Gaussian (normal dağılım) modeline dayalı kümeleme ve ayrıştırma analizi için uygun modeller oluşturmuşlardır. Bu çalışmada Biernacki ve Govaert, Celeux ve Govaert (1993)'in oluşturduğu bu modelleri seçmek için Monte Carlo simülasyon verisini kullanarak bir çok sınıflandırma kriterinin performanslarını karşılaştırmışlardır. Bu kriterler arasında bilgi kriteri olarak Akaike

bilgi kriteri (AIC) ve Bayesçi bilgi kriteri (BIC) ve sınıflandırma kriteri içinde NEC ve çapraz geçerlilik (cross-validation) kriterleri örnek verilebilir. Çapraz geçerlilik kriteri ayrıştırma analizinde iyi sonuçlar vermektedir. Bilgi kriterleri ve BIC hesaplamalarda zaman kazandırdığı için yeterli sonuçlar vermektedir.

Fraley ve Raftery (2002) tarafından model tabanlı sınıflandırma, yoğunluk tahmini ve ayrıştırma analizi yazılımı MCLUST oluşturulmuştur. MCLUST model tabanlı sınıflandırma, yoğunluk tahmini ve ayrıştırma analizi yapmak için oluşturulmuş bir paket programdır. Bu program normal dağılıma dayalı hiyerarşik sınıflandırma algoritmalarına ve EM algoritmasına bağlı olarak çalışmaktadır. Aynı zamanda bu yazılım hiyerarşik sınıflandırma algoritmasını, EM algoritmasını ve Bayesçi bilgi kriterini birleştiren fonksiyonları içermektedir. Bu fonksiyonlar sınıflandırma, yoğunluk tahmini ve ayrıştırma analizi için kullanılabilecek çok yönlü fonksiyonlardır. MCLUST programı ile sınıflandırma sonuçlarını iki ve üç boyutlu grafikler üzerinde görmek mümkündür.

Ju ve ark. (2003), karma normal dağılıma dayalı ayrıştırma analizi ve uzaktan algılamada bir uygulama gerçekleştirmişlerdir. Karma dağılım modelleri için yapılan analizler, uzaktan algılama ile elde edilen görüntülerin incelenmesinde de kullanılmaktadır. Bu analizler, algılanan görüntünün yapısında bulunan farklılıkları belirlemede önemli rol oynar. Uzaktan algılamada karma normal dağılım modellerinin lineer karma dağılım modellerinden lineer olmayan yapay sinir ağı karma dağılım modellerine kadar birçok uygulama alanı vardır. Ju ve ark. (2003) bu çalışmalarında karma ayrıştırma analizi ile yapay sinir ağlarına dayalı ayrıştırma analizini yöntemlerini karşılaştırmış ve uzaktan algılamada karma ayrıştırma analizinin oldukça etkili sonuçlar verdiğini göstermişlerdir.

Halbe ve Aledjem (2005) modele dayalı karma ayrıştırma analizini kullanarak deneysel bir çalışma yapmışlardır. Halbe ve Aladjem (2005) bu çalışmalarında modele dayalı ayrıştırma analizini gerçek ve simülasyonla elde edilmiş çeşitli veri kümeleri üzerinde incelemişlerdir. Az sayıda gözlem içeren veri kümelerindeki bileşenler için uygun varyans-kovaryans yapıları belirleyerek doğru ayrıştırma ve/veya sınıflandırma oranını artırmışlardır. Ayrıca Halbe ve Aladjem bu çalışmalarında model seçimi için Bayesci bilgi kriterini kullanan yeni bir yöntem önermişlerdir.

Li (2005), çok tabakalı karma modele dayalı sınıflandırma geliştirmiştir. Model tabanlı sınıflandırmada her bir kümenin (sınıfın) yoğunluğu genellikle normal dağılım gibi belirli temel parametrik dağılıma sahip olduğu varsayılır. Özellikle çok değişkenli

veri için bir kümenin dağılımı belirlenirken hangi parametrik dağılımın uygun olacağına karar vermek pratikte zordur. Bunun yanında, her bir sınıf kendi içerisinde birden fazla moda sahip olabilir. Dolayısı ile temel parametrik bir dağılımla doğru olarak modellenmeyebilir. Sonuçta ise genel model çok tabakalı karma normal dağılıma sahip olabilir. Modeli tahmin etmek ve sınıflandırma yapabilmek için sınıflandırma en çok olabilirlik kriterine (CML) ve karma en çok olabilirlik kriterine kriterine (CMML) dayalı algoritmalar geliştirilmiştir. Bunun yanında her bir kümedeki normal dağılıma sahip bileşenlerin sayısını belirlemek için BIC ve ICL-BIC gibi bilgi kriterleri incelenmiştir.

Bashir ve Carter (2005), rank indirgeme yöntemi ile karma ayrıştırma analizini incelemişlerdir. Bu makalede çok sayıda özellik vektörünün olması durumunda çok değişkenli Gaussian karma dağılım modelleri kullanılarak yapılan ayrıştırma analizinde indirgenmiş alt uzaylar çalışılmıştır. Bu çalışma Hastie ve Tibshirani (1996)' nin çalışmalarının genelleştirilmiş halidir. Gaussian karma dağılım modellerinin olması durumlarında rank indirgeme yöntemi ile yapılan ayrıştırma analizi ağırlıklandırılmış k ranklı lineer ayrıştırma analizine eşittir. Çok değişkenli Gaussian modellerinin karmalarındaki indirgenmiş rank çözümleri tam ranklı robust karma dağılım modellerinin çözümlerinden elde edilir. Yeni yönlerdeki sınıflandırmalar robust S-tahmin edicileri kullanarak orijinal koordinatlara dayalı ayrıştırma analizi yaklaşımı ile karşılaştırılmaktadır. Birçok durumda rank indirgemeye dayalı ayrıştırma analizi test verileri için daha iyi sonuç vermektedir. Bununla birlikte kovaryans matrisinin köşegen ve ortak olması durumu için rank indirgemeye dayalı karma ayrıştırma analizi, tam ranklı karma ayrıştırma analizinden sınıflandırmada daha az hata ile daha iyi sonuç vermektedir.

Qiao ve Li (2010) makalelerinde, karma ayrıştırma analizi paradigması altında, yüksek boyutlu veri sınıflandırmak için iki yönlü bir Gauss karma modeli önermişlerdir. Bu model gruplar halinde değişkenlerin bölünmesiyle karma bileşenleri düzenler ve daha sonra aynı olan benzer grupların değişkenleri için parametreleri kısıtlar. Burada değişken grupları önceden belirlenmemiş fakat model tahmininin parçası olarak optimize edilmiştir. Makalede genel bir üstel aileden gelen dağılımların iki yönlü karması için bir boyut indirgeme özelliğini ispatlamışlardır. Sonuç olarak, bazı gerçek veri setlerinde iki yönlü karma grup değişkenleri olmayan karma modelden daha iyi performans gösterdiğini belirtmişlerdir. Ayrıca bir yan ürün olarak, önemli bir boyut indirgeme elde etmişlerdir.

3. EKSİK VERİLERİN İNCELENMESİ VE ATAMA YÖNTEMLERİ

Araştırmalarda yararlanılan veri toplama yöntemlerinin çoğunda, gerek araştırmacıdan, gerek araştırılan birimlerden, gerek araştırmanın planından ya da kontrol altına alınamayan nedenlerden dolayı eksik verilerle karşılaşılabilmektedir. Burada önemli olan eksik verinin nasıl oluştuğu değil, eksik veri probleminin uygun bir çözümünün bulunmasıdır. Uygun çözüm yönteminin ortaya konması da eksik veri kavramının doğru olarak algılanması ve eksik veri sürecinin ayrıntılı olarak incelenmesi ile mümkündür.

Eksik veriler yapılacak olan istatistiksel analizlerde önemli problemler yaratmaktadır. Çünkü istatistiksel analizler ve bu analizlerin yapılmasına olanak veren ilgili paket programlar verilerin tam olduğu durumlar için geliştirilmiştir. Eksik veri ya da veriler bir bilgi yokluğunu temsil ederler dolayısıyla bir veri kaybına neden olurlar. Standart istatistiksel yöntemler ve bunların çözümlerini sağlayan paket programlar tam bilgi durumu için düzenlendiklerinden eksik verileri işlem dışı bırakırlar ve dolayısıyla yanlı tahminlere neden olurlar ayrıca veri setinde eksik gözlemlerin çıkartılması hesaplanan istatistiklerin kararlılığını düşürmekte ve çalışmanın geçerliliğini ve genellenebilirliğini etkilemektedirler.

Araştırmalarda öncelikle eksik veri probleminin ortaya çıkması önlenmelidir. Ancak bazen birimlerin bazen de değişkenlerin doğal yapısından dolayı araştırmalarda tam veriye ulaşmak her zaman mümkün olmamaktadır.

Sorulara yanıt vermeme, bulunamama gibi gözlem birimi kaynaklı ya da veri toplama, kayıt ve veri giriş hataları gibi gözlem birimi dışındaki nedenlerden kaynaklanan ve eksik veriye yol açan sürece eksik veri süreci denir. Bazen eksik veri süreciyle karşılaşılabileceği önceden tahmin edilebilir ve araştırma planında gerekli önlemler alınabilir. Ancak özellikle gözlem birimi kaynaklı olarak ortaya çıkan eksik veri yapıları önceden öngörülemeyebilir. Bu durumda veride eksik değerleri ortaya çıkaran bir yapı olup olmadığının araştırmacı tarafından incelenmesi gerekir. Bu incelemede genellikle (Alpar, 2011);

1-Eksik verilerin gözlemlere rasgele dağılıp dağılmadığı ve belirgin bir yapı oluşturup oluşturmadıkları,

2-Eksik verilerin ne kadar olduğu ya da ne sıklıkta ortaya çıktığı sorularına yanıt aranır.

Eksik veri ile ilgili yapılar söz konusu ve eksik verinin araştırma içerisindeki sayısı herhangi bir önlemi almaya gerektirecek seviyede ise veride eksik veri sürecinin varlığından söz edilebilir. Bu durumda eksik verilerle elde edilen istatistiksel sonuçlar çalışmada kullanılan değişkenlerin eksik veri sürecinden etkilendiği derecede yanlı olur, fakat eksik verinin araştırma sonuçlarına etkisi sonuçlar üzerindeki yanlılığı ile sınırlı değildir. Örneğin eksik veri ile ilişkin problemlerin giderilemediği çalışmalarda, eksik verili gözlemlerin çalışmadan çıkartılması söz konusu olur. Bu durum ise gözlem sayısında ciddi bir azalmaya ve araştırmanın başında yeterli olarak düşünülen örneklemin yetersiz sayıda bir örnekleme dönüşmesine neden olur. İncelenen veri setinde bir veya birden çok değişkenin eksik veri içermesi nedeniyle veri setinden çıkartılması sonuçların istatistiksel kararlılığını, geçerliliğini ve genellenebilirliğini etkileyebilir.

3.1. Eksik Veri Kavramı

$\mathbf{Y} = (y_{ij})$, $n \times p$ boyutunda, eksik değer içermeyen tam birim sayısının değişken sayısından fazla olduğu $(n > p)$ dikdörtgen bir veri setini göstersin. Açık matris notasyonu ile veri matrisi

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots\vdots & \vdots & \vdots\vdots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \vdots & \vdots\vdots & \vdots & \vdots\vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nj} & \cdots & y_{np} \end{bmatrix} \tag{3.1}$$

şeklinde gösterilir. Burada y_{ij} *i.* birimin *j.* değişken için almış olduğu değeri göstermektedir. Veri matrisinde gözlem vektörü,

$$\mathbf{y}_{i.} = \begin{pmatrix} y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \end{pmatrix} \tag{3.2}$$

şeklinde ifade edilir. Gözlem vektörü *i.* birimin tüm değişkenler için almış olduğu değerlerden oluşur.

Eksik veri yapısı içeren veri setlerinde eksik veri gösterge (etiket) matrisi $n \times p$ boyutunda $\mathbf{M} = (m_{ij})$ olarak tanımlanır. Eksik veri gösterge matrisi oluşturulurken birimlerin gözlemlenmiş değerleri için 1, gözlemlenmemiş değerleri için 0 değeri verilerek oluşturulur. Örneğin beş birimin üç değişken bakımından gözlemlenmesi ile oluşmuş veri matrisinde eksik veri gösterge matrisi

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

şeklinde tanımlansın. **M** gösterge matrisine göre ilk iki birimin ikinci değişken için, üçüncü birimin birinci değişken için ve beşinci birimin birinci ve ikinci değişken için değerleri gözlemlenememiştir. **M** gösterge matrisine göre tüm birimlerin üçüncü değişken için değerleri gözlemlenmiştir.

Veri setindeki eksik değerlerin giderilmesi öncesinde;

i. Eksik veri çeşitleri
ii. Eksik veri yapısının incelenmesi
iii. Eksik veri setinde rasgeleliğin araştırılması
iv. Eksik veri sorununun çözümlenmesi

adımlarının takip edilmesi gerekmektedir (Bal, 2003).

Eksik veri içeren araştırmalarda öncelikle eksik veri yapısının incelenmesi gerekir. Bu inceleme ile eksik verilerin veri setinde hangi değişkenlerde ve ne miktarlarda yoğunlaştığı, bir değişkende gözlemlenemeyen eksik verinin bir diğer değişken ya da değişkenlerden etkilenip etkilenmediğini gibi birçok durumun araştırılması gerekir.

Eksik veri yapılarının incelenmesi ve eksik veri problemine çözüm aramak amacıyla oluşturulan eksik gözlem içeren örnek veri seti Çizelge 3.1'de verilmiştir.

Çizelge 3.1. Eksik gözlem içeren örnek veri seti

Gözlem No	Y_1	Y_2	Y_3	Y_4	Y_5
1	2	-	-	-	5
2	1	5889	-	-	5
3	23	766	6280	-	17
4	21	832	4210	-	-
5	10	1653	3940	-	46
6	9	-	9260	-	13
7	7	2005	14790	-	7
8	3	4418	19380	13	4
9	1	3636	-	22	5
10	5	3744	15730	27	7
11	3	6554	-	26	4
12	2	4537	-	20	5
13	17	-	3550	25	32
14	1	4187	31680	17	5
15	4	2794	24550	10	-
16	4	2757	16060	26	8
17	6	12179	33660	27	3
18	2	3647	32730	35	-
19	2	3169	28030	8	4
20	4	1946	12880	5	10
21	5	2543	13860	6	9
22	13	1347	7510	1	16
23	20	944	2530	4	20
24	2	5048	-	30	6
25	2	-	-	17	-
26	5	2429	13480	4	7
27	3	2574	-	9	5
28	10	1775	9280	3	17
29	6	4586	11570	8	-
30	4	3503	15510	7	8
31	3	3629	-	5	4
32	3	3340	-	7	5
33	1	5784	-	17	4
34	1	3630	-	22	5
35	11	-	10850	2	29
36	10	3041	5520	4	-
37	1	3894	-	28	6

Çizelge 3.1'de yer alan örnek veri seti için eksik gözlemlerin incelenmesinde öncelikle örnek veri setine ait tanımlayıcı istatistikler değerlendirilir. Çoğu araştırmacı tarafından yaygın bir kullanıma sahip olan SPSS istatistik paket programında "Missing Value Analysis" seçeneği ile gerçekleştirilen analizden elde edilen tanımlayıcı istatistikler Çizelge 3.2'de verilmiştir.

Çizelge 3.2. Eksik gözlem içeren örnek veri seti için tanımlayıcı istatistikler

			Standart	Eksik Gözlem		Aşırı Gözlem Sayısı*	
Değişken	***n***	**Ortalama**	**Sapma**	**Sayı**	**%**	**Küçük**	**Büyük**
Y_1	37	6.1351	5.94494	0	.0	0	3
Y_2	32	3524.3750	2147.21352	5	13.5	0	1
Y_3	24	14451.6667	9477.20637	13	35.1	0	0
Y_4	30	14.5000	10.16332	7	18.9	0	0
Y_5	31	10.3548	9.79302	6	16.2	0	3

(*) $(Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR)$ sınırlarının dışına çıkan gözlem sayısı

Eksik gözlemler için elde edilen tanımlayıcı istatistikler tablosunda, her bir değişken için gözlemlenmiş birim sayısı, gözlemlenmiş değerlerin ortalamaları ve standart sapmaları, her bir değişkende gözlemlenemeyen birim sayısı ve eksik gözlemlerin toplam birim içersindeki yüzdesi yer alır. Ayrıca her bir değişken için Tukey'in kutu grafiği ölçütüne göre belirlenen küçük ve büyük aşırı değerlerin sayısı belirtilmiştir. Tukey'in kutu grafiği ölçütüne göre sınırlar,

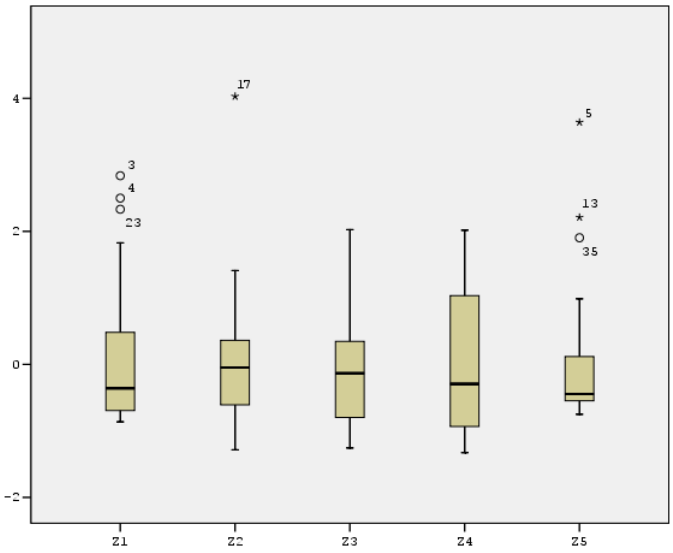
$$(Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR) \tag{3.3}$$

şeklinde tanımlanır. Eşitlikte yer alan Q_1 ve Q_3 sırasıyla birinci çeyreği ve üçüncü çeyreği gösterirken IQR çeyrek aralığı ifade eder ve,

$$IQR = Q_3 - Q_1 \tag{3.4}$$

eşitliği ile elde edilir. Örneğimizde değişkenlerin yapısını aynı anda görebilmek için veriler standartlaştırılmış ve standartlaştırılmış veriler için kutu-çizgi grafiği Şekil 3.1'de verilmiştir.

Şekil 3.1'e göre Y_1 değişkeninde üçüncü, dördüncü ve yirmi üçüncü birimlerin gözlem değerleri aşırı büyük değerlerdir. Aynı şekilde Y_2 değişkeninde on yedinci birimin gözlem değeri aşırı büyük değer iken Y_5 değişkeninde beşinci, on üçüncü ve otuz beşinci birimin gözlem değerleri aşırı büyük değerlerdir.



Şekil 3.1. Örnek veri setinde standartlaştırılmış değerler için kutu-çizgi grafiği

Çizelge 3.2 incelendiğinde Y_1 değişkeninin 37 birim için değeri gözlemlenmiş ve bu değerlerin ortalama ve standart sapması sırasıyla 6.1351 ve 5.94494 bulunmuştur. Y_1 değişkeninde gözlemlenemeyen birim bulunmazken 3 birimin gözlem değeri aşırı büyük değer olarak tanımlanmıştır. Örnek veri seti içersinde yer alan Y_3 değişkenin ise 24 birim için değeri gözlemlenmiş ve bu değerlerin ortalaması ve standart sapması sırasıyla 1451.6667 ve 9477.20637 olarak hesaplanmıştır. Y_3 değişkeninde gözlemlenemeyen 13 birim bulunurken bunların 37 gözlem içersindeki yüzdesi %35.1 olarak belirlenmiştir. Y_3 değişkeninde aşırı küçük ya da büyük gözlem bulunmamaktadır.

Çizelge 3.2'ye göre Y_3 değişkeni 13 eksik gözlem ile veri seti içersinde en fazla eksik gözlem barındıran değişken olmuştur. Y_3 değişkenini %18.9, %16.2 ve %13.5 eksik gözlem barındırma yüzdeleri ile sırasıyla Y_4, Y_5 ve Y_2 değişkenleri takip etmiştir. Araştırmalarda bir değişkenin gözlemlenemeyen birim sayısının toplam içersindeki yüzdesi %15'den fazla ve değişkenin çalışmadan çıkartılmasının çalışmanın yapısını bozmayacağı düşünülüyor ise değişkenin çalışmadan çıkartılması tavsiye edilmektedir. Buna göre çalışmanın yapısını bozmama varsayımı altında sırasıyla Y_3, Y_4 ve Y_5 değişkenlerinden biri veya birkaçı çalışmadan çıkartılabilir.

Eksik gözlemlerin incelenmesinde yararlanılan bir diğer çizelge ise eksik veri ve aşırı değer yapılarıdır. Bu çizelgede tam gözlemler **boşluk** ile, eksik gözlemler "**S**" ile temsil edilirken aşırı değerler ise "+" ile temsil edilir. Örnek veri seti için eksik veri yapısı Çizelge 3.3'de verilmiştir.

Çizelge 3.3. Eksik gözlem içeren örnek veri seti için eksik veri yapısı

	Eksik Veri İçin		**Eksik Veri ve Aşırı Değer Yapıları**					**Değişken Değerleri**				
Göz. No	**Sayı**	**%**	Y_1	Y_2	Y_3	Y_4	Y_5	Y_1	Y_2	Y_3	Y_4	Y_5
1	**3**	**60.0**		**S**	**S**	**S**		**2**	**-**	**-**	**-**	**5**
2	2	40.0			S	S		1	5889	-	-	5
3	1	20.0	+			S		**23**	766	6280	-	17
4	2	40.0	+			S	S	**21**	832	4210	-	-
5	1	20.0				S	+	10	1653	3940	-	**46**
6	2	40.0		S		S		9	-	9260	-	13
7	1	20.0				S		7	2005	14790	-	7
8	0	0						3	4418	19380	13	4
9	1	20.0			S			1	3636	-	22	5
10	0	0						5	3744	15730	27	7
11	1	20.0			S			3	6554	-	26	4
12	1	20.0			S			2	4537	-	20	5
13	1	20.0		S			+	17	-	3550	25	**32**
14	0	0						1	4187	31680	17	5
15	1	20.0					S	4	2794	24550	10	-
16	0	0						4	2757	16060	26	8
17	0	0		+				6	**12179**	33660	27	3
18	1	20.0					S	2	3647	32730	35	-
19	0	0						2	3169	28030	8	4
20	0	0						4	1946	12880	5	10
21	0	0						5	2543	13860	6	9
22	0	0						13	1347	7510	1	16
23	0	0	+					**20**	944	2530	4	20
24	1	20.0			S			2	5048	-	30	6
25	**3**	60.0		S	S		S	2	-	-	17	-
26	0	0						5	2429	13480	4	7
27	1	20.0			S			3	2574	-	9	5
28	0	0						10	1775	9280	3	17
29	1	20.0					S	6	4586	11570	8	-
30	0	0						4	3503	15510	7	8
31	1	20.0			S			3	3629	-	5	4
32	1	20.0			S			3	3340	-	7	5
33	1	20.0			S			1	5784	-	17	4
34	1	20.0			S			1	3630	-	22	5
35	1	20.0		S			+	11	-	10850	2	**29**
36	1	20.0					S	10	3041	5520	4	-
37	1	20.0			S			1	3894	-	28	6

Çizelge 3.3'deki eksik veri yapısı tablosunda her bir birim için hangi değişkenlerin değerlerinin gözlemlenip gözlenemediği, gözlenemeyen değişkenlerin sayısı ve yüzdesi ile aşırı değerli gözlemler belirlenebilmektedir. Buna göre birinci gözlem biriminde 3 değişkenin değeri gözlemlenememiştir. Birinci gözlem biriminde değerleri gözlemlenemeyen değişkenlerin tüm değişkenler içersindeki yüzdesi %60'dır. Birinci gözlem biriminde değerleri gözlemlenemeyen değişkenler Y_2, Y_3 ve Y_4 değişkenleridir. Birinci gözlem biriminde Y_1 ve Y_5 değişkeninin değerleri sırasıyla 2 ve 5 olarak gözlemlenmiştir. Tüm gözlem birimleri içersinde en fazla eksik gözleme sahip olan birimler üçer eksik gözlemle 1. ve 25. gözlem birimleridir. Eksik veri yapısına göre Y_1 değişkeninde aşırı gözlem değerlerine sahip birimler 3., 4. ve 23. birimlerdir. Y_1 değişkeninde aşırı gözlem değerleri sırasıyla 23, 21 ve 20'dir. Y_2 değişkenindeki aşırı gözlem değeri, 17. gözlem birimine ait 12179 değeridir. Y_5 değişkenindeki aşırı gözlem değerleri ise 5., 13. ve 35. gözlem birimlerine ait sırasıyla 46, 32 ve 29 değerleridir.

Eksik veri probleminin çözümde eksik veri yapısına göre eksik gözlemlerin yoğunlaştığı birimler tespit edilerek eksik veri yapısı uygun olma şartı ile çalışmadan çıkartılabilir. Bu yaklaşıma göre, tüm gözlem birimleri içersinde en fazla eksik gözleme sahip 1. ve 25. gözlem birimleri çalışmadan çıkartılabilir.

Eksik gözlemlerin incelenmesinde oluşturulan tablolaştırılmış eksik veri yapılarında değişkenlerin tek başına veya birlikte eksik gözlem barındırdığı durum sayısı, ilgili değişkenlerin silinmesi durumunda tam veri sayısı ve ilgili özellikteki değişkenlerin ortalamaları yer alır. Örnek veri seti için oluşturulmuş tablolaştırılmış eksik veri yapıları Çizelge 3.4'de verilmiştir.

Çizelge 3.4. Örnek veri seti için tablolaştırılmış eksik veri yapıları

Göz. Sayısı	Eksik Veri Yapısı					Değişkenler Silindiğinde	İlgi Özellikteki Birimler İçin Değişken Ortalamaları				
	Y_1	Y_2	Y_4	Y_5	Y_3	Göz. Sayısı	Y_1	Y_2	Y_3	Y_4	Y_5
13						**13**	**6.31**	**3457.00**	**16891.54**	**11.38**	**9.08**
3			X			16	13.33	1474.67	8336.67	.	23.33
1			X	X		21	21.00	832.00	4210.00	.	.
4				**X**		17	**5.50**	**3517.00**	**18592.50**	**14.25**	.
10					X	23	2.00	4262.60	.	18.60	4.90
1			X		X	27	1.00	5889.00	.	.	5.00
1		**X**	**X**		**X**	**31**	**2.00**	.	.	.	**5.00**
1		X	X			19	9.00	.	9260.00	.	13.00
2		X				15	14.00	.	7200.00	13.50	30.50
1		X		X	X	30	2.00	.	.	17.00	.

Çizelge 3.4'e göre tüm değişkenler için değerleri gözlemlenmiş 13 birim bulunmaktadır. Tüm değişkenler için gözlem değerleri olan birimler göz önünde bulundurularak hesaplanan ortalamalara tablonun ilk satırının son bölümünde verilmiştir. Y_2, Y_4 ve Y_3 değişken değerlerinin gözlenemediği 1 birim bulunmaktadır ve çalışmadan Y_2, Y_4 ve Y_3 değişkenleri çıkartıldığında tam gözleme sahip 31 birimlik bir veri elde edilecektir. Y_2, Y_4 ve Y_3 değişken değerlerinin gözlenemediği birimin Y_1 ve Y_5 değişkenleri için almış olduğu değerler sırasıyla 2 ve 5'dir. Sadece Y_5 değişkeninin değerinin gözlemlenemeyip, diğer değişkenlerin değerinin gözlemlendiği 4 gözlem birimi bulunmaktadır. Çalışmadan Y_5 değişkeni çıkartıldığında tam gözleme sahip 17 birim elde edilecektir. İlgili özelliği sağlayan 4 birimin Y_1, Y_2, Y_3 ve Y_4 değişkenleri için gözlemlenmiş değerlerinin ortalaması sırasıyla 5.5, 3517, 18592.5 ve 14.25'dir.

Eksik gözlemlerin incelenmesinde yararlanılan bir diğer çizelge ise eksik gözlem gösterge değişkenlerinin uyuşma yüzdeleri çizelgesidir. Örnek veri seti için oluşturulan eksik gözlem gösterge değişkenlerinin uyuşma yüzdeleri Çizelge 3.5'de verilmiştir.

Çizelge 3.5. Eksik gözlem gösterge değişkenlerinin uyuşma yüzdeleri

Değişkenler	Y_2	Y_4	Y_5	Y_3
Y_2	**13.51**			
Y_4	21.62	18.92		
Y_5	24.32	29.73	16.22	
Y_3	37.84	43.24	**45.95**	35.14

Bu çizelgede esas köşegen üzerinde yer alan yüzdeler ilgili değişkendeki eksik gözlemlerin yüzdesini gösterirken, esas köşegen dışındaki yüzdeler ise ilgili değişken çifti için eksik gözlem gösterge değişkeninin aynı olduğu durumların yüzdesini yani uyuşma yüzdelerini verir. Çizelge 3.5 incelendiğinde Y_2 değişkenindeki eksik gözlemlerin yüzdesi %13.51 iken (Y_3, Y_5) değişken için birlikte değerlerin gözlemlendiği veya gözlenemediği birimlerin tüm birimler içersindeki yüzdesi %45.95'dir.

3.2. Eksik Veri Mekanizmaları

Eksik veri probleminin çözümünde en önemli aşama, eksik veri mekanizmalarının açıklanmasıdır. Çünkü hangi çözüm ve atama yönteminin uygun olabileceği eksik veri mekanizmalarına bağlıdır (Baygül, 2007).

Eksik verilerin analizinde mekanizmanın rolü Rubin'in (1976) teorisinde formulize edilinceye kadar araştırmacılar tarafından ihmal edilmiştir. Rubin'in teorisinde eksik veri göstergeleri rassal değişkenler olarak işlenmiş ve bir dağılıma atanmıştır. Little ve Rubin bu mekanizmaları 3 temel mekanizmaya ayırmıştır;

1- Tamamıyla Rassal Olarak Kayıp (**Missing Completely At Random, MCAR**)
2- Rassal Olarak Kayıp (**Missing At Random, MAR)**
3- Rassal Olarak Kayıp Değil (**Not Missing At Random, NMAR)**

3.2.1. Tamamıyla rassal olarak kayıp mekanizması (TROK)

X ve Y gibi iki değişken ele alındığında, Y değişkenine cevap alınamama olasılığı X değişkeni ile ilişkili değil ise yani başka bir anlatımla Y'deki eksik veri Y değişkeninin kendi yapısından kaynaklanıyorsa bu tür eksik veri mekanizmaları tamamıyla rassal olarak kayıp mekanizması (TROK) olarak tanımlanır (Little ve Rubin, 1987).

Daha önce tanımlandığı gibi, $\mathbf{Y} = (y_{ij})$, $n \times p$ boyutlu veri setini ve $\mathbf{M} = (m_{ij})$ eksik gözlem gösterge matrisini göstersin. Eksik veri mekanizması $\mathbf{Y}$ verilmişken $\mathbf{M}$'nin koşullu dağılımıdır ve $f(\mathbf{M}/\mathbf{Y},\phi)$ şeklinde gösterilir. Burada ϕ bilinmeyen parametreleri göstermektedir. Eğer eksiklik $\mathbf{Y}$'nin değerlerine bağlı değilse,

$$f(\mathbf{M}/\mathbf{Y},\phi) = f(\mathbf{M}/\phi) \qquad \text{tüm } Y,\phi\text{'ler için} \tag{3.5}$$

şeklinde gösterilir ve veriler TROK olarak adlandırılır. Burada dikkat edilmesi gereken varsayım yapının kendisinin rassal olduğu anlamına gelmesidir. Fakat, eksik olma durumu verilerin değerlerine bağlı değildir. Eksik veri, veri setinde bulunan herhangi gözlenen veya gözlenmeyen değerlere bağlı değildir.

TROK koşullarının sağlanıp sağlanmadığı eksik gözlemlerle tamamlanmış gözlemler arasındaki karşılaştırma ile belirlenebilir. TROK sağlanıyorsa tam gözlemlerin kullanımını amaçlayan liste bazında silme veya birim bazında silme yöntemleri iyi bir seçim oluşturacaktır.

Bu yöntemlerin avantajı basitliği ve hesaplama süresinin kısa olmasıdır. Eğer eksik veri mekanizması TROK özelliği taşımıyorsa bu yöntemlerin kullanılması durumunda elde edilen çözümler yanlı olacaktır. TROK özelliği taşımayan eksik veri mekanizmalarında robust yöntemler çözümleme için uygun olmaktadır.

3.2.2. Rassal olarak kayıp mekanizması (ROK)

Kayıp veri olasılığı araştırma konusuna bağlı ise bu kayıp veri türü rassal olarak kayıp (ROK) olarak adlandırılır. Tek taraflı bir bağımsızlık söz konusudur. Yani, X ve Y gibi iki değişken alındığında, X değişkenindeki cevapsızlık olasılığı Y değişkenine bağlı iken, Y değişkenindeki cevapsızlık olasılığı X değişkenine bağlı değildir. $\mathbf{Y}_{göz}$, $\mathbf{Y}$ veri setinde gözlenen kısmı, $\mathbf{Y}_{eks}$, $\mathbf{Y}$ veri setinde eksik olan kısmı göstermek üzere,

$$f(\mathbf{M}/\mathbf{Y},\phi) = f(\mathbf{M}/\mathbf{Y}_{göz},\phi) \qquad \text{tüm } \mathbf{Y}_{eks},\phi\text{'ler için} \tag{3.6}$$

olacaktır. ROK'ta eksik olma durumu eksik veri ile ilişkili değildir, fakat diğer gözlenen değişkenlerle ilişkilidir. Eksik veri mekanizması ROK olarak tanımlandığında EM, regresyon ve çoklu atama yöntemleri eksik veri probleminin çözümünde uygun atama yöntemleridir.

3.2.3. Rassal olarak kayıp değil mekanizması (ROKD)

X ve Y gibi iki değişken ele alındığında cevap olasılığını her iki değişkene de bağlı olması mümkün ise eksik veri rassal olarak kayıp değil (ROKD) kabul edilir, diğer bir ifade ile eksik olma rassal değildir ve eksik veri diğer değişkenleri kullanarak tahmin edilemez. Eksik olma $\mathbf{Y}$ veri setinde hem gözlenen kısma hem de eksik olan kısma bağlı ise

$$f(\mathbf{M}/\mathbf{Y},\phi) = f(\mathbf{M}/\mathbf{Y}_{göz},\mathbf{Y}_{eks},\phi) \qquad \text{tüm } \mathbf{Y}_{göz}\ \mathbf{Y}_{eks},\phi\text{'ler için} \tag{3.7}$$

olacaktır. Atama yöntemlerinin tümü ROKD durumunda yanlı olurlar.

Eksik veri mekanizmalarını daha iyi anlayabilmek için IQ ve iş performans değişkeni değerlerinden oluşan bir örnek veri setinde (Craig, 2010) ilgili mekanizma yapılarının oluşturulduğu Çizelge 3.6'yı inceleyelim.

Çizelge 3.6. Eksik veri mekanizmalarının gösterimi için örnek veri seti

IQ	İş Performansı			
	Tam Veri	**TROK**	**ROK**	**ROKD**
78	9	-	-	9
84	13	13	-	13
84	10	-	-	10
85	8	8	-	-
87	7	7	-	-
91	7	7	7	-
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	-	7	-
99	7	7	7	-
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	-	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	-	12	12

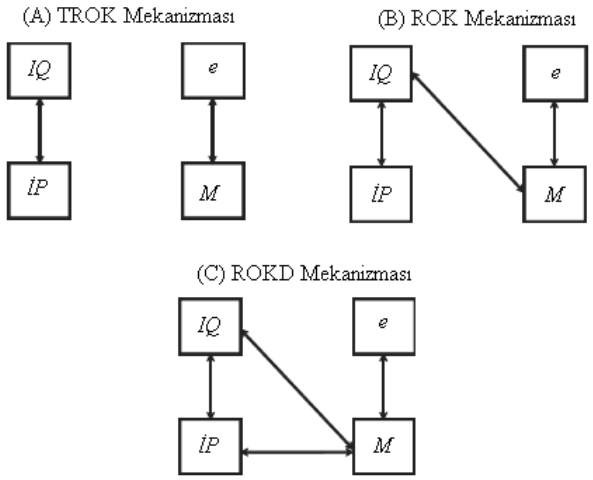
Çizelge 3.6'da yer alan TROK yapısı incelendiğinde, iş performansı değişkenindeki eksik gözlemlerin IQ değişkeninin hem büyük değerleri hem de küçük değerleri için oluştuğu dolayısıyla IQ değişkeninin gözlemlenmiş değerlerinden etkilenmediği açıkça görülmektedir. İş performansı değişkeninde kayıp olan değerlerin, ayrıca iş performansının hem düşük değerlerinde hem de yüksek değerleri içinde oluştuğu görülmektedir. Dolayısıyla TROK yapısında eksik gözlemler veri setinde bulunan herhangi gözlenen veya gözlenmeyen değerlere bağlı değildir (Craig, 2010).

Çizelge 3.6'da yer alan ROK yapısı incelendiğinde iş performansı değişkenindeki kayıp gözlemlerin, IQ değişkeninin düşük değerlerine karşılık geldiği ancak iş performansındaki değerlere bağlı olmadığı görülmektedir.

Çizelge 3.6'da yer alan ROKD yapısı incelendiğinde iş performansı değişkenindeki kayıp gözlemlerin, IQ değişkeninin düşük değerlerine karşılık geldiği ve beraberinde iş performansındaki değerlerin düşük olduğu gözlemlerin kayıp olduğu

görülmektedir. Dolayısıyla eksik gözlemler, hem gözlemlenen hem de gözlemlenemeyen değerlere bağlıdır ve rassal değildir.

Rubin'in eksik veri mekanizmalarını açıklanmasında oluşturduğu grafiksel gösterim Şekil 3.2'de verilmiştir (Rubin, 1976).



Şekil 3.2. Rubin'in kayıp veri mekanizmaları için grafiksel gösterimi

Şekil 3.2' de *IQ* ve *İP* gösterimi ile sırasıyla IQ puanları ve iş performansı skorları, *M* eksik gözlem gösterge matrisini ve *e* gösterimi ile de çalışmaya dahil edilemeyen değişkenleri dolayısıyla rassallığın etkisi temsil edilmektedir.

3.3. Rassallığın İncelenmesi

Rassallığın incelenmesinde yaygın olarak bağımsız iki örneklem t testi, Pearson korelasyon katsayısı ve Little (1988)'ın önerdiği MCAR testi kullanılmaktadır.

3.3.1. Bağımsız iki örneklem t testi

Rassallığın incelenmesinde, bağımsız iki örneklem t testi ile eksik gözlemlerin diğer bir değişkenden etkilenip etkilenmediği test edilir. Bu test öncesinde eksik gözlemlerin bulunduğu birimlere ait diğer değişkenin değerleri birinci grup, eksik gözlemlerin bulunduğu değişkende gözlemlenen birimlere ait diğer değişken değerleri ise ikinci grup olarak ele alınır ve aralarında bağımsız iki örneklem t-testi gerçekleştirilir. Test sonucunda iki grup ortalaması arasındaki fark istatistiksel olarak

anlamlı bulunursa, eksik gözlemlerin etkisi incelenen değişkenden kaynaklandığı ve eksik veri sürecin rassal olmadığına karar verilir.

Oluşturulan iki grup ortalaması arasındaki farkın test edilmesinde kullanılacak test istatistiği,

$$t_{hesap} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.8}$$

ve $\overline{x}_2$ sırasıyla

eşitliği ile hesaplanır. Eşitlikte yer alan $\overline{x}_1$, eksik gözlemlerin bulunduğu birimlere ait etkisi incelenecek değişkenin değerlerinin ortalamasını, $\overline{x}_2$ ise tam gözlemlere karşılık gelen birimlere ait etkisi incelenecek değişkenin değerlerinin ortalamasını gösterir. Eşitlikte yer alan s_1^2 ve s_2^2 gösterimleri sırasıyla eksik gözlemlerin bulunduğu birimlere ait etkisi incelenecek değişken değerlerinin örneklem varyansını, tam gözlemlere karşılık gelen birimlere ait etkisi incelenecek değişken değerlerinin örneklem varyansını gösterir. Benzer şekilde eşitlikte yer alan n_1 ve n_2 gösterimleri de sırasıyla eksik gözlemlere karşılık gelen birim sayısı ile eksik gözlemin bulunduğu değişkendeki gözlemlenen birimlerin sayısını göstermektedir. Elde edilen test istatistiği α anlam düzeyinde,

$$sd = \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{(s_1^2 / n_1)^2 / (n_1 - 1) + (s_2^2 / n_2)^2 / (n_2 - 1)} \tag{3.9}$$

eşitliği ile hesaplanan serbestlik derecesine göre belirlenen Student's t tablo değeri ile karşılaştırılır. Eğer hesaplanan test istatistiğinin değeri tablo değerinden büyük ise yani $t_{hesap} \geq t_{tablo}$ ise ilgili değişkenin eksik gözlem oluşumunda anlamlı bir etkiye sahip olduğuna (1- α) güvenirliliği ile karar verilir.

Çizelge 3.1'de verilen örnek veri seti için eksik gözlemlerin hangi değişken ya da değişkenlerin etkisinin tespiti için gerçekleştirilen bağımsız iki örneklem t testi sonuçları Çizelge 3.7'de verilmiştir.

Çizelge 3.7. Eksik gözlemlerin rassallığının tespitinde bağımsız iki örneklem t testi için sonuçları

		Y_1	Y_2	Y_3	Y_4	Y_5
Y_2	t_{hesap}	-0.8	.	2.4	0.0	-1.6
	sd	5.1	.	6.8	2.4	3.4
	p	0.467	.	0.046	**0.981**	0.193
	n_1	32	32	21	27	27
	n_2	5	0	3	3	4
	$\bar{x}_1$	5.8125	3524.3750	15389.5238	14.4815	8.9630
	$\bar{x}_2$	8.2000	.	7886.6667	14.6667	19.7500
Y_3	t_{hesap}	5.0	-2.1	.	-1.8	3.4
	sd	24.6	30.0	.	24.9	18.2
	p	**0.000**	0.044	.	0.088	0.003
	n_1	24	21	24	19	19
	n_2	13	11	0	11	12
	$\bar{x}_1$	8.4167	3060.2381	14451.6667	12.2105	13.7895
	$\bar{x}_2$	1.9231	4410.4545	.	18.4545	4.9167
Y_4	t_{hesap}	-1.6	1.5	2.8	.	-1.0
	sd	6.9	5.6	14.8	.	5.6
	p	0.160	0.190	0.012	.	0.373
	n_1	30	27	19	30	25
	n_2	7	5	5	0	6
	$\bar{x}_1$	5.1333	3764.2593	16229.4737	14.5000	9.1200
	$\bar{x}_2$	10.4286	2229.0000	7696.0000	.	15.5000
$Y5$	t_{hesap}	-.5	.9	-.3	-.1	.
	sd	6.3	8.6	5.1	5.1	.
	p	0.622	0.417	0.798	0.953	.
	n_1	31	27	19	25	31
	n_2	6	5	5	5	0
	$\bar{x}_1$	5.8710	3625.1852	14118.9474	14.4400	10.3548
	$\bar{x}_2$	7.5000	2980.0000	15716.0000	14.8000	.

Çizelge 3.7'ye göre Y_2 değişkenindeki eksik gözlemlere Y_4 değişkeninin bir etkisinin olup olmadığının belirlenmesi için gerçekleştirilen bağımsız iki örneklem t testi sonucuna göre Y_2 değişkenindeki eksik gözlemlere Y_4 değişkeninin bir etkisinin olmadığına karar verilir ($p = 0.981 > \alpha = 0.05$).

Çizelge 3.7'ye göre Y_3 değişkenindeki eksik gözlemlere Y_1 değişkeninin bir etkisinin olup olmadığının belirlenmesi için gerçekleştirilen bağımsız iki örneklem t testi sonucuna göre Y_3 değişkenindeki eksik gözlemlere, Y_1 değişkeninin istatistiksel olarak anlamlı bir etkisinin olduğuna karar verilir ($p = 0.00 < \alpha = 0.05$).

3.3.2. Korelasyon testi

Rassallığın araştırılmasında kullanılan ikinci yöntem değişkenlerdeki eksik değerleri 0 tam değerleri 1 olarak kodlayarak değişkenler arasındaki korelasyonların hesaplanmasıdır. Eğer hesaplanan korelasyon matrisinde yüksek ilişki veya ilişkiler söz konusu ise eksik veri mekanizması rassal değildir denir. Korelasyon matrisi içerisinde yüksek ilişkiler yoksa bu durum veri yapısının rassallığını göstermektedir.

X ve Y gibi iki rassal değişken arasındaki örneklem Pearson korelasyon katsayısı r_{xy} ile gösterilir ve,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{3.10}$$

eşitliği ile hesaplanır. Eşitlikte yer alan $\overline{x}$, $\overline{y}$ simgeleri sırasıyla X ve Y değişkenlerinin örneklem ortalamalarını göstermektedir.

Çizelge 3.1'de verilen örnek veri seti için eksik gözlemleri 0, tam gözlemleri 1 ile kodlayarak oluşturulan eksik gözlem gösterge değişkenleri arasındaki korelasyon katsayıları hesaplanmış ve sonuçlar Çizelge 3.8'de verilmiştir.

Çizelge 3.8. Eksik gözlem gösterge değişkenleri için Pearson korelasyon katsayıları

Değişkenler		M_2	M_3	M_4	M_5
M_2	r	1	0.040	0.213	0.041
	p		0.813	0.206	0.812
	n	37	37	37	37
M_3	r	0.040	1	-0.066	-0.170
	p	0.813		0.696	0.314
	n	37	37	37	37
M_4	r	0.213	-0.066	1	-0.025
	p	0.206	0.696		0.882
	n	37	37	37	37
M_5	r	0.041	-0.170	-0.025	1
	p	0.812	0.314	0.882	
	n	37	37	37	37

Örnek veri setimizde Y_1 değişkeni tüm gözlem birimleri için gözlemlendiğinden tam veri durumundadır. Dolayısıyla M_1 gösterge değişkeni tamamen 1 değerinden oluşmaktadır. Pearson korelasyon katsayısının hesaplanabilmesi için değişkende en az bir farklı değerin gözlemlenmiş olması gerekir. Bu sebepten dolayı M_1 gösterge değişkeni ile diğer gösterge değişkenleri arasındaki korelasyon hesaplanamamıştır. Hesaplanan Pearson korelasyon matrisi incelendiğinde gösterge değişkenleri arasında istatistiksel olarak anlamlı bir ilişkinin olmadığı görülmektedir ($p > \alpha = 0.05$). Bu yapı eksik gözlem yapısının rassallığını göstermekle birlikte çalışmada yer alan Y_1 değişkeni bu hesaplamalarda dikkate alınamadığından bu veri seti için korelasyon matrisi ile rassallığın incelenmesi çok doğru olmayacaktır.

3.3.3. Little'nin MCAR testi

Little (1988) rassallığın incelenmesi için, veri seti içersindeki tüm değişkenler arasındaki ortalama farkları aynı anda değerlendirecek şekilde t testi yaklaşımının çok değişkenli genelleştirilmiş halini önerdi. Little'nin önerdiği yaklaşım veri seti içersindeki tüm değişkenleri aynı anda ele aldığından tümel bir testtir ve bu özelliği ile tek değişkenli t testinden ayrılmaktadır. Önerilen bu tümel test rassallığın tüm değişkenler için aynı anda incelenmesinde korelasyon yaklaşımından daha yararlıdır. Çünkü örnek veri setimiz için rassallığın incelenmesinde de görüldüğü gibi korelasyon yaklaşımı her zaman uygulanabilir değildir.

Little'nin MCAR testi, t testindeki benzer şekilde ortalama farkların incelenmesinde eksik gözlem içeren veri setini alt gruplara ayırır. Test istatistiği alt grup ortalamaları ve genel ortalamalar arasındaki standartlaştırılmış farkların ağırlıklı toplamıdır ve,

$$\chi^2_{MCAR} = \sum_j n_j (\hat{\mu}_j - \hat{\mu}_j^{(ML)})^T \hat{\Sigma}_j^{-1} (\hat{\mu}_j - \hat{\mu}_j^{(ML)}) \tag{3.11}$$

eşitliği ile hesaplanır. Eşitlikte yer alan n_j, veri seti içersindeki *j.* alt grubun gözlem sayısı, $\hat{\mu}_j$, *j.* alt gruba ait ortalama vektörü, $\hat{\mu}_j^{(ML)}$ simgesi de *j.* alt gruptaki ilgili değişkenler için ortalama vektörün EM algoritmasından elde edilen en çok olabilirlik

tahminini göstermektedir. Eşitlikte yer alan $\hat{\Sigma}_j$ ise j. alt gruptaki ilgili değişkenler için kovaryans matrisinin EM algoritmasından elde edilen en çok olabilirlik tahmini göstermektedir.

Little'ın MCAR testinde tüm değişkenlerdeki eksik veri yapısı analiz edilir ve rassal eksik veri süreci için beklenen yapı ile karşılaştırılır (Craig, 2010). Eğer anlamlı farklılıklar yoksa eksik veri mekanizması TROK olarak adlandırılır, şayet anlamlı farklılıklar bulunursa rassal olmayan eksik veri sürecinin hangi değişkenler nedeniyle ortaya çıktığı, ilk iki yöntem kullanılarak belirlenebilir.

Little'nin MCAR testinin uygulamasını eksik veri mekanizmalarının açıklanmasında kullanılan ve Çizelge 3.6'da verilen veri setinde iş performansı değişkeninin ROK yapısı gösterdiği durum için gerçekleştirelim.

Çizelge 3.9. IQ ve iş performansı değişkenlerinden oluşan veri seti

IQ	İş Performansı
78	-
84	-
84	-
85	-
87	-
91	7
92	9
94	9
94	11
96	7
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

Veri setindeki eksik gözlemlerin rassallığı incelenirken veri seti öncelikle alt gruplara bölünür. İlgili veri setinde iş performansının eksik gözlemlerine karşılık gelen IQ değerleri (78, 84, 84, 85, 87) bir veri yapısı, IQ ve iş performansı değişkeninin birlikte gözlemlendiği veri ikinci veri yapımızı oluşturmaktadır. iş performansının eksik gözlemlerine karşılık gelen IQ değerlerine karşılık gelen alt grup ortalaması $\hat{\mu}_1 = 83.6$

ve diğer alt gruba ait ortalama vektörü ise $\hat{\mu}_2 = [105.467 \quad 10.667]$ olarak hesaplanır. Eşitlik 3.11'de verilen formülün uygulanabilmesi için eksik verilerin EM algoritması ile tahmininden elde edilen en çok olabilirlik tahminleri

$$\hat{\mu}^{(ML)} = [100 \quad 10.667] \qquad \hat{\Sigma} = \begin{bmatrix} 199.5789 & 29.85577 \\ 29.85577 & 9.151386 \end{bmatrix}$$

elde edilir. Test istatistiği

$$\chi^2_{MCAR} = 5 \times (85.6 - 100) \times 199.5789^{-1} \times (85.6 - 100)$$
$$+ 15 \times \begin{bmatrix} 105.467 - 100 & 10.667 - 10.667 \end{bmatrix} \times \begin{bmatrix} 199.5789 & 29.85577 \\ 29.85577 & 9.151386 \end{bmatrix}^{-1} \times \begin{bmatrix} 105.467 - 100 \\ 10.667 - 10.667 \end{bmatrix}$$

$$\chi^2_{MCAR} = 8.984248$$

şeklinde hesaplanır. Hesaplanan test istatistiği alt grup sayısından 1 çıkartılarak elde edilen serbestlik derecesi ve α anlam düzeyinde χ^2_{Tablo} değeri ile karşılaştırılır. Örneğimizde sd=1 ve $\alpha = 0.05$ için belirlenen tablo değeri 3.841 olarak belirlenir. Test istatistiğinin değeri tablo değerinden büyük olduğundan H_0 hipotezi red edilir ve sürecin TROK olmadığına karar verilir.

3.4. Eksik Veri Sorununun Giderilmesine Yönelik Yöntemler Ve Atama Yöntemleri

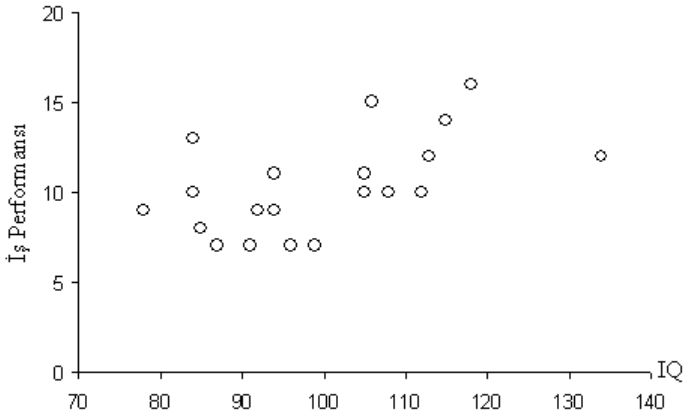
Eksik verinin rassal olup olmadığını belirledikten sonra eksik veri probleminin giderilmesine yönelik yaklaşımlarda bulunulur. Eksik veri probleminin giderilmesine yönelik çözümlemeler eksik veri sürecinin rassallığı ve eksik veri tahmininde kullanılan yöntemlere en genel anlamda silme ve atama yöntemleri olarak iki alt başlıkta incelenir.

3.4.1. Liste bazında silme yöntemi(Listwise Method)

Liste bazında silme yöntemi aynı zamanda tam gözlemlerin kullanılması yöntemi olarak da bilinir. Eksik veri durumunda akla ilk gelen bu yöntem kolaylığı bakımından oldukça yaygın bir kullanıma sahiptir. Bu yöntemde eksik gözlem içeren birimler ya da değişkenler veriden silinerek veri tam veri durumuna dönüştürülür. Bu yöntemin olumsuz tarafı eksik gözlem içeren birimler ya da değişkenlerin çıkartılmasının neden olduğu bilgi kaybıdır. Eksik gözlemlerin bir değişken ya da değişken grubunda yoğunlaşmadığı ve birim bazında silme işleminin gerçekleştiği uygulamalarda örneklem büyüklüğü küçülecek ve başlangıçta örneklem büyüklüğü yeterli iken silme sonucunda yetersiz duruma gelebilecektir.

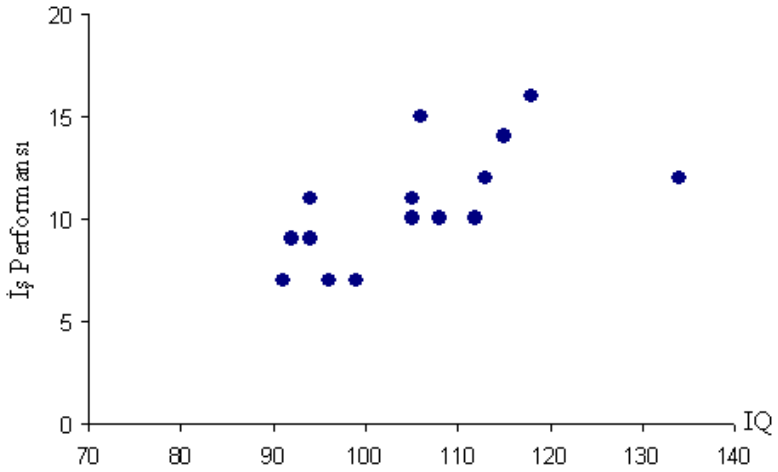
Liste bazında silme yönteminin en büyük avantajı ise uygulama kolaylığıdır. İstatistiksel yazılımların çoğunda yer alan ve eksik veri sorununun çözümünde en basit ve doğrudan bir yaklaşım olan bu yöntem araştırmalarda olur olmaz kullanılmaktadır. Fakat bu doğru bir yaklaşım değildir. Çünkü liste bazında silme yönteminin kullanılabilmesi için eksik veri sürecinin TROK olması gerekir. Ancak TROK varsayımı altında tam gözlemlerin orijinal gözlemlerin rasgele bir örnekleme olduğu ve dikkate alınmayan eksik verili gözlemlerin tahminleri yanlılaştıramayacağı sadece değişkenlere ilişkin güven aralıklarında artışların görülebileceği belirtilmektedir. Eksik veri süreci TROK değilse tam gözlemlerin kullanılması durumunda çözümlemelerde yanlı sonuçlara neden olacak ve elde edilen sonuçlar ana kitleye genellenemeyecektir.

Eksik veri sürecinin TROK olmadığı durumlarda liste bazında silme yönteminin sakıncasını IQ ve iş performansı değişkenlerinden oluşan veri setinde inceleyelim. Çizelge 3.6'da verilen IQ ve iş performansı değişkenlerinin tam veri durumundaki saçılım grafiği Şekil 3.3'de verilmiştir.



Şekil 3.3. Tam veri durumunda IQ ve iş performansı değişkenlerinin saçılım grafiği

Tam veri durumunda IQ ve iş performansı değişkenlerinin saçılım grafiği incelendiğinde değişkenler arasında yüksek bir korelasyon bulunmadığı görülmektedir. İlgili veri setinde eksik veri süreci ROK durumunu göstermek için ilk beş gözlemin eksik olduğu varsayılarak oluşturulan ve Çizelge 3.9'da verilen veri setinin saçılım grafiği Şekil 3.4'de verilmiştir.



Şekil 3.4. Eksik veri sürecinin ROK olduğu ve liste bazında silme sonrası IQ ve iş performansı değişkenlerinin saçılım grafiği

Eksik veri sürecinin ROK olduğu durumda liste bazında silme yönteminin uygulanması sonrası elde edilen saçılım grafiği incelendiğinde IQ ve iş performansı değişkenleri arasında gerçekte olmadığı halde pozitif yönlü önemli bir korelasyon olduğu görülmektedir. Açıkça görüldüğü gibi eksik gözlem sürecinin TROK olmadığı durumlarda liste bazında silme yöntemi uygulanması sonrasında elde edilen sonuçlar yanlı olmakta ve sonuçların genellenebilirlik özelliği ortadan kalmaktadır.

Liste bazında silme yönteminde veriden çıkartılacak uygun gözlem birimi ya da değişken miktarının ne olacağı konusunda herkesçe kabul görmüş, kesin bir yaklaşım ne yazık ki bulunmamaktadır. Araştırmacıların bazıları bir değişkendeki eksik gözlemlerin oranının %15'in üzerinde olması durumunda bu değişkenin araştırmadan çıkartılmasına uygun olacağını ifade etmişlerdir. Araştırmacılar eksik verilerin oldukça fazla olduğu bağımsız değişken veya değişkenlerin diğer değişkenler ile yüksek derecede ilişkili olması durumunda bu tür değişkenlerin veriden çıkartılmasında bilgi kaybı minimum olacağını belirtmektedirler. Çünkü ilgili değişkenin açıklayıcılığının büyük bir bölümü yüksek derecede ilişkili bulunduğu diğer değişkenler tarafından da açıklanabilmektedir.

Eksik verilerin bulunduğu değişken, analizlerde bağımlı değişken olarak seçilirse (örneğin: regresyon analizinde), eksik gözlemlerin bulunduğu birimlerin araştırmadan çıkartılması önerilir. Böylece bu tür çalışmalarda eksik gözlemlerin atama yöntemleri ile tahmin edilmesi durumunda karşılaşılan yapay açıklayıcılık artışının önüne geçilmiş olur. Eksik verinin önemli bir bölümünü az sayıda gözlem birimi üzerine yoğunlaştığı ya da eksik gözlem sayısının az ve veriden çıkartılması durumunda örneklem büyüklüğü etkilemeyecek düzeyde ise liste bazında silme yöntemi birim bazında uygulanarak gerçekleştirilir.

Liste bazında silme yöntemini uygulamadan önce, özellikle eksik veri sürecinin TROK olması varsayımının sağlanıp sağlanmadığı kontrol edilmeli ve silme işlemi için örneklem büyüklüğünün yeterli olup olmadığına dikkat edilmelidir. Araştırmacılar liste bazında silme yönteminin çalışmalarına sağlayacağı katkı ile silme durumunda karşılaşacakları kayıpların çalışmalarına olan etkilerini kıyaslayarak bu yöntemi kullanıp kullanmayacaklarına karar vermelidirler.

Çizelge 3.1'de verilen örnek veri setindeki eksik gözlem probleminde çözüm yöntemi olarak liste bazında silme yöntemi uygulanarak elde edilen veri seti, tanımlayıcı istatistikler, korelasyon ve kovaryans matrisleri sırasıyla Çizelge 3.10, 11, 12 ve 13'de verilmiştir.

Çizelge 3.10. Eksik gözlem içeren birimlerin silinmesi (liste bazında silinme) ile elde edilen tam veri seti

Gözlem No	Y_1	Y_2	Y_3	Y_4	Y_5
8	3	4418	19380	13	4
10	5	3744	15730	27	7
14	1	4187	31680	17	5
16	4	2757	16060	26	8
17	6	12179	33660	27	3
19	2	3169	28030	8	4
20	4	1946	12880	5	10
21	5	2543	13860	6	9
22	13	1347	7510	1	16
23	20	944	2530	4	20
26	5	2429	13480	4	7
28	10	1775	9280	3	17
30	4	3503	15510	7	8

Çizelge 3.11. Eksik gözlem içeren birimlerin silinmesi (liste bazında silinme) ile elde edilen tam veri seti

	Y_1	Y_2	Y_3	Y_4	Y_5
Ortalama	6.307692	3457	16891.54	11.38462	9.076923
Standart Sapma	5.202317	2827.821	9234.044	9.674179	5.376826

Çizelge 3.12. Liste bazında silinme yönteminde elde edilen kovaryans matrisi

	Y_1	Y_2	Y_3	Y_4	Y_5
Y_1	27.0641				
Y_2	-4673	7996569			
Y_3	-33590.5	19787193	85267564		
Y_4	-19.9615	17800.42	51783.53	93.58974	
Y_5	24.64103	-9488.42	-42091	-30.1154	28.91026

Çizelge 3.13. Liste bazında silinme yönteminde elde edilen korelasyon matrisi

	Y_1	Y_2	Y_3	Y_4	Y_5
Y_1	1				
Y_2	-0.31765	1			
Y_3	-0.69924	0.757775	1		
Y_4	-0.39663	0.650675	0.579676	1	
Y_5	0.880919	-0.62405	-0.84776	-0.57896	1

Çizelge 3.10'da verilen liste bazında silme sonrası elde edilen veri seti 13 gözlem birimi içermektedir. 37 gözlem biriminden eksik gözlemler nedeni ile 24 birim veriden silinmiştir. Veriden silinen birimlerin oranı yaklaşık %65'dir. Bu oran oldukça yüksektir ve analizlerde yetersiz örneklem hacmi problemi ile karşılaşılaşılabilir.

3.4.2. Çiftler bazında silme yöntemi (Pairwise Method)

Çiftler bazında silme yöntemi aynı zamanda eldeki mümkün tüm verinin kullanılması yöntemi olarak da bilinir. Bu yöntemde ilgili hesaplamalar veri seti içersindeki eksik gözlemleri dikkate almayarak verinin mümkün olanı ile gerçekleştirilir. Çiftler bazında silme yönteminde aslında veri setinden gözlem biriminin veya değişkenlerin silinmesi söz konusu değildir. Yöntemde sadece hesaplamalarda eksik veriler göz önünde bulundurulmamaktadır.

Liste bazında silme yönteminde bir birimin gözlemlenen değerleri bir değişkendeki gözlemlenemeyen değerinden dolayı çıkartıldığından araştırmacılar eldeki mümkün tüm verinin kullanılması yöntemine sıklıkla başvurmaktadır. Bu yöntemde, eksik gözlemler atanmış değerlerle doldurulmamakta, eldeki mümkün tüm değerler yardımıyla dağılımın ortalama, standart sapma gibi tanımlayıcı istatistikleri, korelasyon ve kovaryans gibi ilişki ölçüleri hesaplanmaktadır. Bu yöntemde aşama aşama her bir değişkenin ortalama ve varyansları ayrı ayrı hesaplanır, ayrıca çift olarak ele alındıklarında mümkün olan tüm çiftler için kovaryans ve korelasyonlar hesaplanır.

X ve Y çiftler bazında kovaryans ve korelasyon değerini hesaplayacağımız iki değişken olsun. Eksik gözlem içeren veri setlerinde çiftler bazında silme yöntemi ile X ve Y değişkenleri arasındaki kovaryans,

$$\hat{\sigma}_{XY} = \frac{\sum (x_i - \hat{\mu}_x^{(P)})(y_i - \hat{\mu}_y^{(P)})}{N_c - 1} \tag{3.12}$$

eşitliği ile hesaplanır. Eşitlikte yer alan $\hat{\mu}_x^{(P)}$, X değişkeninin gözlemlenen değerlerinin ortalamasını gösterirken $\hat{\mu}_y^{(P)}$ gösterimi de benzer şekilde Y değişkeninin gözlemlenen değerlerinin ortalamasıdır. Eşitlikte N_c gösterimi X ve Y değişkenlerinin birlikte gözlemlendiği durum sayısını ifade etmektedir. İki değişken arasındaki korelasyon katsayısı çiftler bazında silme yönteminden elde edilen $\hat{\sigma}_{XY}$, $\hat{\sigma}_X$ ve $\hat{\sigma}_Y$ tahminleri kullanılarak,

$$r = \frac{\hat{\sigma}_{XY}}{\sqrt{\hat{\sigma}_X^2 \hat{\sigma}_Y^2}} \tag{3.13}$$

eşitliği ile hesaplanır.

Çizelge 3.1'de verilen örnek veri setindeki eksik gözlem probleminde çözüm yöntemi olarak yöntemi uygulandığında oluşan frekans dağılımı Çizelge 3.14'de verilmiştir.

Çizelge 3.14. Çiftler bazında silme yönteminde oluşan frekans dağılımı

	Y_1	Y_2	Y_3	Y_4	Y_5
Y_1	**37**				
Y_2	32	32			
Y_3	24	**21**	24		
Y_4	30	27	19	30	
Y_5	31	27	19	25	31

Frekans çizelgesinde esas köşegen üzerindeki frekanslar değişkenlerin tek tek ele alındığı durumdaki gözlem sayılarını, çizelgedeki diğer değerler ise değişken çiftlerinin birlikte gözlemlendiği birim sayılarını vermektedir. Buna göre Y_1

değişkeninin gözlemlendiği 37 birim bulunurken, Y_2 değişkeni ile Y_3 değişkeninin aynı anda birlikte gözlemlenebildiği 21 birim bulunmaktadır.

Çizelge 3.1'de verilen örnek veri setindeki eksik gözlem probleminde çözüm yöntemi olarak çiftler bazında silme yöntemi uygulandığında elde edilen tanımlayıcı istatistikler Çizelge 3.15'de verilmiştir.

Çizelge 3.15. Çiftler bazında silme yönteminde elde edilen ortalama ve standart sapmalar

Çiftler Bazında Ortalamalar						**Çiftler Bazında Standart Sapmalar**					
	Y_1	Y_2	Y_3	Y_4	Y_5		Y_1	Y_2	Y_3	Y_4	Y_5
Y_1	**6.14**	3524.38	14451.67	14.50	**10.35**	Y_1	5.94	2147.21	9477.21	10.16	9.79
Y_2	5.81	3524.38	15389.52	14.48	8.96	Y_2	5.92	2147.21	9717.40	10.23	8.71
Y_3	8.42	3060.24	14451.67	12.21	13.79	Y_3	6.28	2388.21	9477.21	10.57	11.29
Y_4	5.13	3764.26	16229.47	14.50	9.12	Y_4	4.80	2104.15	9708.30	10.16	7.77
Y_5	**5.87**	3625.19	14118.95	14.44	10.35	Y_5	5.76	2265.92	8927.65	9.99	9.79

Çizelge 3.15'de Y_1 değişkeni tek başına değerlendirildiğinde gözlemlenebildiği 37 birimin almış olduğu değerlerin ortalaması 6.14'dür. Y_1 ve Y_5 değişkeninin birlikte gözlemlendiği 31 birimin Y_5 değişkeni için almış olduğu değerlerin ortalaması 10.35, Y_1 değişkeni için almış olduğu değerlerin ortalaması 5.87'dir.

Çizelge 3.1'de verilen örnek veri setindeki eksik gözlem probleminde çözüm yöntemi olarak çiftler bazında silme yöntemi uygulandığında elde edilen kovaryans ve korelasyon matrisi Çizelge 3.16'de verilmiştir.

Çizelge 3.16. Çiftler bazında silme yönteminde elde edilen kovaryans ve korelasyon matrisleri

Kovaryans Matrisi						**Korelasyon Matrisi**					
	Y_1	Y_2	Y_3	Y_4	Y_5		Y_1	Y_2	Y_3	Y_4	Y_5
Y_1	35.34					Y_1	1.00				
Y_2	-6315.83	4610525.92				Y_2	-0.50	1.00			
Y_3	-43918.1	15492326.1	89817440.5			Y_3	-0.74	0.67	1.00		
Y_4	-18.28	11628.45	54014.01	103.29		Y_4	-0.37	0.54	0.53	1.00	
Y_5	38.21	-9407.11	-70605.23	-21.35	95.90	Y_5	0.68	-0.48	-0.70	-0.27	1.00

Çiftler bazında silme yöntemi, bir değişkendeki bir eksik nedeniyle o gözlemin tamamen veriden çıkarılmasını sorununu ortadan kaldırması ve kullanılan veriyi en fazlaya çıkarmasına rağmen bazı sorunları da beraberinde getirebilmektedir. Bu sorunların en belirgini, korelasyonların beklenen sınırlarının dışında elde edilmesi ve

hesaplanan korelasyonların, korelasyon matrisindeki diğer korelasyonlarla tutarsız olmasıdır. X ve Y değişkenleri arasındaki korelasyon, bu değişkenlerin Z gibi bir değişken ile yaptığı korelasyonlar ile sınırlı olup, bu sınırlar

$$r_{xy} \text{ 'nin sınırları} = r_{XZ}r_{YZ} \pm \sqrt{(1-r_{xz}^2)(1-r_{yz}^2)} \tag{3.14}$$

eşitliği ile tanımlanır. Matristeki korelasyonlara ilişkin sınırları tanımlayan bu formüle göre X ve Y değişkenlerinin diğer değişkenlerle olan korelasyonu artıkça X ve Y değişkenleri arasındaki korelasyon değerinin olası sınırları azalır. Bu durum, bir gözlem kümesinden elde edilen korelasyonlarla tutarsız olması olasılığını arttırır. Örneğin X ve Y değişkenlerinin Z değişkeni ile olan korelasyonu sırasıyla 0.90 ve 0.80 olsun. Buna göre X ve Y değişkenleri arasındaki korelasyonun olası sınırları,

$$r_{xy} \text{ 'nin sınırları} = 0.90 \times 0.80 \pm \sqrt{(1-0.90^2)(1-0.80^2)} = [0.4585; 0.9185]$$

şeklinde oluşur. Dolayısıyla bu sınırlar dışında bir korelasyon katsayısının elde edilmesi, sonuçların matematiksel olarak tutarsız olmasına neden olur.

Korelasyon katsayılarındaki bu soruna bağlı olarak ortaya çıkan br diğer sorun ise, korelasyon matrisinin özdeğerlerinin negatif olabilmesidir. Bu durumda, korelasyon matrisinin varyans özellikleri de değişir.

3.4.3. Atama yöntemleri (Imputation Methods)

Eksik verinin giderilmesine yönelik olarak başvurulacak bir yöntem örneklemdeki tam veri içeren değişken ya da gözlem değerlerini kullanarak eksik verilerin tahmin edilmesidir. Eksik değerlerin tahmini için birçok yöntem bulunmaktadır. Bu yöntemler atama yöntemleri olarak isimlendirilir. Bu yöntemler ilk bakışta çekici yöntemler olmakla beraber analiz sonuçlarına etkileri nedeni ile tehlikeli olabilecek yöntemlerdir. Atama yöntemleri örneklemin tümünü genelleştirmek amacıyla gözlemlerin bir alt kümesinden elde edilen bilgi yardımıyla eksik verilerin tahmini şeklinde tanımlanabilir.

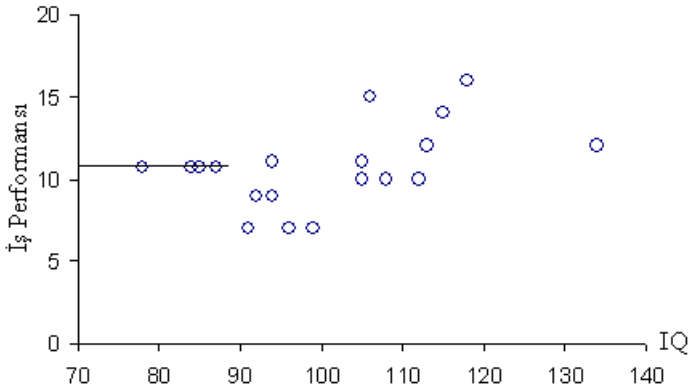
3.4.3.1. Ortalama değerin atanması (Mean Imputation)

Eksik gözlemler yerine ortalama değerin atanması yöntemi Wilks (1932) tarafından önerilen en eksi eksik gözlem çözüm yöntemlerinden biridir ve işlem kolaylığı nedeni ile birçok araştırmacı tarafından yaygın olarak kullanılmaktadır. Bu yöntemin temeli, eksik gözlemler yerine eksik gözlemin veya gözlemlerin bulunduğu değişkenin veya değişkenlerin gözlemlenmiş değerlerinin ortalamasının kullanılmasıdır. Hesaplama kolaylığı bakımından sıklıkla tercih edilen bu yöntemin dezavantajı ise parametre tahminlerindeki bozulmalardır. Bu yöntemde elde edilen ortalama tam gözlemlerin ortalamasına eşit olmasına karşın standart sapma değeri küçüleceğinden değişkenler arasındaki korelasyon katsayıları da daha düşük elde edilecektir. Bilindiği gibi örneklem standart sapması,

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}} \tag{3.15}$$

eşitliği ile hesaplanmaktadır. Eksik gözlemler yerine ortalamanın yani $\overline{x}$ değerinin atanması durumunda, eksik gözlemler yerine konan ortalama değer için $(x_i - \overline{x})^2 = 0$ olacaktır. Eşitlik 3.15'de toplam değeri aynı kalırken ortalama ile tahmin edilen eksik gözlemler örneklem hacmini artıracak ve bölen değeri büyüyecektir. Bu durumda da standart sapma değerimiz gerçekte olduğundan daha küçük çıkacaktır. Standart sapmanın gerçeğinden farklı olarak küçük çıkması, varyansın küçülmesine buda değişkenlerin birlikte değişimini gösteren kovaryans ve korelasyon değerlerinin olduğundan küçük çıkmasına neden olur. Ayrıca varyansın olduğundan küçük çıkması nedeniyle değişkenler gerçek dağılımından uzaklaşacaktır.

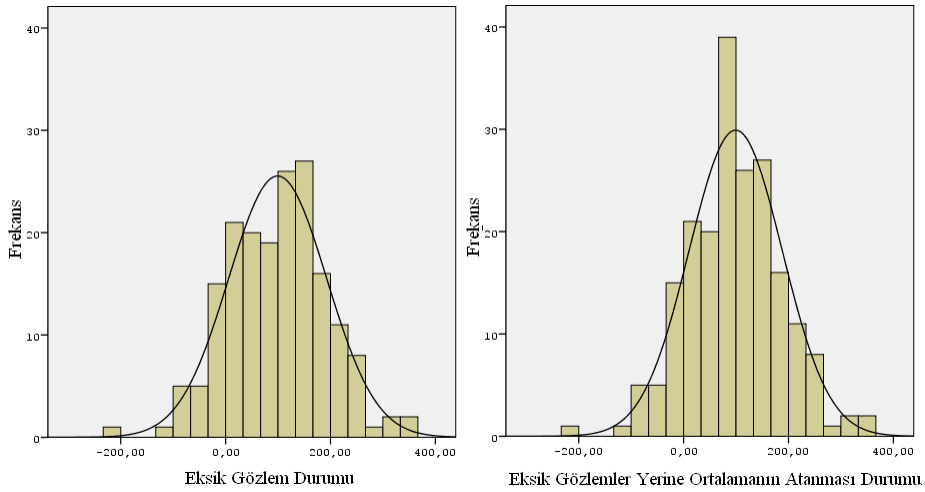
Çizelge 3.9'da verilen IQ ve iş performansı değişkenlerinden oluşan veri setinde eksik gözlemlerin yerine ortalama değerin atanması durumunda saçılım grafiğindeki değişim Şekil 3.5'de verilmiştir.



Şekil 3.5. Ortalama atama sonrasında IQ ve iş performansı değişkenlerinin saçılım grafiği

Şekil 3.5'de iş performansının gözlemlenen değerlerinin ortalaması 10.667 değeri düz çizgi ile gösterilmiştir. Veri setinde iş performansının gözlemlenemeyen ilk beş değeri 10.667 değeri ile tamamlanmıştır. Eksik gözlemler yerine ortalama değerin kullanılması ile iki değişken arasındaki korelasyonun azaldığı açıkça görülmektedir.

Eksik gözlemler yerine ortalamanın atanması durumunda değişkenin gerçek dağılımından uzaklaşması durumunu grafiksel olarak göstermek için normal dağılımdan üretilen 200 birimlik veriden 20 birimin gözlem değerinin silinmesi ile oluşan veri setinin histogramı ile eksik gözlemler yerine ortalamanın atanması ile oluşan tamamlanmış verinin histogramının karşılaştırılması Şekil 3.6'da verilmiştir.



Şekil 3.6. Eksik gözlemler yerine ortalamanın atanmasının değişkenin dağılımı üzerindeki etkisi

Şekil 3.6 incelendiğinde eksik gözlemler yerine ortalanmanın atanması durumunda değişkenin varyansının küçüldüğü ve dağılım şeklinin değiştiği açıkça görülmektedir.

3.4.3.2. En benzer birime ya da birimlere benzetme (Hot Deck Imputation)

Bu atama yönteminde eksik gözlemin bulunduğu birime en benzer birimler tespit edilerek bu birimin o değişken için almış olduğu değer eksik gözlem tahmini olarak kabul edilir. Dolayısıyla bu yöntemin kullanılabilmesi için eksik gözlem içeren birime benzerliği hesaplanacak birimin eksik gözlemin bulunduğu değişken için değerinin gözlemlenmiş olması gerekir. Eksik gözlem yerine, eksik gözlemin bulunduğu birime en benzer birimin gözlem değerinin kullanılması, aynı verinin birden fazla kez kullanılması anlamına gelecek ve veride ortalama atama yöntemine benzer avantaj ve dezavantajları ortaya çıkaracaktır. Bu yöntemde eksik gözlem yerine en benzer bir birimin değerinin atanması yerine birden fazla benzer birimin o değişken için almış olduğu değerlerin atanması son dönemde daha yaygın olarak kullanılmaktadır. Böylelikle aynı değerin tekrar tekrar kullanılmasının önüne geçilmiştir. Buradaki en önemli problemde uygun benzer birim sayısının ne olacağı problemidir. Burada araştırmacılar benzerlik olarak hesaplanan uzaklık ya da ilişki katsayılarındaki sıçramaları dikkate alırlar. Ancak burada genel kabul görmüş bir yaklaşım bulunmamaktadır.

En benzer birimin ya da birimlerin tespitinde benzerlik hesaplanırken genellikle Öklid uzaklığı kullanılır. Öklid uzaklığı birimlerin aynı değişken için almış olduğu değerler arasındaki farkların kareleri toplamının kareköküdür. Yani *i.* ve *k.* birimler arasındaki Öklid uzaklığı *j* değişkenleri göstermek üzere,

$$d_{ik} = d_{ki} = \sqrt{\sum \left(x_{ij} - x_{kj} \right)^2} \tag{3.16}$$

eşitliği ile hesaplanır. Eşitlikte de gösterildiği üzere, *i.* ve *k.* birimler arasındaki Öklid uzaklığı, *k.* ve *i.* birimler arasındaki Öklid uzaklığına eşittir. Birimler arasındaki benzerliğin uzaklık ölçüleri ile belirlenmesi durumunda en küçük uzaklığa sahip iki birim en benzerdir, denir. Dolayısıyla uzaklık büyüdükçe benzerlik azalmakta, uzaklık değeri küçüldükçe de benzerlik artmaktadır.

En benzer birime benzetme yönteminin uygulaması için Çizelge 3.17'de beş birimden oluşan küçük bir veri seti verilmiştir.

Çizelge 3.17. En benzer birime benzetme yönteminin uygulaması için örnek veri seti

Göz. No	Y_1	Y_2	Y_3	Y_4	Y_5
1	5	47	25	28	5
2	3	39		26	4
3	2	45		20	5
4	7		35	25	
5	1	41	31	17	6

Çizelge 3.17'de verilen veri setinde ikinci ve üçüncü gözlem birimlerinin Y_3 değişkeni için, dördüncü biriminde Y_2 ve Y_5 değişkenleri için gözlem değerleri elde edilememiştir. Eksik gözlemlerin tahmininde, gözlem değerleri tam olan birinci ve beşinci birimler kullanılacaktır. Burada ikinci ve üçüncü birimlerin eksik gözlemlerinin tahmininde dördüncü biriminde kullanılabileceği düşünülebilir. Ancak gözlemleri tam olan birimler varken daha az gözlem değerine sahip birimlerin kullanılması önerilmemektedir. Eksik gözlemlerin tahmininde öncelikle, eksik gözlem barındıran birimlerin gözlem değerleri tam olan birinci ve beşinci birimlere olan Öklid uzaklıklarının belirlenmesi gerekir.

$$d_{12} = \sqrt{(5-3)^2 + (47-39)^2 + (28-26)^2 + (5-4)^2} = 8.544$$

$$d_{52} = \sqrt{(1-3)^2 + (41-39)^2 + (17-26)^2 + (6-4)^2} = 9.644$$

$$d_{13} = \sqrt{(5-2)^2 + (47-45)^2 + (28-20)^2 + (5-5)^2} = 8.775$$

$$d_{53} = \sqrt{(1-2)^2 + (41-45)^2 + (17-20)^2 + (6-5)^2} = 5.196$$

$$d_{14} = \sqrt{(5-7)^2 + (25-35)^2 + (28-25)^2} = 10.630$$

$$d_{54} = \sqrt{(1-7)^2 + (31-35)^2 + (17-25)^2} = 10.770$$

Elde edilen uzaklıklar incelendiğinde eksik gözlem barındıran ikinci birime birinci birimin daha benzer olduğu ($d_{12} < d_{52}$), üçüncü birime beşinci birimin daha benzer olduğu ($d_{13} > d_{53}$) ve dördüncü birime de birinci birimin daha benzer olduğu ($d_{14} < d_{54}$) görülmektedir. Buna göre ikinci birimin gözlemlenemeyen Y_3 değeri birinci birimin Y_3 değişkeninin 25 değeriyle, üçüncü birimin gözlemlenemeyen Y_3 değeri beşinci birimin Y_3 değişkeninin 31 değeriyle doldurulur. İki eksik gözlemi olan dördüncü birimin eksik olan Y_2 ve Y_5 değişken değerleri birinci birimin ilgili değişken

için almış olduğu 47 ve 5 değeri ile doldurulur. En benzer birime benzetme yöntemi ile elde edilen tamamlanmış veri seti Çizelge 3.18'de verilmiştir.

Çizelge 3.18. En benzer birime benzetme yöntemi ile tamamlanmış veri seti

Göz. No	Y_1	Y_2	Y_3	Y_4	Y_5
1	5	47	25	28	5
2	3	39	**25**	26	4
3	2	45	**31**	20	5
4	7	**47**	35	25	**5**
5	1	41	31	17	6

3.4.3.3. Dış kaynaklı atama (Cold Deck Imputation)

Bu yöntemde önceki çalışmalardan aynı çalışmanın bir önceki uygulamasından ya da dış kaynaklardan elde edilen sabit bir değer ilgili değişkendeki tüm eksik değerler yerine atanır. Bu yöntemin ortalama değeri atama yönteminden tek farkı atanacak değerin kaynağıdır. Araştırmacı dış kaynaklardan elde edilen değeri eldeki veriler yardımıyla elde edilecek ortalama gibi bir değerden daha geçerli olabileceğinden emin olmalıdır. Bu yöntemde ortalama değerin atanması yöntemine benzer olumsuzluklara sahiptir.

3.4.3.4. Regresyon atama (Regression Imputation)

Regresyon çözümlenmesinin temel amacı bir ya da daha fazla bağımsız değişken yardımıyla bağımlı değişken değerlerinin tahmin edilmesidir. Regresyon atama yönteminde eksik gözlemlerin bulunduğu değişken bağımlı değişken olarak ele alınırken gözlemleri tam olan ve eksik gözlemlerin tahmininde kullanılacak veri setindeki diğer değişken (veya değişkenler) bağımsız değişken (veya değişkenler) olarak ele alınır. Bağımlı ve bağımsız değişkenlerin birlikte gözlemlenmiş değerleri göz önünde bulundurularak oluşturulacak regresyon modeli ile eksik gözlemler tahmin edilmeye çalışılır. Bu yaklaşımda eksik gözlemlerin bulunduğu değişken ölçüm değişkeni ise basit veya çoklu regresyon yöntemi kullanılırken eksik gözlemlerin bulunduğu değişken kategorik ise lojistik regresyon uygulanır. Bu yöntemin verimli olabilmesi için oluşturulan regresyon modelinin açıklama oranının yüksek ve istatistiksel olarak anlamlı olması gerekir. Böyle bir açıklama oranı söz konusu değilse

bu yöntemin kullanılması avantajlı değildir. Regresyon yöntemi ile atamanın dezavantajları ise özetle;

i. Eksik değerler tahmininde var olan ilişkiden yararlanıldığında değişkenler arasındaki ilişki olduğundan daha da kuvvetlenmiş olur.
ii. Bu yöntemle tahmin edilen eksik gözlem sayısı fazla ise veri kendine özgü bir yapıya ulaşabilir ve böylece sonuçların genellenebilirliği azalır.
iii. Regresyon yöntemi ile elde edilen tahminler veride var olan sınırların dışında bulunabilir. Örneğin eksik gözlemin bulunduğu değişken 0 ila 20 arasında değer alırken oluşturulan regresyon denklemi ile oluşan değer 21 olarak bulunabilir.

şeklinde verilebilir.

Regresyon atama yönteminde eksik gözlemlerin tahmininde uygun regresyon modeli basit doğrusal regresyon olarak belirlendiğinde aşağıdaki gösterim ve formüllerden yararlanılmaktadır.

X ve Y gibi iki değişkenden oluşan veri setinde, X değişkeninin tüm birimler için değerinin gözlemlendiği yani tam gözlemli, Y değişkeninde de eksik gözlemlerin bulunduğunu varsayalım. İlgili veri setinde Y değişkenindeki eksik gözlemler regresyon atama yöntemi ile tahmin edilmek istendiğinde, tam gözleme sahip olan X değişkeni bağımsız değişken, eksik gözlemleri regresyon yöntemi ile tahmin edinilmek istenen Y değişkeni bağımlı değişken olarak ele alınır. İki değişkenin birlikte gözlemlendiği birimler göz önünde bulundurularak oluşturulacak regresyon modeli,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3.17}$$

şeklindedir. Burada $\hat{\beta}_0$ ve $\hat{\beta}_1$ tahminlerini gözlem değerlerinden yararlanarak tahmin edinilmek istenirse, iki değişkenin birlikte gözlemlendiği birimleri ifade etmek için m_{ij}, *i.* birimin $j.$ değişken için eksik gözlem gösterge değeri olmak üzere,

$$\delta_i = \begin{cases} 1, & m_{i1} = 0 \ ve \ m_{i2} = 0 \ ise \\ 0, & diğer \ durumlarda \end{cases} \tag{3.18}$$

şeklinde yeni bir gösterge değişkeni tanımlayalım. Bu tanımlama altında $\hat{\beta}_0$ ve $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\delta_i \sum_{i=1}^{n}\delta_i x_i y_i - \sum_{i=1}^{n}\delta_i x_i \sum_{i=1}^{n}\delta_i y_i}{\sum_{i=1}^{n}\delta_i \sum_{i=1}^{n}\delta_i x_i^2 - \left(\sum_{i=1}^{n}\delta_i x_i\right)^2} \tag{3.19}$$

$$\hat{\beta}_0 = \overline{y}^{(c)} - \hat{\beta}_1 \overline{x}^{(c)} \tag{3.20}$$

eşitlikleri ile elde edilir. Eşitlik 3.20'de yer alan $\overline{y}^{(c)}$ ve $\overline{x}^{(c)}$,

$$\overline{y}^{(c)} = \frac{\sum_{i=1}^{n}\delta_i y_i}{\sum_{i=1}^{n}\delta_i} \tag{3.21}$$

$$\overline{x}^{(c)} = \frac{\sum_{i=1}^{n}\delta_i x_i}{\sum_{i=1}^{n}\delta_i} \tag{3.22}$$

şeklinde tanımlanır. Regresyon modelindeki parametrelerin tahminleri elde edildikten sonra, Y değişkenindeki eksik gözlemlere karşılık gelen X değişkeni değerleri yardımıyla eksik gözlemler regresyon modeline göre tahmin edilir.

IQ ve iş performansı değişkenlerinden oluşan veri setinde eksik gözlemlerin tahminini regresyon atama yöntemi ile gerçekleştirelim. İlgili veri setinde IQ değişkeni tam gözlemli, iş performansı değişkeni eksik gözlemlere sahipti. Dolayısıyla regresyon atama yönteminde, eksik gözlemleri tahmin edilmek istenen iş performansı değişkeni bağımlı değişken, gözlemleri tam olan IQ değişkeni de bağımsız değişken olarak ele alınacaktır. Regresyon modelindeki parametrelerin tahmini için gerçekleştirilen hesaplamalar Çizelge 3.19'da verilmiştir.

Çizelge 3.19. Regresyon modelindeki parametrelerin tahmini için gerçekleştirilen hesaplamalar

IQ (x_i)	**İş Per** (y_i)	δ_i	$\delta_i x_i$	$\delta_i y_i$	$\delta_i x_i^2$	$\delta_i x_i y_i$
78	-	0	0	0	0	0
84	-	0	0	0	0	0
84	-	0	0	0	0	0
85	-	0	0	0	0	0
87	-	0	0	0	0	0
91	7	1	91	7	8281	637
92	9	1	92	9	8464	828
94	9	1	94	9	8836	846
94	11	1	94	11	8836	1034
96	7	1	96	7	9216	672
99	7	1	99	7	9801	693
105	10	1	105	10	11025	1050
105	11	1	105	11	11025	1155
106	15	1	106	15	11236	1590
108	10	1	108	10	11664	1080
112	10	1	112	10	12544	1120
113	12	1	113	12	12769	1356
115	14	1	115	14	13225	1610
118	16	1	118	16	13924	1888
134	12	1	134	12	17956	1608
Toplam		15	1582	160	168802	17167

Çizelge 3.19'daki değerlere göre $\hat{\beta}_1$ ve $\hat{\beta}_0$,

$$\hat{\beta}_1 = \frac{15 \times 17167 - 1582 \times 160}{15 \times 168802 - 1582^2} = 0.149628$$

$$\hat{\beta}_0 = \frac{160}{15} - 0.149628 \times \frac{1582}{15} = -5.11411$$

şeklinde elde edilir. Parametre tahminleri sonrasında iş performansı değişkenindeki eksik değerlerin tahmininde kullanılacak regresyon modeli,

$$(\dot{I}\check{s}Per)_i = -5.11411 + 0.149628 \times (IQ)_i$$

şeklinde oluşturulur. İş performansı değişkendeki eksik gözlem tahminleri,

$$(\dot{I}\check{s}Per)_1 = -5.11411 + 0.149628 \times (78) = 6.56$$

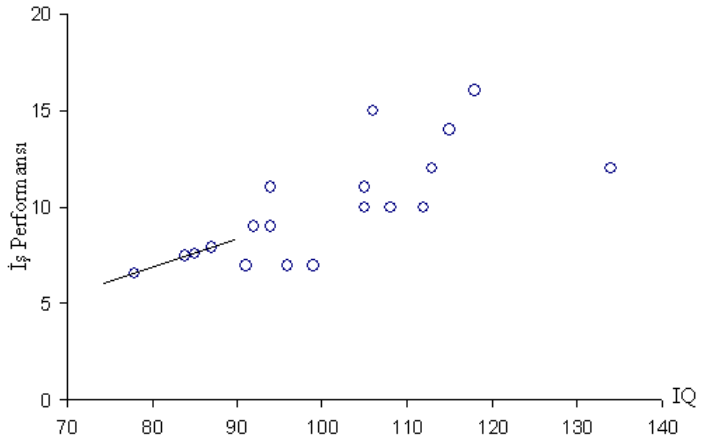
$$(\dot{I}\check{s}Per)_2 = -5.11411 + 0.149628 \times (84) = 7.45$$

$$(\dot{I}\check{s}Per)_3 = -5.11411 + 0.149628 \times (84) = 7.45$$

$$(\dot{I}\check{s}Per)_4 = -5.11411 + 0.149628 \times (85) = 7.60$$

$$(\dot{I}\check{s}Per)_5 = -5.11411 + 0.149628 \times (87) = 7.90$$

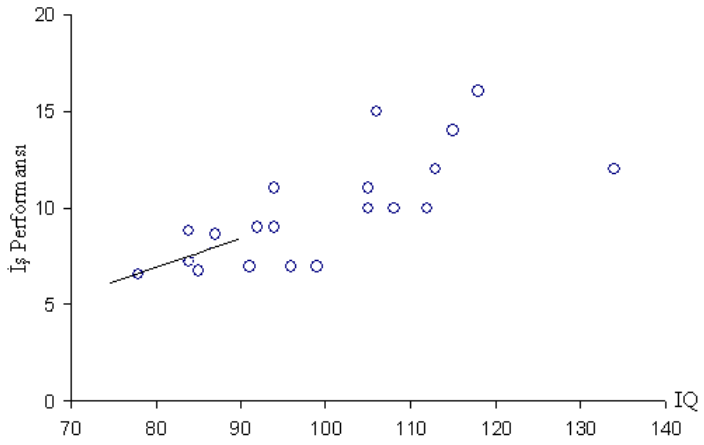
eşitlikleri ile tahmin edilir. Eksik gözlemlerin regresyon ataması sonrası IQ değişkeni ile iş performansı değişkeninin saçılım grafiği Şekil 3.7'de verilmiştir.



Şekil 3.7. Regresyon atama yöntemi sonrasında IQ ve iş performansı değişkenlerinin saçılım grafiği

Şekil 3.7'de iş performansının eksik gözlemlerinin tahmininde kullanılan regresyon doğrusu düz çizgi ile gösterilmiştir. Eksik gözlemlerin regresyon ataması yöntemi ile tahmininden sonra ortalama atama yönteminin aksine ilgili değişkenler arasındaki ilişkinin arttığı açıkça görülmektedir. Araştırmacıların bir kısmı bu dezavantajın etkisini azaltmak için tahmin değerlerine hata terimi eklemeyi önermektedir. Literatürde regresyon tahminlerine hata teriminin eklenmesi yöntemi stokastik regresyon atama yöntemi olarak isimlendirilmektedir. Eklenecek hata terimi için araştırmacıların bir kısmı $N(0,1)$ dağılımından rassal sayı üretmeyi önerirken, bir kısmı da $N(0,\hat{\sigma}_e^2)$ dağılımından rassal sayı üretmeyi önermektedir.İ ş performansının eksik gözlemlerinin tahminine eklenmek üzere $N(0,1)$ dağılımından üretilen beş rassal sayının tahminlere eklenmesi ile eksik gözlem tahminlerimiz $(6.56 - 0.01 = 6.55)$, $(7.45 + 1.36 = 8.81)$, $(7.45 - 0.21 = 7.24)$, $(7.60 - 0.89 = 6.71)$ ve $(7.90 + 0.70 = 8.60)$ şeklinde gerçekleşir.

Eksik gözlemlerin stokastik regresyon atama yöntemi ile tahmininden sonra IQ ile iş performansı değişkenlerine göre elde edilen saçılım grafiği Şekil 3.8'de verilmiştir.



Şekil 3.8. Stokastik regresyon atama yönteminde IQ ve iş performansı değişkenlerinin saçılım grafiği

Regresyon atama yönteminde eksik gözlemlerin tahmini için birden fazla bağımsız değişken olması durumunda çoklu regresyon analizi gerçekleştirilebileceği gibi işlem kolaylığı sağlanması amacı ile bu bağımsız değişkenler arasından bağımlı değişken ile en yüksek korelasyona sahip olan bağımsız değişken tespit edilerek basit doğrusal regresyon analizi gerçekleştirilebilir. Bağımlı değişken ile bağımsız değişken arasındaki ilişkinin doğrusal regresyon modelleri ile açıklanması yeterli olmadığı durumlarda da doğrusal olmayan regresyon modelleri tercih edilebilmektedir.

3.4.3.5. EM (Expactation-Maximization) yöntemi

EM yöntemi eksik veri probleminin çözümünde kullanılan etkin yöntemlerden biridir. Yöntemde eksik gözlemlerin tahmininde kullanılan en çok olabilirlik tahminlerinin elde edildiği EM algoritması kullanılır. EM algoritmasının başlangıcı 1970'lere kadar gider (Orchard ve Woodbury, 1972; Beale ve Little, 1975; Dempster ve ark., 1977). Dempster ve ark. (1977)'nın çalışması EM algoritmasının gelişiminde önemli bir rol oynamıştır. EM algoritmasının ilk uygulaması eksik gözlem durumunda ortalama vektörün ve kovaryans matrisinin en çok olabilirlik tahminlerinin elde edilmesi olsa da sonradan araştırmacılar çok düzeyli modeller, sonlu karma modeller ve

yapısal eşitlik modellemesi gibi tam veri durumundaki zor problemlerin çözümünde de EM algoritmasını kullanmışlardır.

EM algoritması, **E** ve **M** aşamalarından oluşan yinelemeli bir süreçtir. EM algoritması ile tahmin süreci ortalama vektörü ile kovaryans matrisinin başlangıç tahminlerinin elde edilmesi ile başlar. Algoritmanın **E** adımında tam gözlemleri kullanarak eksik gözlemleri tahmin etmek için oluşturulan ortalama vektörü ve kovaryans matrisi ile stokastik regresyon modeli oluşturulur ve model yardımı ile eksik gözlemler tahmin edilir. Algoritmanın **E** adımının sonunda, eksik gözlemlerin tahmini ile veri seti tamamlanmış veri durumuna getirilir.

Algoritmanın **M** adımında, tamamlanmış veri durumundaki prosedürler uygulanarak ortalama vektörü ve kovaryans matrisi tahminleri yenilenir. Bu süreç tahmin değerlerindeki ya da logaritmik olabilirlik değerlerinin toplamları arasındaki değişimin önemsenmeyecek derecede azalmasına kadar devam eder.

Bu işlemler doğrusallık ve normal dağılım koşulu ile uygulanır. EM yaklaşımı da diğer atama yöntemleri gibi geçerli olmayan standart hatalara yol açar. EM yaklaşımı çoklu atama yöntemi için genişletildiğinde standart hatanın geçerli tahminlerini elde etmek mümkün olabilir (Enders ve Bandalos, 2001).

X değişkeni tam gözlemli, Y değişkeni eksik gözlemli olmak üzere iki değişkenli durum için m_{i1}, Y değişkenine ait eksik gözlem göstergesi, m_{i2}, X değişkenine ait eksik gözlem göstergesi, ve δ_i iki değişkenin birlikte gözlemlendiğinde 1, diğer durumda 0 değerini alan gösterge değişkeni olarak tanımlandığında EM algoritmasında başlangıç tahminleri,

$$\hat{\mu}_Y^{(0)} = \frac{\sum_{i=1}^{n} m_{i1} y_i}{\sum_{i=1}^{n} m_{i1}} \tag{3.23}$$

$$\hat{\mu}_X^{(0)} = \frac{\sum_{i=1}^{n} m_{i2} x_i}{\sum_{i=1}^{n} m_{i2}} \tag{3.24}$$

$$\hat{\sigma}_Y^{2(0)} = \frac{1}{\sum_{i=1}^{n} m_{i1}} \left(\frac{\sum_{i=1}^{n} m_{i1} y_i^2}{\sum_{i=1}^{n} m_{i1}} - \frac{\left(\sum_{i=1}^{n} m_{i1} y_i \right)^2}{\sum_{i=1}^{n} m_{i1}} \right) \tag{3.25}$$

$$\hat{\sigma}_X^{2\,(0)} = \frac{1}{\sum_{i=1}^{n} m_{i2}} \left(\frac{\sum_{i=1}^{n} m_{i2} x_i^2}{\sum_{i=1}^{n} m_{i2}} - \frac{\left(\sum_{i=1}^{n} m_{i2} x_i \right)^2}{\sum_{i=1}^{n} m_{i2}} \right) \tag{3.26}$$

$$\hat{\sigma}_{XY}^{(0)} = \frac{1}{\sum_{i=1}^{n} \delta_i} \left(\frac{\sum_{i=1}^{n} \delta_i x_i y_i}{\sum_{i=1}^{n} \delta_i} - \frac{\sum_{i=1}^{n} \delta_i x_i \sum_{i=1}^{n} \delta_i y_i}{\sum_{i=1}^{n} \delta_i} \right) \tag{3.27}$$

eşitlikleri ile elde edilir. Bu tahminler ile eksik gözlemlerin tahmininde kullanılacak regresyon modeli

$$\hat{y}_i^{(0)} = \hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_i \tag{3.28}$$

şeklinde olup, modeldeki parametre tahminleri,

$$\hat{\beta}_1^{(0)} = \frac{\hat{\sigma}_{XY}^{(0)}}{\hat{\sigma}_X^{2\,(0)}} \tag{3.29}$$

$$\hat{\beta}_0^{(0)} = \hat{\mu}_Y^{(0)} - \hat{\beta}_1^{(0)} \hat{\mu}_X^{(0)} \tag{3.30}$$

eşitlikleri ile elde edilir. Bazı araştırmacılar başlangıç tahminlerinde iki değişken arasında korelasyonun sıfır olduğunu varsayarak yani $\hat{\sigma}_{XY}^{2\,(0)} = 0$ alarak eksik gözlemlerin tahmininde kullanılacak regresyon modelini

$$\hat{y}_i^{(0)} = \hat{\mu}_Y^{(0)} \tag{3.31}$$

şeklinde oluşturmaktadır. Eşitlik 3.31'de verilen model incelendiğinde bu modeli tercih eden araştırmacıların aslında başlangıçta regresyon atama yöntemi yerine işlem kolaylığı nedeni ile ortalama atama yöntemini tercih ettiği görülmektedir. Bu tercih EM

algoritması sonucu elde edilecek tahminleri değiştirmeyip sadece iterasyon sayısını etkileyecektir.

Algoritmanın **E** adımının sonunda Eşitlik (3.28) veya Eşitlik (3.31) kullanılarak eksik veriler tahmin edilerek veri tamamlanmış veri durumuna getirilir.

Algoritmanın **M** adımında tamamlanmış veri göz önünde bulundurularak ortalama vektörü ve kovaryans matrisi yinelenir. Tamamlanmış veri durumunda ortalama vektörünün ve kovaryans matrisinin elemanları,

$$\hat{\mu}_Y = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3.32}$$

$$\hat{\mu}_X = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.33}$$

$$\hat{\sigma}_Y^2 = \frac{1}{n}\left(\frac{\sum_{i=1}^{n} (y_i^2 + (1-\delta_i)\hat{\sigma}_{Y|X}^{2\,*})}{n} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \right) \tag{3.34}$$

$$\hat{\sigma}_X^2 = \frac{1}{n}\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \right) \tag{3.35}$$

$$\hat{\sigma}_{XY} = \frac{1}{n}\left(\frac{\sum_{i=1}^{n} x_i y_i}{n} - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} \right) \tag{3.36}$$

eşitlikleri ile yenilenir. Eşitlik (3.34)'de yer alan $\hat{\sigma}_{Y|X}^{2\,*}$ değeri bir önceki adımdaki tahminler kullanılarak

$$\hat{\sigma}_{Y|X}^{2\,*} = \hat{\sigma}_Y^{2*} - \hat{\beta}_1^{2*}\hat{\sigma}_X^{2\,*} \tag{3.37}$$

eşitliği ile elde edilir. Bu tahminler ile eksik gözlemlerin tahmininde kullanılacak regresyon modelindeki parametre tahminleri

$$\hat{\beta}_1 = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} \tag{3.38}$$

$$\hat{\beta}_0 = \hat{\mu}_Y - \hat{\beta}_1 \hat{\mu}_X \tag{3.39}$$

eşitlikleri ile yenilenir. EM algoritması ardışık iki iterasyon tahminleri arasındaki fark veya logaritmik olabilirlik değerlerinin toplamları arasındaki fark önemsenmeyecek kadar küçük oluncaya kadar devam eder.

$\mathbf{Y} = (y_{ij})$, $n \times p$ boyutlu bir veri matrisi olmak üzere *i.* birim için logaritmik olabilirlik değeri,

$$\log L_i = \frac{p}{2}\log(2\pi) - \frac{1}{2}\log\left|\hat{\Sigma}\right| - \frac{1}{2}\left[Y_{i.} - \hat{\mu}\right]^T \hat{\Sigma}^{-1}\left[Y_{i.} - \hat{\mu}\right]$$

eşitliği ile elde edilir.

IQ ve iş performansı değişkenlerinden oluşan ve Çizelge 3.9'da verilen veri setindeki eksik gözlemlerin tahmininde EM atama yönteminde başlangıç tahminleri,

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\mu}_X^{(0)} & \hat{\mu}_Y^{(0)} \end{bmatrix} = \begin{bmatrix} 100 & 10.667 \end{bmatrix} \text{ ve } \hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_X^{2\,(0)} & \hat{\sigma}_{XY}^{(0)} \\ \hat{\sigma}_{XY}^{(0)} & \hat{\sigma}_Y^{2(0)} \end{bmatrix} = \begin{bmatrix} 189.600 & 19.489 \\ 19.489 & 7.810 \end{bmatrix}$$

şeklinde elde edilmiştir. Regresyon modelinin parametre tahminleri,

$$\hat{\beta}_1^{(0)} = \frac{19.489}{189.600} = 0.1028$$

$$\hat{\beta}_0^{(0)} = 10.667 - 0.1028 \times 100 = 0.387$$

şeklinde elde edilir. Böylece eksik gözlemlerin tahmininde kullanılacak regresyon modeli

$$\hat{y}_i^{(0)} = 0.387 + 0.1028 x_i$$

şeklindedir.

Eksik gözlem tahminleri ve parametre tahminlerinin yenilenmesi için gerekli terimlerin hesaplanması ile ilgili işlemler Çizelge 3.20'de verilmiştir.

Çizelge 3.20. EM atama yönteminde parametre tahminlerinin yenilenmesi için gerekli hesaplamalar

Göz. No	**IQ** (x_i)	**İş Per.** (y_i)	δ_i	$\hat{y}_i^2 + (1-\delta_i)\hat{\sigma}_{Y\|X}^{2\,*}$	x_i^2	$x_i y_i$
1	78	8.41	0	70.65+5.8063	6084	655.62
2	84	9.02	0	81.40+5.8063	7056	757.86
3	84	9.02	0	81.40+5.8063	7056	757.86
4	85	9.13	0	83.27+5.8063	7225	775.63
5	87	9.33	0	87.06+5.8063	7569	811.76
6	91	7	1	49	8281	637
7	92	9	1	81	8464	828
8	94	9	1	81	8836	846
9	94	11	1	121	8836	1034
10	96	7	1	49	9216	672
11	99	7	1	49	9801	693
12	105	10	1	100	11025	1050
13	105	11	1	121	11025	1155
14	106	15	1	225	11236	1590
15	108	10	1	100	11664	1080
16	112	10	1	100	12544	1120
17	113	12	1	144	12769	1356
18	115	14	1	196	13225	1610
19	118	16	1	256	13924	1888
20	134	12	1	144	17956	1608
Toplam	2000	204.91	15	2225.583	203792	20925.74

Çizelge 3.20'de yer alan hesaplamalarda $(y_i^2 + (1-\delta_i)\hat{\sigma}_{Y|X}^{2\,*})$ teriminin hesaplanması için gerekli $\hat{\sigma}_{Y|X}^{2\,*}$,

$$\hat{\sigma}_{Y|X}^{2\,*} = 7.810 - 0.1028^2 \times 189.6 = 5.8063$$

şeklinde hesaplanır.

EM atama yönteminde ortalama vektörü ve kovaryans matrisi,

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\mu}_X & \hat{\mu}_Y \end{bmatrix} = \begin{bmatrix} 100 & 10.245 \end{bmatrix} \text{ ve } \hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_X^2 & \hat{\sigma}_{XY} \\ \hat{\sigma}_{XY} & \hat{\sigma}_Y^2 \end{bmatrix} = \begin{bmatrix} 189.600 & 21.670 \\ 21.670 & 6.314 \end{bmatrix}$$

şeklinde elde edilmiştir.

Regresyon modelinin parametre tahminleri,

$$\hat{\beta}_1 = \frac{21.670}{189.600} = 0.1148$$

$$\hat{\beta}_0 = 10.245 - 0.1148 \times 100 = -1.232$$

şeklinde elde edilir. Eksik gözlemlerin tahminlerinin yenilenmesinde kullanılacak regresyon modeli,

$$\hat{y}_i = -1.232 + 0.1148 x_i$$

şeklinde elde edilir. Model ile eksik gözlemler 7.72, 8.41, 8.41, 8.53 ve 8.76 olarak tahmin edilir.

3.4.3.6. Çoklu atama yöntemi (Multiple Imputation)

Bu yöntemin temeli eksik gözlemlerin iki ya da daha fazla atama yöntemini birlikte kullanarak tahmin edilmeye çalışılmasıdır. Dolayısıyla bu yaklaşım karma bir tahmin değeri elde etmeyi amaçlar. Bu tahmin değeri genellikle iki yada daha fazla yöntemle elde edilmiş tahmin değerlerinin ortalamasıdır. Çoklu atama yöntemi bir çok araştırmacı tarafından tek bir yöntemle elde edilen tahmin yöntemlerine göre daha güvenilir bulunmaktadır.

Çoklu atama yönteminin çok sayıda avantajı bulunmaktadır. En önemlisi anlaşılır bir yöntemdir. Aynı zamanda analiz değeri alan değişkenlerin normalliğinin ihmal edildiği durumlarda da robust sonuçlar vermektedir. Liste bazında veri silme, çiftler bazında veri silme ve yerine ortalamayı koyma yöntemlerinden pek çok durumda üstün olmaktadır. Dezavantajı ise atama işleminin uzun sürmesidir.

4. KARMA DAĞILIM MODELİNE DAYALI AYRIŞTIRMA ANALİZİ

Ayrıştırma analizi, analiz öncesi birimlerin gruplarının bilindiği $n \times p$ boyutlu veri setinde birimleri ait oldukları gruplara en az hata ile atamak üzere ayırma fonksiyonlarının oluşturulduğu ve ayırma fonksiyonları yardımı ile sonradan gözlemlenen birimlerin gruplarının belirlenmeye çalışıldığı çok değişkenli istatistiksel bir analizdir.

Ayrıştırma analizi uygulamalarından biri olan doğrusal ayrıştırma analizinde ayırma fonksiyonları, çok değişkenli normal dağılım varsayımı ve grupların kovaryans matrislerinin eşitliği varsayımı altında, tahmin değişkenlerinin doğrusal bileşenleri şeklinde elde edilir. Doğrusal ayrıştırma analizinin taşıdığı varsayımlar nedeni ile kullanılabilirliği oldukça zordur. Doğrusal ayrıştırma analizinin taşıdığı bu dezavantajlar nedeni ile birçok farklı yaklaşım geliştirilmiştir. Bu yaklaşımlardan biride karma dağılım modellerine dayalı ayrıştırma analizidir.

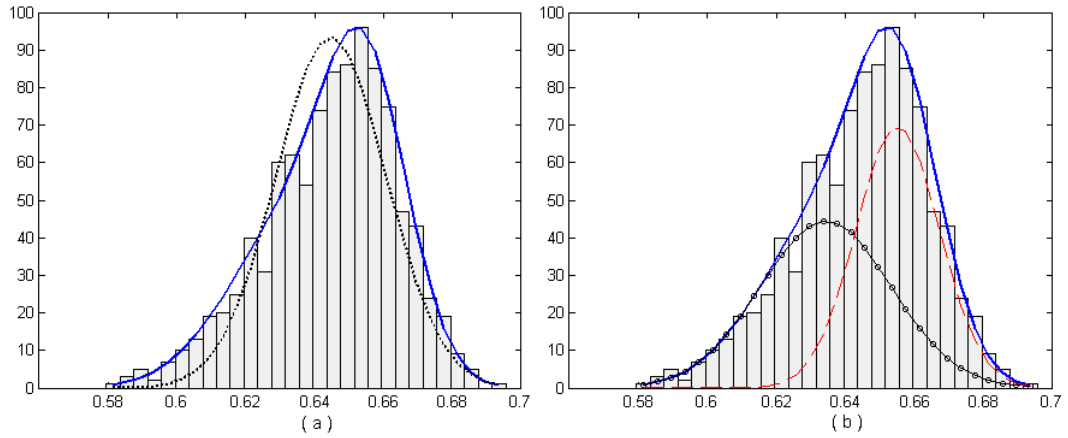
Farklı karakteristik özellikler taşıyan dağılımların bir arada kullanılabildiği karma dağılım modelleri veri modellemede standart olasılık dağılımlarına göre daha esnek ve kullanışlıdır. Dağılım modelleme konusundaki esneklik avantajı nedeni ile doğrusal ayrıştırma analizinin taşıdığı varsayımlar karma dağılım modeline dayalı ayrıştırma analizinde söz konusu değildir. Karma dağılım modeline dayalı ayrıştırma analizinde yaygın olarak çok değişkenli normal dağılımlarının karması kullanılmaktadır. Bu bölümde öncelikle, karma dağılım modelleri tanıtılacak ve sonrasında da çok değişkenli normal dağılımların karmasına dayalı ayrıştırma analizi ele alınacaktır.

4.1. Karma Dağılım Modelleri

İki ya da daha fazla bileşenden oluşan dağılımlar, karma dağılımlar olarak adlandırılır. Karma dağılım modelleri birçok alanda rasgelelik içeren doğal olayların farklı özellikleri hakkında toplanan ölçüm değerlerine istatistiksel olarak model oluşturmada matematiksel bir yaklaşım sağlar. Bu nedenlerden dolayı karma dağılım modellerinin hem teorik hem de uygulamalı birçok çalışma alanında önemi günden güne artmaktadır (Çalış 2005).

Karma dağılım modelleri ile ilgili ilk çalışma yaklaşık 120 yıl kadar önce ünlü biyometrici Karl Pearson tarafından yapılmıştır. Pearson (1894) yengeç verisi için birbirinden farklı iki tek değişkenli normal dağılımın karma modelini incelemiştir. Bu

veriler Weldon (1892, 1893) tarafından ölçülmüştür. Weldon (1893) verilerin histogramı asimetrik olduğundan ya da simetrik olmadığından dolayı, yengeç kitlesinin iki farklı türden oluştuğunu iddia etmiştir. Fakat matematiksel yönü zayıf olduğu için Karl Pearson'dan yardım istemiştir. Bu veri önce histogramda sınıf sayısı 29 alınarak bir tek normal dağılımla modellenmiştir. Daha sonra simetrik olmayan yapı nedeniyle Şekil 4.1'de görüldüğü gibi normal dağılımların karması kullanılarak modellenmiştir.



Şekil 4.1. (a) Yengeç verisi için uydurulan tek bileşenli normal dağılım (nokta nokta) ve iki bileşenli normal dağılım (düz çizgi) eğrileri. (b) Yengeç verisi için uydurulan iki bileşenli normal dağılım (düz çizgi) ve her bir bileşene ait normal dağılım eğrileri

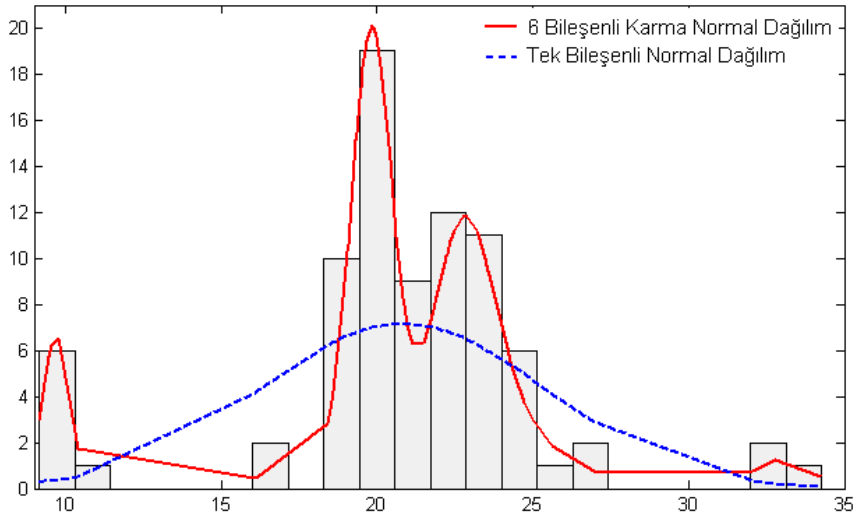
Yengeç verisi iki bileşenli normal dağılımların karması ile modellenirken kitle içerisinde farklı iki alt grup olduğu varsayılmış ve X yengeç vücut genişliği rassal değişkeni olmak üzere karma dağılım modelinin olasılık yoğunluk fonksiyonu

$$f(x) = \pi f_1(x) + (1-\pi) f_2(x) \tag{4.1}$$

şeklinde tanımlanmıştır. Eşitlik (4.1)'de yer alan π, $\pi \in (0,1)$ olmak üzere birinci bileşene ait karma oranıdır. Modelde yer alan $f_1(x)$ ve $f_2(x)$ her bir alt grubun modellenmesinde kullanılan normal dağılımlara ait olasılık yoğunluk fonksiyonlarıdır.

Tahmin edilmesi gereken parametre sayısının fazla olması dezavantajı nedeni ile başlangıçta fazla ilgilenilmeyen karma dağılım modelleri, EM algoritmasının karma dağılım modellerindeki parametrelerinin tahminindeki etkinliği görüldükten sonra çok geniş bir uygulama alanına sahip olmuştur (Erişoğlu, 2011). Çünkü klasik olasılık dağılımları, frekans dağılımının tek modlu olduğu durumlarda kullanışlı iken karma dağılımlar frekans dağılımının iki veya daha fazla moda sahip olan hatta tek modlu

olması durumunda bile verinin modellenmesinde daha kullanışlı ve daha yararlıdır (Everitt ve Hand 1981). Karma dağılım modellerinin, klasik istatistik modellerine göre çok modlu verilerin modellenmesindeki başarısı Roeder (1990) tarafından tanımlanan Galaxy veri setinin modellenmesinde açıkça görülmektedir. Galaxy veri seti için uydurulan 6 bileşenli karma normal dağılım ve tek bileşenli normal dağılımın yoğunluk eğrileri Şekil 4.2'de verilmiştir.



Şekil 4.2. Roeder (1990)'in Galaxy veri seti için uydurulan 6 bileşenli karma normal dağılım (düz çizgi) ve tek bileşenli normal dağılımın (kesikli çizgi) yoğunluk eğrileri

Şekil 4.2 incelendiğinde Galaxy verisinin tek bileşenli normal dağılımla modellenmesi durumunda yapılacak hatanın, 6 bileşenli karma normal dağılımla modellenmesi durumunda yapılacak olan hatadan çok daha fazla olduğu açıkça görülmektedir.

4.1.1. Karma dağılım modelleri için temel gösterimler

$X_1,\ldots,X_n$ ile n birimlik bir rasgele örneklem gösterilsin. Her bir X_i, $\Re^p$ uzayında tanımlı, p-boyutlu bir rasgele vektör olsun. Rasgele vektör X_i'nin olasılık yoğunluk fonksiyonu $f(x_j)$ ile gösterilsin. Bu durumda rasgele vektör X_i, ilgilenilen i. nesne ya da birim üzerinden gözlenen p tane değişkene ait ölçüm değerlerini içerir. Tüm örneklem X ile gösterilirse, $X=\left(X_1^T,\ldots X_n^T\right)^T$ şeklinde yazılır. Gözlenmiş bir rasgele örneklem $x=\left(x_1^T,\ldots x_n^T\right)^T$ ile ifade edilsin. Burada x_i, X_i rasgele vektörünün

gözlenen değeridir. Karma dağılım modelleri içerisinde g (≥ 2) tamsayı olmak üzere, g tane farklı alt grup veya sınıfın olduğu varsayımı yapılır. Sonlu karma dağılım modelinin olasılık yoğunluk fonksiyonu

$$f(x_i) = \pi_1 f_1(x_i) + \pi_2 f_2(x_i) + \cdots + \pi_g f_g(x_i)$$

$$f(x_i) = \sum_{k=1}^{g} \pi_k f_k(x_i) \quad i = 1,2,\ldots,n \tag{4.2}$$

biçiminde verilsin. Burada π_k katsayı değerleri negatif olmayan $(0 \leq \pi_k \leq 1)$ ve toplamları 1 $(\sum_{k=1}^{g} \pi_k = 1)$ olan karma oranları ifade eder. Eşitlik (4.2)'de yer alan $f_k(x_i)$, $k.$ alt grupta x_i rassal vektörün olasılık yoğunluk fonksiyonunu ifade eder.

4.1.2. Karma dağılım modelleri için parametrik gösterimler

Birçok uygulamada karma dağılımın $f_k(x_i)$ bileşen olasılık yoğunluk fonksiyonları belirli bir parametrik olasılık dağılım ailesinden alınır. Bu durumda $f_k(x_i)$ bileşen olasılık yoğunluk fonksiyonu, $f_k(x_i;\boldsymbol{\theta}_k)$ şeklinde parametrik olarak ifade edilir. Burada $\boldsymbol{\theta}_k$ karma dağılım modelinin $k.$ bileşen olasılık yoğunluk fonksiyonun bilinmeyen parametrelerini içeren bir vektördür. Karma olasılık dağılımı $f(x_i)$ parametrik olarak,

$$f(x_i;\boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_k f_k(x_i;\boldsymbol{\theta}_k) \tag{4.3}$$

şeklinde ifade edilir(Servi, 2009). Burada $\boldsymbol{\Psi}$, karma olasılık dağılım modelinde yer alan tüm bilinmeyen parametrelerin bir vektörüdür ve

$$\boldsymbol{\Psi} = \left(\pi_1,\ldots,\pi_{g-1},\boldsymbol{\xi}^T\right)^T \tag{4.4}$$

şeklinde açık olarak yazılır. Burada $\boldsymbol{\xi}$, $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_g$ bileşen dağılımların bilinmeyen parametre vektörlerini içeren bir vektördür.

Karma dağılım modelini oluşturan bileşenlerden her birinin çok değişkenli normal dağılıma sahip olması durumunda olasılık yoğunluk fonksiyonu,

$$f(\mathbf{x}_i;\boldsymbol{\Psi}) = \sum_{k=1}^{g} \pi_k \Phi_k(\mathbf{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \tag{4.5}$$

ile verilir. Burada $k = 1,2,\ldots,g$ ve $i = 1,2,\ldots,n$ olmak üzere $\Phi_k(\mathbf{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ fonksiyonu ortalama vektörü $\boldsymbol{\mu}_i$ kovaryans matrisi $\boldsymbol{\Sigma}_i$ olan,

$$\Phi_k(\mathbf{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = (2\pi)^{-\frac{p}{2}} \left|\boldsymbol{\Sigma}_k\right|^{-\frac{1}{2}} e^{\left\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\right\}} \tag{4.6}$$

şeklinde ifade edilen çok değişkenli normal dağılım fonksiyonudur. Bu durumda karma modelin tüm bilinmeyen parametreler vektörü $\boldsymbol{\Psi} = \left(\pi_1,\ldots,\pi_{g-1},\boldsymbol{\xi}^T\right)^T$ şeklinde yazılır. Burada $\boldsymbol{\xi}$, karma dağılım modelindeki bileşen olasılık yoğunluk fonksiyonlarının parametreleri olan bileşen ortalama vektörleri $\boldsymbol{\mu} = (\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_g)$ ve bileşen kovaryans matrisleri $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\ldots,\boldsymbol{\Sigma}_g)$ 'dan oluşur.

4.1.3. Karma dağılım modelinde parametre tahmini

EM algoritması, Expectation (Beklenen değeri hesaplama adımı) ve Maximization (Maksimum yapma adımı) sözcüklerinin ilk harflerinin kullanılmasıyla ifade edilmiştir. EM algoritması, Dempster ve ark. (1977) tarafından geliştirilmiştir. EM algoritmasıyla ilgili ilk çalışmalar; Baum ve ark. (1970), Orchard ve Woodbury (1972) ve Sundberg (1974) tarafından yapılmıştır. Karma dağılım modellerinde parametre tahmini için en uygun yöntem, en çok olabilirlik yöntemiyle birlikte EM algoritmasının kullanılmasıdır. EM algoritmasında temel amaç, $\mathbf{x}$ tamamlanmamış veri kümesi olmak üzere,

$$L(\boldsymbol{\Psi}) = f(\mathbf{x};\boldsymbol{\Psi}) = \sum_{k=1}^{g} \pi_k f_k(\mathbf{x};\theta_k) \tag{4.7}$$

olabilirlik fonksiyonunu maksimum yapan $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}$ parametre tahmini hesaplamaktır.

EM algoritmasının temelinde tamamlanmış veri yapısı vardır. $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ veri kümesine tamamlanmamış veri kümesi denir. $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ veri kümesine tamamlanmamış veri kümesi denilmesinin nedeni, $\mathbf{x}_i$ gözlem vektörünün karma dağılım modelinde hangi bileşene ait olduğunun bilinmemesidir. $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ veri kümesi için bileşen etiket vektörleri,

$$z_{ik} = \begin{cases} 1 & \mathbf{x}_i \text{ gözlem vektörü k. bileşenine ait ise} \\ 0 & \mathbf{x}_i \text{ gözlem vektörü k. bileşenine ait değilse} \end{cases} \tag{4.8}$$

dir. $i = 1,2,...,n$ ve $k = 1,2,...,g$ olmak üzere, $\mathbf{x}_i$ gözlem vektörü ile birlikte karşılık gelen z_{ik} etiketinin oluşturdukları,

$$(\mathbf{x}_i, z_{1i}, z_{2i}, ..., z_{gi}) \tag{4.9}$$

veri kümesine tamamlanmış veri kümesi denir. Bu durumda $\boldsymbol{\Psi}$ için logaritması alınmış tamamlanmış verinin olabilirlik fonksiyonu,

$$\log L_c(\boldsymbol{\Psi}) = \sum_{k=1}^{g} \sum_{i=1}^{n} z_{ki} \left[\log \pi_k + \log f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \right] \tag{4.10}$$

şeklinde olur.

E adım: Tamamlanmış verinin logaritmik olabilirlik fonksiyonu, z_{ki} etiket değerleri cinsinden lineer olduğundan E adımda; $\mathbf{x}$ gözlenmiş değeri verilmişken $\mathbf{Z}_{ki}$ kategorik değişkeninin anlık koşullu beklenen değeri hesaplanır. Burada $\mathbf{Z}_{ki}$, z_{ki}'lere karşılık gelen rastgele değişkenlerdir. $\boldsymbol{\Psi}$ parametre vektörü için $\boldsymbol{\Psi}^{(0)}$ ile ifade edilen başlangıç değeri atanır. EM algoritmasının birinci döngüsünde E adımı için $\mathbf{x}$ verilmişken $\log L_c(\boldsymbol{\Psi})$ ifadesinin koşullu beklenen değeri, $\boldsymbol{\Psi}^{(0)}$ başlangıç değerleri ile

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(0)}) = E_{\boldsymbol{\Psi}^{(0)}} \left\{ \log L_c(\boldsymbol{\Psi}) \mid \mathbf{x} \right\} \tag{4.11}$$

biçiminde hesaplanır. EM algoritmasının $(r+1)$. döngüsündeki E adımında $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(r)})$ ifadesinin bulunmasını gerektirir. Burada $\boldsymbol{\Psi}^{(r)}$, EM algoritmasının r. adımında elde edilen $\boldsymbol{\Psi}$ vektörünün değeridir. EM algoritmasının $(r+1)$. döngüsündeki E adımında $k=1,2,\ldots,g$ ve $i=1,2,\ldots,n$ olmak üzere

$$E_{\boldsymbol{\Psi}^{(k)}}(Z_{ki}\mid \mathbf{x}) = z_{ki}^{(r)} = \frac{\pi_k^{(r)} f_k(\mathbf{x}_i;\boldsymbol{\theta}_k^{(r)})}{\sum_{m=1}^{g} \pi_m^{(r)} f_m(\mathbf{x}_i;\boldsymbol{\theta}_m^{(r)})} \tag{4.12}$$

değeri hesaplanır. Burada z_{ki} ifadesi i. gözlem vektörü $\mathbf{x}_i$'nin karmanın k. bileşenine ait olma olasılığıdır. (4.12)'deki eşitlik kullanılarak (4.11)'deki ifadenin $\mathbf{x}$ verilmişken koşullu beklenen değeri,

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(r)}) = \sum_{k=1}^{g}\sum_{i=1}^{n} z_{ki}^{(r)} \left\{\log \pi_k + \log f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)\right\} \tag{4.13}$$

ile hesaplanır.

M adım: EM algoritmasının $(r+1)$. döngüsündeki M adımında $\boldsymbol{\Omega}$ parametre uzayında tanımlı $\boldsymbol{\Psi}$ vektörünün $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(r)})$ fonksiyonunu maksimum yapan $\boldsymbol{\Psi}^{(r+1)}$ güncellenmiş tahmin değeri bulunur. Sonlu karma olasılık dağılım modelinde π_k'nın $\pi_k^{(r+1)}$ güncel tahmini, bileşen yoğunluklarındaki bilinmeyen parametrelerin vektörü $\boldsymbol{\xi}$'nın güncellenmesinden bağımsız olarak yapılır.

z_{ki}'ler gözlenmiş olsaydı tamamlanmış veri için π_k'nin en çok olabilirlik tahmini,

$$\hat{\pi}_k = \sum_{i=1}^{n} \frac{z_{ki}}{n} \qquad (k=1,2,\ldots,g) \tag{4.14}$$

şeklinde bulunur.

EM algoritmasının $(r+1)$. iterasyonundaki M-adımda $\boldsymbol{\xi}$'nın, $\boldsymbol{\xi}^{(r+1)}$ güncel değeri,

$$\sum_{i=1}^{g}\sum_{j=1}^{n} z_{ki}^{(r)} \frac{\partial \log f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)}{\partial \boldsymbol{\xi}} = \mathbf{0} \tag{4.15}$$

şeklinde elde edilen eşitliğin uygun bir köküne eşittir.

Karma dağılım modeli çok değişkenli normal dağılımların karmasından oluşuyor ise, EM algoritmasının $(r+1)$. iterasyonundaki M-adımında güncellenmiş karma oranları π_k, ve ortalama vektörü $\boldsymbol{\mu}_k$ 'nin en çok olabilirlik kestiricileri sırasıyla,

$$\pi_i^{(k+1)} = \sum_{i=1}^{n} \frac{z_{ki}^{(r)}}{n} \tag{4.16}$$

$$\boldsymbol{\mu}_k^{(r+1)} = \frac{\sum_{i=1}^{n} z_{ki}^{(r)} \mathbf{x}_i}{\sum_{i=1}^{n} z_{ki}^{(r)}} \tag{4.17}$$

ifadeleri ile bulunur. Bileşen olasılık yoğunluklarının kovaryans matrisi $\boldsymbol{\Sigma}_k$ 'nin güncel tahmini de

$$\boldsymbol{\Sigma}_k^{(r+1)} = \frac{\sum_{i=1}^{n} z_{ki}^{(r)} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(r+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(r+1)})^T}{\sum_{i=1}^{n} z_{ki}^{(r)}} \quad (k = 1,2,\ldots,g) \tag{4.18}$$

ifadesi ile hesaplanır.

EM algoritmasında E ve M adımları bir yakınsama kriteri sağlanıncaya kadar tekrarlanır. Yakınsama için uygun bir durdurma kuralı olarak $\|.\|$ uygun bir norm olmak üzere $\left\|\boldsymbol{\Psi}^{(r+1)} - \boldsymbol{\Psi}^{(r)}\right\| < \varepsilon$ ve $\varepsilon > 0$ olma koşulunu sağlayan bir ε değeri alınabilir veya $L(\boldsymbol{\Psi}^{(r+1)}) - L(\boldsymbol{\Psi}^{(r)})$ farkı oldukça küçükse ya da durağanlaşıyorsa algoritma sonlandırılır.

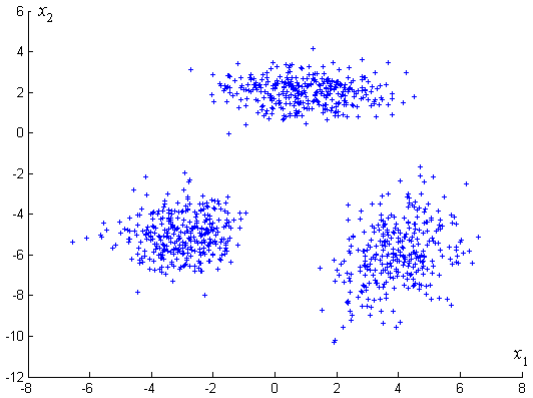
Çok değişkenli normal dağılımların karmasındaki parametrelerin tahmininde EM algoritmasının uygulamasını göstermek için

$$\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix} \quad p_1 = \frac{1}{3}$$

$$\mu_2 = \begin{bmatrix} -3 \\ -5 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad p_2 = \frac{1}{3}$$

$$\mu_3 = \begin{bmatrix} 4 \\ -6 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix} \quad p_3 = \frac{1}{3}$$

parametreleriyle her biri çok değişkenli normal dağılıma sahip 3 bileşenden oluşan çok değişkenli normal dağılımdan rassal olarak 1000 birimlik örneklem üretilmiştir. Gözlemlerin saçılım grafiği Şekil 4.3' de verilmiştir.



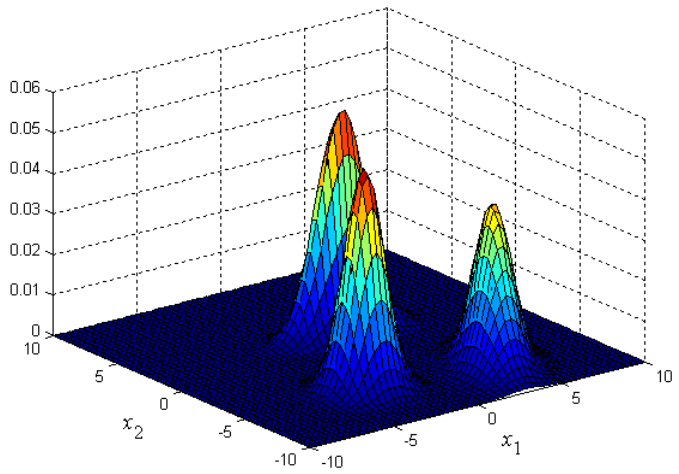
Şekil 4.3. Çok değişkenli normal dağılımların karmasından rassal olarak üretilen veri setinin saçılım grafiği

Rassal olarak üretilen veri setinin çok değişkenli normal dağılımların karması ile modellenmesinde EM algoritması ile elde edilen parametre tahminleri Çizelge 4.1'de verilmiştir.

Çizelge 4.1. EM algoritması ile elde edilen parametrelerin en çok olabilirlik tahminleri

	Bileşen 1		**Bileşen 2**		**Bileşen 3**	
$\hat{p}_k$	0.3333		0.3333		0.3333	
$\hat{\mu}_k$	0.9679	2.0138	-3.0014	-5.0266	3.9841	-5.9610
$\hat{\Sigma}_k$	1.8369	-0.0492	0.9298	0.0769	0.9928	0.4907
	-0.0492	0.4363	0.0769	0.9561	0.4907	2.2551

Çizelge 4.1'deki parametre tahmin değerlerine göre Matlab programı ile elde edilen çok değişkenli normal dağılımların karmasının olasılık yoğunluk fonksiyonun grafiği Şekil 4.4'de verilmiştir.

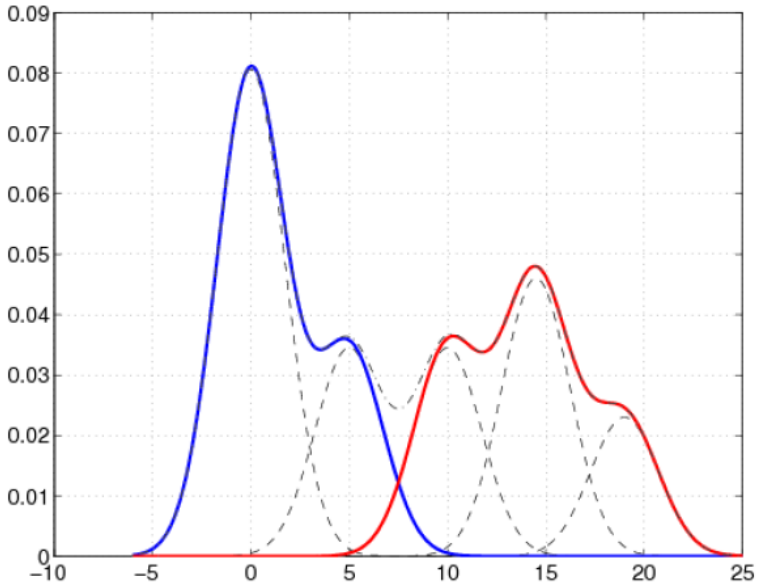


Şekil 4.4. Çok değişkenli normal dağılımların karmasının olasılık yoğunluk fonksiyonu

4.2. Karma Dağılım Modellerine Dayalı Ayrıştırma Analizi

Ayrıştırma analizinin temel yapısı, gruplar için normal dağılım modeli varsayımına dayalıdır. Gruplar için, normal dağılıma sahip olmayan küme yapıları söz konusu olduğunda, bu kümeleri en uygun biçimde sınıflara ayırmak için karma dağılım modellerine dayalı ayrıştırma analizi yöntemleri geliştirilebilir. Veri grupları için tek bir normal dağılımla modellemenin yeterli olmadığı durumlarda, normal dağılımların karmaları kullanılarak veri yapısına daha uygun modeller oluşturulabilir. Hastie ve Tibshirani (1996) özellikle verideki kümelenmelerin belirgin olduğu durumda normal olmayan bu veri kümesini etkili bir şekilde sınıflandırmak için karma normal dağılım yaklaşımını incelemişlerdir. Hastie ve Tibshirani, bu çalışmalarında verideki her bir alt kümenin μ_{kr} ortalamalı ve Σ ortak kovaryans matrisli çok değişkenli normal dağılıma sahip olduğu varsayımı ile veri için çok değişkenli karma normal dağılım modeli oluşturmuşlardır. Oluşturdukları modele dayalı olarak ta veriyi sınıflandırarak karma normal dağılım modellerine dayalı karma ayrıştırma analizini yapmışlardır.

Karma dağılım modellerine dayalı ayrıştırma analizinin daha iyi anlaşılması için tek değişkenli normal dağılımların karmalarının karması ile elde edilen olasılık yoğunluk fonksiyonunun grafiği Şekil 4.5'de verilmiştir.



Şekil 4.5. Normal dağılımların karmalarının karmasından oluşan karma dağılım modelinin olasılık yoğunluk fonksiyonu eğrisi (Li, 2005).

Şekil 4.3'e göre iki alt sınıftan oluşan kitlenin alt sınıflarından biri iki bileşenli normal dağılımla, diğerinin de üç bileşenli normal dağılımla modellendiği görülmektedir. Karma dağılım modellerine dayalı ayrıştırma analizinde daha çok çok değişkenli normal dağılımlarının karması kullanılmaktadır.

Çok değişkenli normal dağılımların karmasına dayalı karma ayrıştırma analizinde $n \times p$ boyutlu $\mathbf{X}$ rastgele vektörünün her alt sınıfı çok değişkenli normal dağılımların karması ile modellenir. Bu durumda yığını oluşturan k. alt sınıfın olasılık yoğunluk fonksiyonu

$$f_k(\mathbf{x}|\mu_k, \Sigma_k) = \sum_{r=1}^{R_k} \pi_{kr} \Phi_{kr}(\mathbf{x}; \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}_{kr}) \tag{4.19}$$

şeklinde ifade edilir. Eşitlikte yer alan π_{kr}, k. alt sınıfın r. alt sınıfının karma oranını, $\boldsymbol{\mu}_{kr}$, k. alt sınıfın r. alt sınıfına ait ortalama vektörünü ve $\boldsymbol{\Sigma}_{kr}$, k. alt sınıfın r. alt sınıfına ait kovaryans matrisini göstermektedir.

Çok değişkenli normal dağılımların karmalarının karmasına dayanan karma ayrıştırma analizinde Eşitlik (4.19)'da verilen modelde k. alt sınıfın r. alt sınıfına ait kovaryans matrisleri farklı yapılarda kullanılarak karma ayrıştırma analizi gerçekleştirilmektedir. Bu yapılar içersinde alt kitlelerin karma dağılım ile

modellenmesinde $k.$ alt sınıfın $r.$ alt sınıfına ait kovaryans matrisleri eşit kabul edilerek yani ortak kovaryans varsayımı ile ayrıştırma analizi gerçekleştirilebilmektedir.

Çok değişkenli normal dağılımların karmasına dayalı karma ayrıştırma analizinde $\mathbf{x}_i$ gözlem vektörünün $k.$ sınıfa ait olma olasılığı

$$p_{i,k} = \frac{\pi_k \sum_{r=1}^{R_k} \pi_{kr} \Phi_{kr}(\mathbf{x}; \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}_{kr})}{\sum_{m=1}^{k} \pi_m \sum_{r=1}^{R_k} \pi_{mr} \Phi_{mr}(\mathbf{x}; \boldsymbol{\mu}_{mr}, \boldsymbol{\Sigma}_{mr})} \tag{4.20}$$

eşitliği ile hesaplanır. Birimlerin sınıflara atanmasında farklı yaklaşımlar bulunmakla birlikte en yaygın kullanıma sahip olanı birimlerin alt sınıflara ait olma olasılıklarına göre atanması yaklaşımıdır. Bu yaklaşıma göre $\mathbf{x}_i$ gözlem vektörüne sahip birim en yüksek ait olma olasılığa sahip alt sınıfa atanır.

4.3. Uygulama

İris veri kümesi (Raftery ve Dean (2006) sınıflandırma algoritmalarında kullanılan İris süsen çiçeğinin üç farklı türüne ait 150 gözlemlik bir veri kümesidir (Murphy ve Aha, 1995). İris veri kümesinde eğitim verisi, İris setosa türüne ait 24, İris versicolor türüne ait 26 ve İris Virginia türüne ait 25 gözlem olmak üzere toplam 75 gözlemden oluşmaktadır. Test veri kümesi geriye kalan 75 gözlemden oluşmaktadır (Çalış, 2011). İris verisi çanak yaprağı uzunluğu ve genişliği, taç yaprağı uzunluğu ve genişliği olmak üzere dört değişkenden oluşmaktadır.

İris verisinin karma ayrıştırma analizi yapılırken eğitim verisindeki her bir grup için karma normal dağılım modelleri oluşturulmuştur. Oluşturulan karma normal dağılım modelleri ayrıştırma analizi için kullanılmıştır. Test veri kümesindeki her bir gözlemin alt kümelerden hangisine ait olduğuna maksimum aitlik olasılıklarına bakılarak karar verilmiştir.

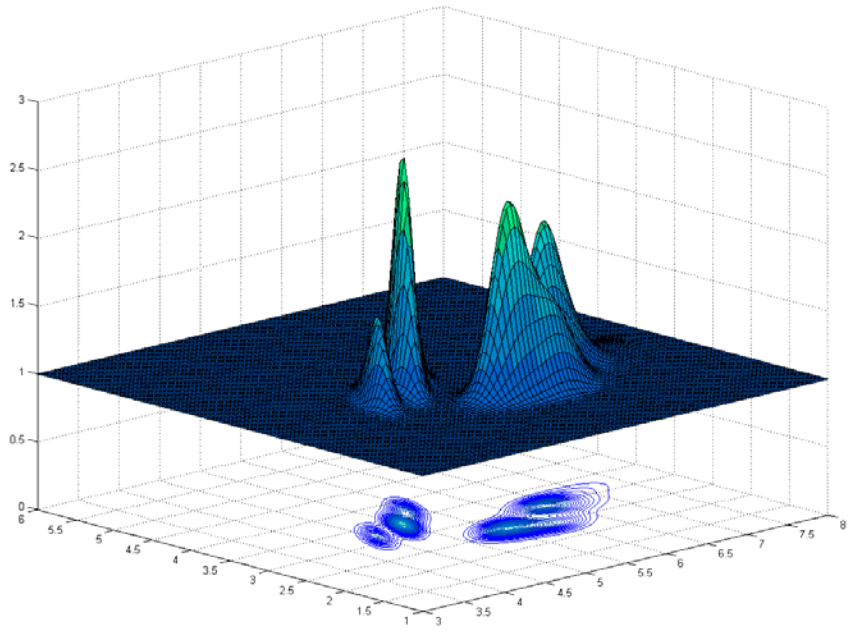
İris verisi için, eğitim veri kümesindeki her bir alt grup için ortak kovaryans matris yapısı ($\Sigma_{1r} = \Sigma_{2r}$) varsayımı altında EM algoritmasıyla elde edilen parametre tahminleri kullanılarak gerçekleştirilen karma ayrıştırma analizi sonucunda yanlış sınıflandırma oranı %5.33 olarak bulunmuştur. İris verisi için, ortak kovaryans matris

yapısı varsayımı altında EM Algoritması kullanılarak yapılan karma ayrıştırma analizi sonuçları Çizelge 4.2'de verilmiştir.

Çizelge 4.2. İris verisi için EM Algoritması kullanılarak ortak kovaryans matris yapısı varsayımı altında yapılan karma ayrıştırma analiz sonuçları

Yöntem		**Setosa**	**Versicolor**	**Virgina**	**Yanlış Sın. Oranı**
EM Algoritması-MDA (Ortak Kovaryans)	**Setosa**	26	0	0	**5.33%**
	Versicolor	0	20	4	
	Virginia	0	0	25	

Ortak kovaryans matris yapısı ($\Sigma_{1r} = \Sigma_{2r}$) varsayımı altında EM algoritmasıyla elde edilen parametre tahminleri kullanılarak oluşturulan karma dağılım modellerine ait olasılık yoğunluk fonksiyonu için Matlab programında oluşturulan grafik Şekil 4.6'da verilmiştir.



Şekil 4.6. İris verisi için ortak kovaryans matris yapısı varsayımı altında EM Algoritması kullanılarak oluşturulan karma dağılım modeline ait olasılık yoğunluk fonksiyonu için oluşturulan grafik

5. EKSİK GÖZLEM ÇÖZÜM YÖNTEMLERİNİN KARMA AYRIŞTIRMA ANALİZİNDEKİ ETKİNLİĞİN İNCELENMESİ

Eksik gözlem tahmin yöntemlerinin karma ayrıştırma analizinde etkinliği incelenirken literatürde yaygın bir kullanıma sahip olan cam (glass), iris ve şarap (wine) verileri kullanılacaktır. Çalışmada öncelikle tam veri durumunda karma ayrıştırma analizi ile hatalı sınıflandırma yüzdeleri belirlenecektir. Tam veri durumunda karma ayrıştırma analizi gerçekleştirildikten sonra veri setlerinde %5, %10 ve %15 eksik gözlemler oluşturulacaktır.

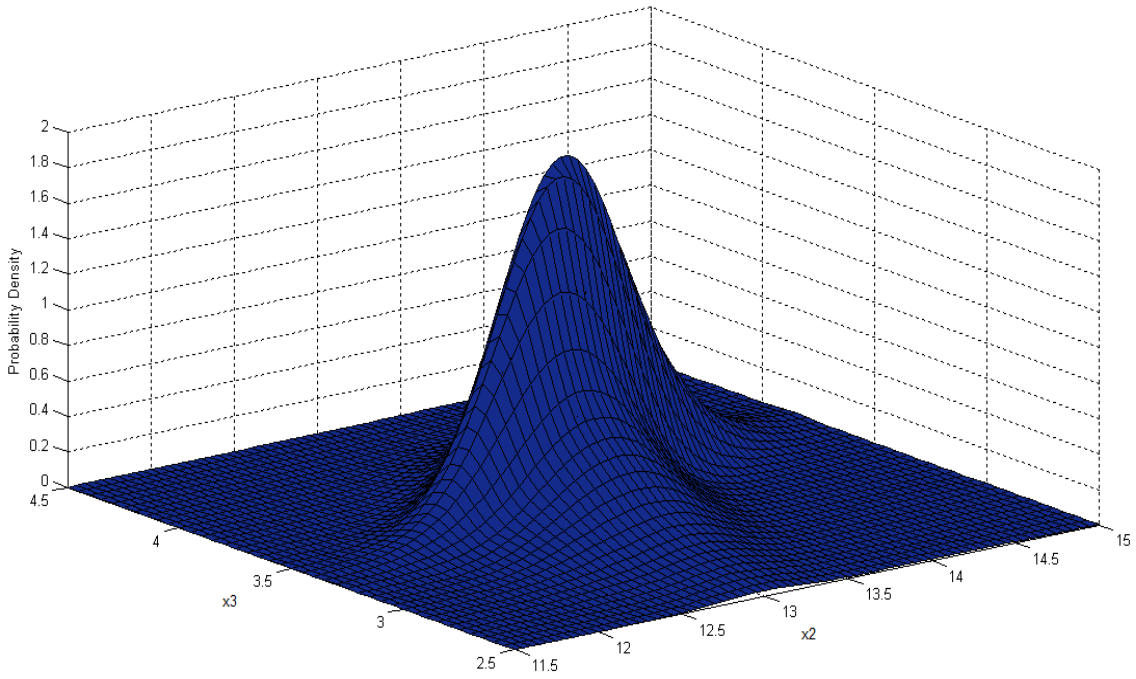
Eksik gözlemli veri setlerindeki eksik veri problemi liste bazında silme ile ortalama atama, grup ortalamalarının atanması, regresyon atama ve EM atama yöntemleriyle çözümlenerek tamamlanmış veri durumuna getirilecektir. Tamamlanmış veri durumuna getirilen veri setlerinde gerçekleştirilen karma ayrıştırma analizi sonucu elde edilen hatalı sınıflandırma yüzdelikleri göz önünde bulundurularak eksik veri probleminin çözümünde farklı eksik gözlem oranlarında hangi yöntemlerin etkin olduğu belirlenecektir. Karma ayrıştırma analizinde grupların alt sınıflarının kovaryans matrislerinin farklı olduğu varsayılacaktır.

5.1. Cam (Glass) Verisi

Cam veri kümesi (http://archive.ics.uci.edu/ml/datasets/ Glass+Identification) adli araştırmalarda farklı cam türlerinin belirlenmesinde kullanılan 163 gözlemlik bir veri kümesidir (Murphy ve Aha, 1995). Cam veri kümesinde eğitim verisi, düz pencere camına ait 70 birim ve düz olmayan pencere camına ait 61 birim olmak üzere toplam 131 gözlemden oluşmaktadır. Test veri kümesi geriye kalan 32 gözlemden oluşmaktadır. Cam verisinde dokuz değişken bulunmaktadır. Bu çalışmada 8. ve 9. değişkenler kovaryans matrisinin hesaplanmasında sorun teşkil ettiğinden dolayı ayrıştırma analizinde ilk yedi değişken kullanılacaktır.

Cam verisinin karma ayrıştırma analizi yapılırken eğitim verisindeki her bir grup için iki alt grubun olduğu varsayılmıştır. Test veri kümesindeki her bir gözlemin alt kümelerden hangisine ait olduğuna maksimum aitlik olasılıklarına bakılarak karar verilmiştir. Cam verisi için ayrıca, eğitim veri kümesindeki her bir alt grup için farklı kovaryans matrisleri ($\Sigma_{1r} \neq \Sigma_{2r}$) varsayımı altında, EM algoritmasıyla elde edilen parametre tahminleri kullanılarak karma ayrıştırma analizi yapılmıştır. Farklı kovaryans

matris yapısı ($\Sigma_{1r} \neq \Sigma_{2r}$) varsayımı altında EM algoritmasıyla elde edilen parametre tahminleri kullanılarak oluşturulan karma dağılım modellerine ait olasılık yoğunluk fonksiyonu için Matlab programında oluşturulan grafik Şekil 5.1'de verilmiştir.



Şekil 5.1. Cam verisi için farklı kovaryans matris yapısında oluşturulan karma dağılım modeline ait olasılık yoğunluk fonksiyonu için oluşturulan grafik

Cam verisinde tam veri durumu için eğitim verisi modellendikten sonra oluşturulan model ile test verisinin sınıflandırılması gerçekleştirilmiştir. Test verisinin sınıflandırılması sonucunda %9.924'lük bir hatalı sınıflandırma yüzdesi elde edilmiştir.

Tam veri durumunda eğitim verisinin modellenmesi ve test verisinin sınıflandırılması sonrasında eğitim verisinde ikinci değişken dışındaki tüm değişkenlerden rassal olarak yaklaşık %5, %10 ve %15'lik oranlarda eksik veri yapısı oluşturulmuştur. Eksik veri çözümünde kullanılan yöntemler ile veri setleri tamamlanmış veri durumuna getirilerek karma ayrıştırma analizi gerçekleştirilmiştir. Karma ayrıştırma analizinde gruplardan her birinin iki alt gruba sahip olduğu varsayılmıştır. Cam veri setinde eksik gözlem oranı %5 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri Çizelge 5.1'de verilmiştir.

Çizelge 5.1. Cam veri setinde eksik gözlem oranı %5 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri

Tam Veri Durumu		
Eğitim	**I**	**II**
I	64	6
II	7	54
Hatalı Sınıf. Yüzdesi=		**9.924**
Test	**I**	**II**
I	16	1
II	1	14
Hatalı Sın. Yüzdesi=		**6.25**

EM Atama		
Eğitim	**I**	**II**
I	65	5
II	8	53
Hatalı Sınıf. Yüzdesi=		**9.924**
Test	**I**	**II**
I	17	0
II	1	14
Hatalı Sın. Yüzdesi=		**3.125**

Regresyon Atama		
Eğitim	**I**	**II**
I	68	2
II	18	43
Hatalı Sın. Yüzdesi=		**15.267**
Test	**I**	**II**
I	17	0
II	4	11
Hatalı Sınıf. Yüzdesi=		**12.5**

Grup Ortalamalarının Atanması		
Eğitim	**I**	**II**
I	63	7
II	6	55
Hatalı Sın. Yüzdesi=		**9.924**
Test	**I**	**II**
I	14	3
II	0	15
Hatalı Sınıf. Yüzdesi=		**9.375**

Ortalama Atama		
Eğitim	**I**	**II**
I	59	11
II	9	52
Hatalı Sınıf. Yüzdesi=		**15.267**
Test	**I**	**II**
I	17	0
II	4	11
Hatalı Sınıf. Yüzdesi=		**12.5**

Liste Bazında Silme		
Eğitim	**I**	**II**
I	53	1
II	16	27
Hatalı Sınıf. Yüzdesi=		**17.526**
Test	**I**	**II**
I	17	0
II	8	7
Hatalı Sınıf. Yüzdesi=		**25**

Cam veri setine tam veri durumunda uygulanan karma ayrıştırma analizi sonrası test verisinin sınıflandırılmasındaki hatalı sınıflandırma yüzdesi %6.25'dir. Çizelge 5.1 incelendiğinde test veri setinin sınıflandırılmasında hatalı sınıflandırma yüzdesi bakımından en başarılı yöntem %3.125 hatalı sınıflandırma yüzdesi ile EM atama yöntemi olmuştur. EM atama yönteminden sonraki en başarılı yöntem %9.375'lik hatalı sınıflandırma yüzdesi ile grup ortalamalarının atanması yöntemi olmuştur. Test veri setinin sınıflandırılmasında en kötü performans gösteren yöntem ise %25'lik hatalı sınıflandırma yüzdesi ile liste bazında silme yöntemi olmuştur.

Cam veri setinde eksik gözlem oranı %10 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri Çizelge 5.2'de verilmiştir.

Çizelge 5.2. Cam veri setinde eksik gözlem oranı %10 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri

Tam Veri Durumu		
Eğitim	**I**	**II**
I	64	6
II	7	54
Hatalı Sınıf. Yüzdesi=		**9.924**
Test	**I**	**II**
I	16	1
II	1	14
Hatalı Sınıf. Yüzdesi=		**6.25**

EM Atama		
Eğitim	**I**	**II**
I	62	8
II	12	49
Hatalı Sınıf. Yüzdesi=		**15.267**
Test	**I**	**II**
I	16	1
II	2	13
Hatalı Sınıf. Yüzdesi=		**9.375**

Regresyon Atama		
Eğitim	**I**	**II**
I	63	7
II	20	41
Hatalı Sınıf. Yüzdesi=		**20.611**
Test	**I**	**II**
I	15	2
II	1	14
Hatalı Sınıf. Yüzdesi=		**9.375**

Grup Ortalamalarının Atanması		
Eğitim	**I**	**II**
I	63	7
II	12	49
Hatalı Sınıf. Yüzdesi=		**14.504**
Test	**I**	**II**
I	17	0
II	1	14
Hatalı Sınıf. Yüzdesi=		**3.125**

Ortalama Atama		
Eğitim	**I**	**II**
I	62	8
II	11	50
Hatalı Sınıf. Yüzdesi=		**14.504**
Test	**I**	**II**
I	17	0
II	2	13
Hatalı Sınıf. Yüzdesi=		**6.25**

Liste Bazında Silme		
Eğitim	**I**	**II**
I	31	8
II	7	27
Hatalı Sınıf. Yüzdesi=		**20.548**
Test	**I**	**II**
I	16	1
II	3	12
Hatalı Sınıf. Yüzdesi=		**12.5**

Çizelge 5.2 incelendiğinde test veri setinin sınıflandırılmasında hatalı sınıflandırma yüzdesi bakımından en başarılı yöntem %3.125 hatalı sınıflandırma yüzdesi ile grup ortalamalarının atanması yöntemi olmuştur. Grup ortalamalarının atanması yönteminden sonraki en başarılı yöntem %6.25'lik hatalı sınıflandırma yüzdesi ile ortalama atama yöntemi olmuştur. Test veri setinin sınıflandırılmasında en kötü performans gösteren yöntem ise %12.5'lik hatalı sınıflandırma yüzdesi ile liste bazında silme yöntemi olmuştur.

Cam veri setinde eksik gözlem oranı %15 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri Çizelge 5.3'de verilmiştir.

Çizelge 5.3. Cam veri setinde eksik gözlem oranı %15 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri

Tam Veri Durumu		
Eğitim	**I**	**II**
I	64	6
II	7	54
Hatalı Sınıf. Yüzdesi=		9.924
Test	**I**	**II**
I	16	1
II	1	14
Hatalı Sınıf. Yüzdesi=		6.25

EM Atama		
Eğitim	**I**	**II**
I	63	7
II	7	54
Hatalı Sınıf. Yüzdesi=		10.687
Test	**I**	**II**
I	16	1
II	2	13
Hatalı Sınıf. Yüzdesi=		9.375

Regresyon Atama		
Eğitim	**I**	**II**
I	64	6
II	7	54
Hatalı Sın. Yüzdesi=		9.924
Test	**I**	**II**
I	14	3
II	1	14
Hatalı Sınıf. Yüzdesi=		12.5

Grup Ortalamalarının Atanması		
Eğitim	**I**	**II**
I	62	8
II	7	54
Hatalı Sınıf. Yüzdesi=		11.450
Test	**I**	**II**
I	14	3
II	1	14
Hatalı Sınıf. Yüzdesi=		12.5

Ortalama Atama		
Eğitim	**I**	**II**
I	62	8
II	9	52
Hatalı Sınıf. Yüzdesi=		12.977
Test	**I**	**II**
I	14	3
II	1	14
Hatalı Sınıf. Yüzdesi=		12.5

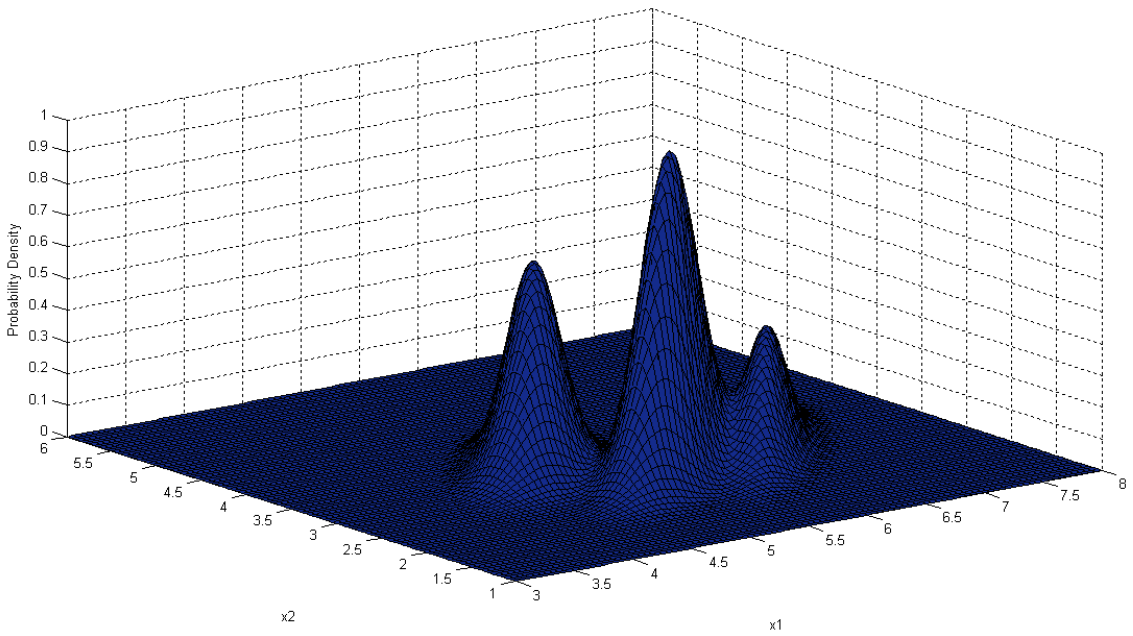
Liste Bazında Silme		
Eğitim	**I**	**II**
I	28	0
II	3	21
Hatalı Sın. Yüzdesi=		5.769
Test	**I**	**II**
I	16	1
II	4	11
Hatalı Sınıf. Yüzdesi=		15.625

Çizelge 5.3 incelendiğinde test veri setinin sınıflandırılmasında hatalı sınıflandırma yüzdesi bakımından en başarılı yöntem %9.375 hatalı sınıflandırma yüzdesi ile EM atama yöntemi olmuştur. Test veri setinin sınıflandırılmasında en kötü performans gösteren yöntem ise %15.625'lik hatalı sınıflandırma yüzdesi ile liste bazında silme yöntemi olmuştur. Cam veri setindeki %15'lik eksik gözlem oranında regresyon atama yöntemi, ortalama atama yöntemi ve grup ortalamalarının atanması yöntemi %12.5'lik hatalı sınıflandırma yüzdesine sahiptir.

Cam verisinde eksik gözlem probleminin çözümünde kullanılan yöntemlerin karma ayrıştırma analizindeki etkinliği incelendiğinde EM atama yönteminin diğer yöntemlere göre daha başarılı olduğu ve tüm eksik gözlem oranlarında en başarısız yöntemin liste bazında silme yöntemi olduğu görülmüştür.

5.2. İris Verisi

İris veri kümesi (Raftery ve Dean (2006) sınıflandırma algoritmalarında kullanılan İris süsen çiçeğinin üç farklı türüne ait 150 gözlemlik bir veri kümesidir (Murphy ve Aha, 1995). İris veri kümesinde eğitim verisi, her bir türden 40'ar gözlem olmak üzere toplam 120 gözlemden oluşmaktadır. Test veri kümesi geriye kalan 30 gözlemden oluşmaktadır. İris verisi dört değişkenden oluşmaktadır. Eksik gözlemler oluşturulurken üçüncü değişken dışındaki diğer değişkenlerden rassal olarak %5, %10 ve %15'lik gözlem eksiltilmiştir.İris verisinin karma ayrıştırma analizi yapılırken eğitim verisindeki her bir grubun iki alt sınıftan oluştuğu varsayılmıştır. Oluşturulan karma normal dağılım modelleri ayrıştırma analizi için kullanılmıştır. Test veri kümesindeki her bir gözlemin alt kümelerden hangisine ait olduğuna maksimum aitlik olasılıklarına bakılarak karar verilmiştir. İris verisi için ayrıca, eğitim veri kümesindeki her bir alt grup için farklı kovaryans matrisleri ($\Sigma_{1r} \neq \Sigma_{2r}$) varsayımı altında, EM algoritmasıyla elde edilen parametre tahminleri kullanılarak karma ayrıştırma analizi yapılmıştır. Farklı kovaryans matris yapısı ($\Sigma_{1r} \neq \Sigma_{2r}$) varsayımı altında EM algoritmasıyla elde edilen parametre tahminleri kullanılarak oluşturulan karma dağılım modellerine ait olasılık yoğunluk fonksiyonu için Matlab programında oluşturulan grafik Şekil 5.2'de verilmiştir.



Şekil 5.2. İris verisi için farklı kovaryans matris yapısı varsayımı altında EM Algoritması kullanılarak oluşturulan karma dağılım modeline ait olasılık yoğunluk fonksiyonu için oluşturulan grafik

İris veri setinde eksik gözlem oranı %5 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri Çizelge 5.4'de verilmiştir.

Çizelge 5.4. İris veri setinde eksik gözlem oranı %5 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri

Tam Veri Durumu			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	0	40
Hatalı Sınıf. Yüzdesi=			**0.833**
Test	**I**	**II**	**III**
I	10	0	0
II	0	8	2
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**6.667**

EM Atama			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	1	39
Hatalı Sınıf. Yüzdesi=			**1.667**
Test	**I**	**II**	**III**
I	10	0	0
II	0	7	3
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**10**

Regresyon Atama			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	0	40
Hatalı Sınıf. Yüzdesi=			**0.833**
Test	**I**	**II**	**III**
I	10	0	0
II	0	8	2
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**6.667**

Grup Ortalamalarının Atanması			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	0	40
Hatalı Sınıf. Yüzdesi=			**0.833**
Test	**I**	**II**	**III**
I	10	0	0
II	0	9	1
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**3.333**

Ortalama Atama			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	0	40
Hatalı Sınıf. Yüzdesi=			**0.833**
Test	**I**	**II**	**III**
I	10	0	0
II	0	7	3
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**10**

Liste Bazında Silme			
Eğitim	**I**	**II**	**III**
I	35	0	0
II	0	33	1
III	0	1	36
Hatalı Sınıf. Yüzdesi=			**1.887**
Test	**I**	**II**	**III**
I	10	0	0
II	0	7	3
III	0	1	9
Hatalı Sınıf. Yüzdesi=			**13.333**

İris veri setine tam veri durumunda uygulanan karma ayrıştırma analizi sonrası test verisinin sınıflandırılmasındaki hatalı sınıflandırma yüzdesi %6.667'dir. Çizelge 5.4 incelendiğinde test veri setinin sınıflandırılmasında hatalı sınıflandırma yüzdesi bakımından en başarılı yöntem %3.333 hatalı sınıflandırma yüzdesi ile grup ortalamalarının atanması yöntemi olmuştur. Grup ortalamalarının atanması

yönteminden sonraki en başarılı yöntem %6.667'lik hatalı sınıflandırma yüzdesi ile regresyon atama yöntemi olmuştur. Test veri setinin sınıflandırılmasında en kötü performans gösteren yöntem ise %13.333'lük hatalı sınıflandırma yüzdesi ile liste bazında silme yöntemi olmuştur.

İris veri setinde eksik gözlem oranı %10 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri Çizelge 5.5'de verilmiştir.

Çizelge 5.5. İris veri setinde eksik gözlem oranı %10 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri

Tam Veri Durumu			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	0	40
Hatalı Sınıf. Yüzdesi=			**0.833**
Test	**I**	**II**	**III**
I	10	0	0
II	0	8	2
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**6.667**

EM Atama			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	0	40
Hatalı Sınıf. Yüzdesi=			**0.833**
Test	**I**	**II**	**III**
I	10	0	0
II	0	8	2
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**6.667**

Regresyon atama			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	0	40
Hatalı Sınıf. Yüzdesi=			**0.833**
Test	**I**	**II**	**III**
I	10	0	0
II	0	9	1
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**3.333**

Grup Ortalamalarının Atanması			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	0	40
Hatalı Sınıf. Yüzdesi=			**0.833**
Test	**I**	**II**	**III**
I	10	0	0
II	0	9	1
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**3.333**

Ortalama Atama			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	0	40
Hatalı Sınıf. Yüzdesi=			**0.833**
Test	**I**	**II**	**III**
I	10	0	0
II	0	7	3
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**10**

Liste Bazında Silme			
Eğitim	**I**	**II**	**III**
I	28	0	0
II	0	28	1
III	0	1	26
Hatalı Sınıf. Yüzdesi=			**2.381**
Test	**I**	**II**	**III**
I	10	0	0
II	0	7	3
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**10**

Çizelge 5.5 incelendiğinde test veri setinin sınıflandırılmasında hatalı sınıflandırma yüzdesi bakımından en başarılı yöntem %3.333 hatalı sınıflandırma yüzdesi ile grup ortalamalarının atanması yöntemi ve regresyon atama yöntemi olmuştur. Test veri setinin sınıflandırılmasında en kötü performans gösteren yöntem ise %10'luk hatalı sınıflandırma yüzdesi ile ortalama atama yöntemi ve liste bazında silme yöntemi olmuştur. İris veri setinde eksik gözlem oranı %15 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri Çizelge 5.6'de verilmiştir.

Çizelge 5.6. İris veri setinde eksik gözlem oranı %15 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri

Tam Veri Durumu			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	0	40
Hatalı Sınıf. Yüzdesi=			**0.833**
Test	**I**	**II**	**III**
I	10	0	0
II	0	8	2
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**6.667**

EM Atama			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	1	39
Hatalı Sınıf. Yüzdesi=			**1.667**
Test	**I**	**II**	**III**
I	10	0	0
II	0	9	1
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**3.333**

Regresyon Atama			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	1	39
Hatalı Sınıf. Yüzdesi=			**1.667**
Test	**I**	**II**	**III**
I	10	0	0
II	0	8	2
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**6.667**

Grup Ortalamalarının Atanması			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	40	0
III	0	1	39
Hatalı Sınıf. Yüzdesi=			**0.833**
Test	**I**	**II**	**III**
I	10	0	0
II	0	8	2
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**6.667**

Ortalama Atama			
Eğitim	**I**	**II**	**III**
I	40	0	0
II	0	39	1
III	0	1	39
Hatalı Sınıf. Yüzdesi=			**1.667**
Test	**I**	**II**	**III**
I	10	0	0
II	0	7	3
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**10**

Liste Bazında Silme			
Eğitim	**I**	**II**	**III**
I	26	0	0
II	0	22	2
III	0	2	23
Hatalı Sınıf. Yüzdesi=			**5.333**
Test	**I**	**II**	**III**
I	10	0	0
II	0	6	4
III	0	0	10
Hatalı Sınıf. Yüzdesi=			**13.333**

Çizelge 5.6 incelendiğinde test veri setinin sınıflandırılmasında hatalı sınıflandırma yüzdesi bakımından en başarılı yöntem %3.333 hatalı sınıflandırma yüzdesi ile EM atama yöntemi olmuştur. EM atama yönteminden sonraki en başarılı yöntem %6.667'lik hatalı sınıflandırma yüzdesi ile regresyon atama yöntemi ve grup ortalamalarının atanması yöntemi olmuştur. Test veri setinin sınıflandırılmasında en kötü performans gösteren yöntem ise %13.333'lük hatalı sınıflandırma yüzdesi ile liste bazında silme yöntemi olmuştur.

İris verisinde eksik gözlem probleminin çözümünde kullanılan yöntemlerin karma ayrıştırma analizindeki etkinliği incelendiğinde %5 ve %10'luk eksik gözlem oranında grup ortalamalarının atanması yöntemi, %15'lik eksik gözlem oranında EM atama yöntemi en başarılı yöntemler olarak tespit edilmiştir. Tüm eksik gözlem oranlarında en başarısız yöntem ise liste bazında silme yöntemi olmuştur.

5.3. Şarap (Wine) Verisi

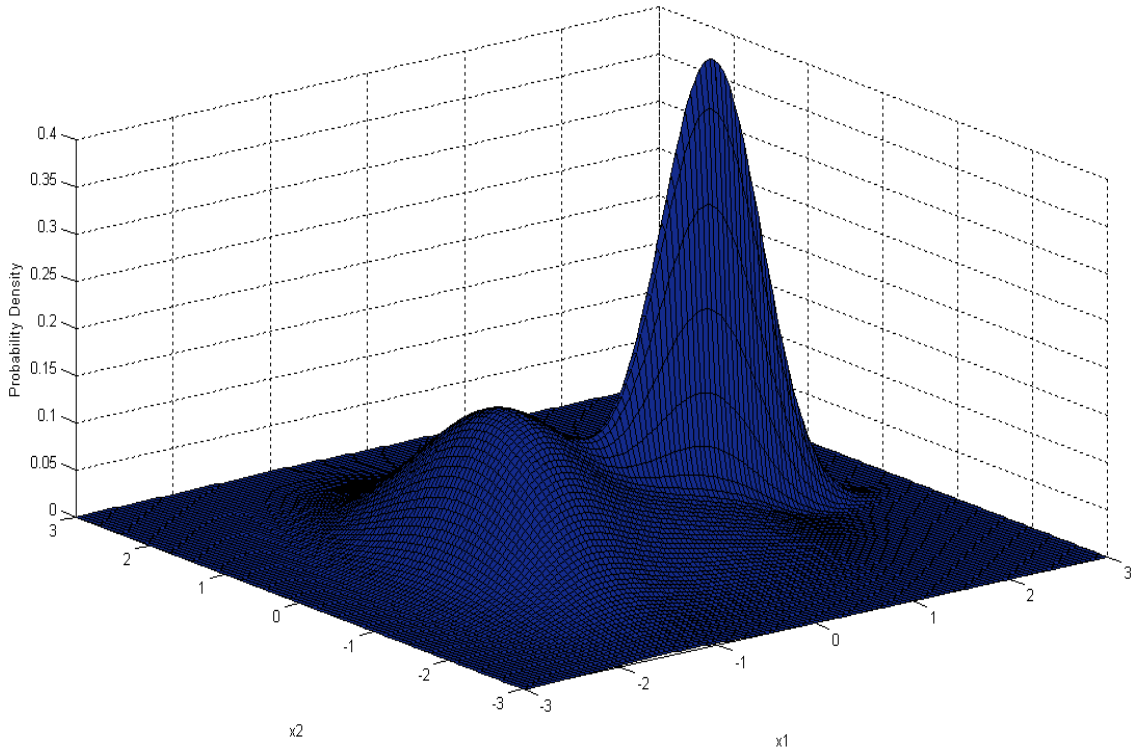
Şarap (Wine) veri kümesi (http://archive.ics.uci.edu/ml/datasets/Wine) sınıflandırma algoritmalarında kullanılan yöntemlerin karşılaştırılmasında yaygın olarak kullanılan on üç değişkenli ve 178 gözlemden oluşan bir veri kümesidir. Literatür incelendiğinde Şarap verisinin sınıflandırılmasında ham veri yerine standartlaştırılmış veri kullanıldığında daha düşük hatalı sınıflandırma yüzdesi elde edilmektedir. Bundan dolayı şarap verisine karma ayrıştırma analizi uygulanmadan önce standartlaştırma işlemi yapılmış ve ham veri yerine standartlaştırılmış veri ile çalışılmıştır. Şarap veri kümesinde 59, 71 ve 48 gözlem içeren üç alt grup bulunmaktadır. Şarap veri kümesinde eğitim verisi, birinci alt gruba ait 47 ikinci alt gruba ait 59 ve üçüncü alt gruba ait 37 gözlem olmak üzere toplam 143 gözlemden oluşmaktadır. Test veri kümesi geriye kalan 35 gözlemden oluşmaktadır.

Şarap verisinde eksik gözlem oluşturulurken birinci değişken dışındaki diğer değişkenlerin her birinde %5, %10 ve %15'lik eksik gözlemler oluşturulmuştur.

Şarap verisinin karma ayrıştırma analizi yapılırken eğitim verisindeki her bir grup için iki bileşenli karma çok değişkenli normal dağılım modelleri oluşturulmuştur. Oluşturulan karma normal dağılım modelleri ayrıştırma analizi için kullanılmıştır. Test veri kümesindeki her bir gözlemin alt kümelerden hangisine ait olduğuna maksimum aitlik olasılıklarına bakılarak karar verilmiştir.

Şarap verisinde eğitim verisine uygulanan karma ayrıştırma analizi sonucu oluşturulan model ile tüm gözlemler doğru olarak sınıflandırılmış ve hatalı sınıflandırma yüzdesi %0 olarak hesaplanmıştır. Eğitim verisinden elde edilen karma çok değişkenli normal dağılımların karması ile test verisinin sınıflandırılmasında 2 gözlem hatalı olarak sınıflandırılmış ve hatalı sınıflandırma yüzdesi %5.714 olarak hesaplanmıştır.

Şarap verisi için, eğitim veri kümesindeki her bir alt grup için farklı kovaryans matrisleri ($\Sigma_{1r} \neq \Sigma_{2r}$) varsayımı altında, EM algoritmasıyla elde edilen parametre tahminleri kullanılarak karma ayrıştırma analizi yapılmıştır. Farklı kovaryans matris yapısı ($\Sigma_{1r} \neq \Sigma_{2r}$) varsayımı altında EM algoritmasıyla elde edilen parametre tahminleri kullanılarak oluşturulan karma dağılım modellerine ait olasılık yoğunluk fonksiyonu için Matlab programında oluşturulan grafik Şekil 5.3'de verilmiştir.



Şekil 5.3. Şarap verisi için farklı kovaryans matris yapısı varsayımı altında EM Algoritması kullanılarak oluşturulan karma dağılım modeline ait olasılık yoğunluk fonksiyonu için oluşturulan grafik

Şarap veri setinde eksik gözlem oranı %5 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri Çizelge 5.7'de verilmiştir.

Çizelge 5.7. Şarap veri setinde eksik gözlem oranı %5 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri

Tam Veri Durumu			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	11	1	0
II	0	12	0
III	0	1	10
Hatalı Sınıf. Yüzdesi=			**5.714**

EM Atama			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	12	0	0
II	0	12	0
III	0	1	10
Hatalı Sınıf. Yüzdesi=			**2.857**

Regresyon Atama			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	12	0	0
II	0	12	0
III	0	1	10
Hatalı Sınıf. Yüzdesi=			**2.857**

Grup Ortalamalarının Atanması			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	11	1	0
II	0	12	0
III	0	0	11
Hatalı Sınıf. Yüzdesi=			**2.857**

Ortalama Atama			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	1	58	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0.699**
Test	**I**	**II**	**III**
I	12	0	0
II	0	12	0
III	0	0	11
Hatalı Sınıf. Yüzdesi=			**0**

Liste Bazında Silme			
Eğitim	**I**	**II**	**III**
I	23	2	1
II	2	33	0
III	1	0	14
Hatalı Sınıf. Yüzdesi=			**7.8945**
Test	**I**	**II**	**III**
I	10	2	0
II	0	11	1
III	1	1	9
Hatalı Sınıf. Yüzdesi=			**14.286**

Çizelge 5.7 incelendiğinde test veri setinin sınıflandırılmasında hatalı sınıflandırma yüzdesi bakımından en başarılı yöntem tüm gözlemlerin doğru olarak sınıflandırılmasını sağlayan ortalama atama yöntemi olmuştur. Test veri setinin sınıflandırılmasında en kötü performans gösteren yöntem ise %14.286'lık hatalı sınıflandırma yüzdesi ile liste bazında silme yöntemi olmuştur

Şarap veri setinde eksik gözlem oranı %10 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri Çizelge 5.8'de verilmiştir.

Çizelge 5.8. Şarap veri setinde eksik gözlem oranı %10 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri

Tam Veri Durumu			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	11	1	0
II	0	12	0
III	0	1	10
Hatalı Sınıf. Yüzdesi=			**5.714**

EM Atama			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	12	0	0
II	0	12	0
III	0	0	11
Hatalı Sınıf. Yüzdesi=			**0**

Regresyon Atama			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	12	0	0
II	0	12	0
III	0	0	11
Hatalı Sınıf. Yüzdesi=			**0**

Grup Ortalamalarının Atanması			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	12	0	0
II	0	12	0
III	0	1	10
Hatalı Sınıf. Yüzdesi=			**2.857**

Ortalama Atama			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	10	2	0
II	0	12	0
III	0	0	11
Hatalı Sınıf. Yüzdesi=			**5.714**

Liste Bazında Silme			
Eğitim	**I**	**II**	**III**
I			
II			
III			
Hatalı Sınıf. Yüzdesi=			
Test	**I**	**II**	**III**
I			
II			
III			
Hatalı Sınıf. Yüzdesi=			

Çizelge 5.8 incelendiğinde test veri setinin sınıflandırılmasında hatalı sınıflandırma yüzdesi bakımından en başarılı yöntem tüm gözlemlerin doğru olarak sınıflandırılmasını sağlayan EM atama yöntemi ve regresyon atama yöntemi olmuştur. Liste bazında silme yöntemi sonucu elde edilen veri setine karma ayrıştırma analizi gözlem sayısı değişken sayısından az olduğundan dolayı uygulanamamıştır. Liste bazında silme yönteminin en büyük dezavantajı olan örneklem hacminin yetersiz duruma düşmesi problemi bu veri setinde ortaya konmuştur.

Şarap veri setinde eksik gözlem oranı %15 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri Çizelge 5.9'da verilmiştir.

Çizelge 5.9. Şarap veri setinde eksik gözlem oranı %15 olduğunda, eksik veri durumunda kullanılan yöntemlere göre karma ayrıştırma analizinde elde edilen hatalı sınıflandırma yüzdeleri

Tam Veri Durumu			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	11	1	0
II	0	12	0
III	0	1	10
Hatalı Sınıf. Yüzdesi=			**5.714**

EM Atama			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	11	1	0
II	0	12	0
III	0	0	11
Hatalı Sınıf. Yüzdesi=			**2.857**

Regresyon Atama			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	12	0	0
II	0	12	0
III	1	1	9
Hatalı Sınıf. Yüzdesi=			**5.714**

Grup Ortalamalarının Atanması			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	12	0	0
II	0	12	0
III	0	1	10
Hatalı Sınıf. Yüzdesi=			**2.857**

Ortalama Atama			
Eğitim	**I**	**II**	**III**
I	47	0	0
II	0	59	0
III	0	0	37
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I	11	1	0
II	0	12	0
III	0	0	11
Hatalı Sınıf. Yüzdesi=			**2.857**

Liste Bazında Silme			
Eğitim	**I**	**II**	**III**
I			
II			
III			
Hatalı Sınıf. Yüzdesi=			**0**
Test	**I**	**II**	**III**
I			
II			
III			
Hatalı Sınıf. Yüzdesi=			

Çizelge 5.9 incelendiğinde test veri setinin sınıflandırılmasında hatalı sınıflandırma yüzdesi bakımından en başarılı yöntem %2.857'lik hatalı sınıflandırma yüzdesi bakımından EM atama yöntemi, grup ortalamalarının atanması ve ortalama atama yöntemi olmuştur. Bir önceki eksik gözlem oranında olduğu gibi burada da liste

bazında silme yöntemi sonucu elde edilen veri setine karma ayrıştırma analizi gözlem sayısı değişken sayısından az olduğundan dolayı uygulanamamıştır.

Eksik gözlem probleminin çözümünde kullanılan yöntemlerin şarap verinde karma ayrıştırma analizindeki etkinliğinin incelenmesinde liste bazında silme yöntemi dışındaki diğer yöntemler bir başka ifade ile atama yöntemlerinin başarılı olduğu gözlemlenmiştir. Yöntemlerin uygulanması sonrası ya tam veri durumunda elde edilen sonuçlar elde edilmiş ya da daha başarılı sınıflandırma sonuçları elde edilmiştir.

5.4. Genel Değerlendirme

Cam, iris ve şarap verilerinden eksik gözlem oranlarına göre elde edilen hatalı sınıflandırma yüzdelerinin ortalamaları Çizelge 5.10'da verilmiştir.

Çizelge 5.10. Eksik gözlem oranlarına göre hatalı sınıflandırma yüzdelerinin ortalaması

Yöntem	**Eksik Gözlem Oranı**		
	%5	**%10**	**%15**
EM Atama	5.327	5.347	5.188
Regresyon Atama	7.341	4.236	8.294
Grup Ortalamalarının Atanması	5.188	3.105	7.341
Ortalama Atama	7.5	7.321	8.452
Liste Bazında Silme	17.54	11.25	14.479

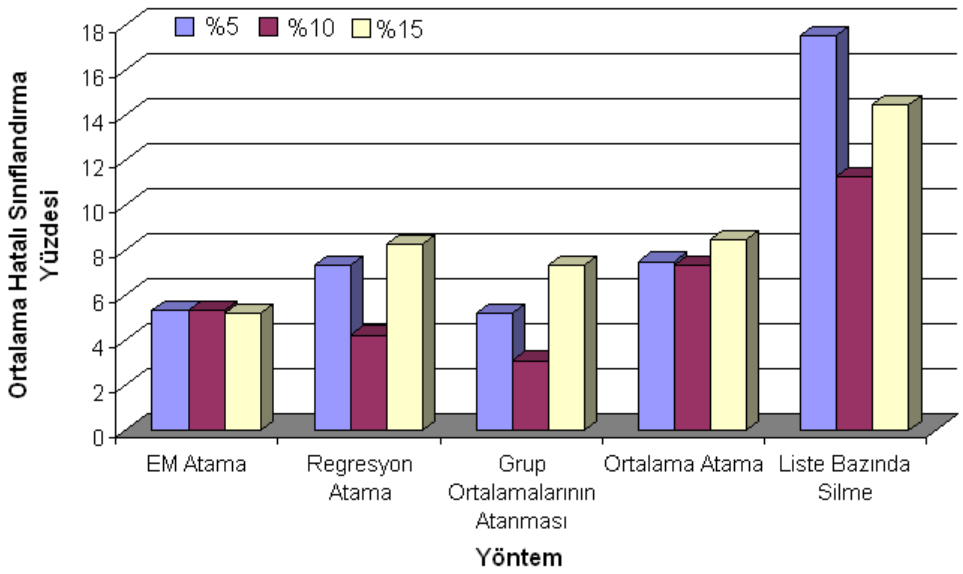
Çizelge 5.10 incelendiğinde %5 oranındaki eksik gözlem durumunda karma ayrıştırma analizinde en başarılı yöntem %5.188 ortalama hatalı sınıflandırma yüzdesi ile grup ortalamalarının atanması yöntemi olduğu görülmektedir. EM atama yöntemi %5.327 ortalama hatalı sınıflandırma yüzdesi ile karma ayrıştırma analizinde başarılı performans gösteren atama yöntemlerinden biri olmuştur.

Eksik gözlem oranı %10 olduğunda da %3.105 ortalama hatalı sınıflandırma yüzdesi ile grup ortalamalarının atanması yöntemi karma ayrıştırma analizinde en başarılı yöntem olmuştur. Eksik gözlem oranı %10 olduğunda da ikinci en başarılı yöntemde %4.236 ile regresyon atama yöntemi olmuştur.

Eksik gözlem oranı %15 olduğunda ise karma ayrıştırma analizinde %5.188'lik ortalama hatalı sınıflandırma yüzdesi ile en başarılı yöntem olarak tespit edilmiştir. Eksik gözlem oranı %15 olduğunda karma ayrıştırma analizinde en başarılı ikinci yöntem %7.341'lik ortalama hatalı sınıflandırma yüzdesi ile grup ortalamalarının atanması yöntemi olmuştur.

Ortalama hatalı sınıflandırma yüzdeleri incelendiğinde tüm eksik gözlem oranlarında karma ayrıştırma analizinde en başarısız yöntem liste bazında silme olmuştur. Yüksek hatalı sınıflandırma yüzdesi yanında liste bazında silme durumunda örneklem hacminin yetersiz olması probleminden dolayı şarap verisinde %10 ve %15'lik eksik gözlem oranında karma ayrıştırma analizi uygulanamamıştır.

Çizelge 5.10'da verilen eksik gözlem oranlarına göre elde edilen ortalama hatalı sınıflandırma yüzdelerinin grafiksel karşılaştırılması Şekil 5.4'de verilmiştir.



Şekil 5.4. Eksik gözlem oranlarına göre eksik gözlem probleminin çözümünde kullanılan yöntemlerin karma ayrıştırma analizindeki etkinliğinin ortalama hatalı sınıflandırma yüzdesi bakımından grafiksel karşılaştırılması

Şekil 5.4 incelendiğinde EM atama yönteminin karma ayrıştırma analizinde hatalı sınıflandırma yüzdesi bakımından eksik gözlem oranından diğer yöntemlere göre daha az etkilendiği açıkça görülmektedir.

6. SONUÇLAR VE ÖNERİLER

6.1. Sonuçlar

Veri setinin doğal yapısından ya da gözlem kaynaklı nedenlerden dolayı çoğu araştırmada tüm birimlerin tüm değişkenler için değerleri her zaman gözlemlenememektedir. Veri setinde gözlem değerlerinin gözlenememesi durumu literatürde eksik gözlem, kayıp gözlem veya tamamlanamamış veri problemi olarak adlandırılmaktadır. Gerek istatistiksel analizler gerek analizlerin uygulanmasını sağlayan istatistik paket programları tam veri varsayımı altında geliştirildiklerinden veri setinde gözlemlenemeyen değerler önemli problemler oluşturabilmektedir. Araştırmacılar eksik veri probleminin çözümünde günümüze kadar birçok yöntem kullanmışlardır. Bu yöntemler incelendiğinde yöntemlerin en genel anlamda silme yöntemleri ve atama yöntemleri olarak iki alt başlıkta ele alınabileceği görülür. Silme yöntemleri, eksik gözlemin bulunduğu değişken ya da birimlerin veri setinden dışlandığı liste bazında silme ve eksik gözlemlerin göz ardı edilerek gözlemlenen tüm değerler kullanılarak istatistiksel hesaplamaların yapıldığı çiftler bazında silme yönteminden oluşmaktadır. Atama yöntemleri eksik verileri tahmin ederek veri setini tamamlanmış veri durumuna getirmeyi amaçlayan yöntemlerdir. Bu yöntemler sabit bir değerin atanması yöntemi, ortalama değerin atanması yöntemi, en benzer birime benzetme yöntemi, regresyon atama yöntemi, EM atama yöntemi ve çoklu atama yöntemidir.

Eksik veri probleminin çözümünde öncelikle veride eksik değerleri ortaya çıkaran bir yapı olup olmadığının araştırmacı tarafından incelenmesi gerekir. Bu incelemede eksik verinin rassallığı incelenerek eksik veri yapısı tanımlanır ve uygun çözüm yöntemi belirlenir. Eksik veri probleminde eksik veri yapısı tamamıyla rassal olarak kayıp (TROK), rassal olarak kayıp (ROK) ve rassal olarak kayıp değil (ROKD) şeklinde üçe ayrılmıştır. Eksik veri yapısı TROK ise, tam gözlemlerin kullanımını amaçlayan liste bazında silme veya birim bazında silme yöntemleri iyi bir seçim olacaktır. Eğer eksik veri mekanizması ROK olarak tanımlanırsa EM, regresyon ve çoklu atama yöntemleri eksik veri probleminin çözümünde uygun atama yöntemleri olarak tanımlanmıştır. Tüm atama yöntemleri eksik veri yapısı ROKD olduğunda yanlı sonuç verirler.

Ayrıştırma analizi, birimlerin doğal gruplarının bilindiği çok değişkenli veri setlerinde birimlerin sınıflandırılmasını sağlayacak ayırma fonksiyonlarının oluşturulması ve elde edilen ayırma fonksiyonları yardımı ile sonradan gözlemlenen ve grubu belli olmayan birimlerin sınıflandırılmasını amaçlar. Karma ayrıştırma analizi, çok değişkenli normal dağılım varsayımı ve eşit varyans varsayımının sağlanamadığı durumlarda kullanılan ayrıştırma analizi yöntemlerinden biridir. Karma ayrıştırma analizinde genellikle çok değişkenli normal dağılımların karmasına dayalı karma ayrıştırma analizi daha yaygın kullanılmaktadır.

Karma dağılım modellerine dayalı ayrıştırma analizinin uygulanabilmesi için veri setinin tamamlanmış olması gerekmektedir. Bu çalışmada eksik veri yöntemlerinin karma ayrıştırma analizindeki etkinliği incelenmiştir.

Çalışmanın üçüncü bölümünde eksik veri kavramı tanımlanarak eksik veri yapıları, eksik veri sürecinin rassallığının incelenmesi ve eksik veri probleminin çözümünde kullanılan yöntemler incelenmiştir.

Çalışmanın dördüncü bölümünde karma dağılım modeline dayalı ayrıştırma analizi için gerekli tanımlama ve gösterimler yapılarak karma dağılım modelleri, parametre tahmini ve çok değişkenli normal dağılımlarının karmasının karması şeklinde tanımlanabilecek karma dağılım modeline dayalı ayrıştırma analizi anlatılmıştır.

Çalışmanın beşinci bölümünde sınıflama literatüründe yaygın bir kullanıma sahip cam (glass), iris ve şarap (wine) verileri kullanılarak eksik gözlem probleminin çözümünde kullanılan yöntemlerin karma ayrıştırma analizindeki etkinliği incelenmiştir. Karma ayrıştırma analizi uygulamasında kovaryans matrislerinin farklı olduğu ve her grubun iki alt gruba sahip olduğu varsayılmıştır. Eksik gözlem probleminin çözüm yöntemlerinin karma ayrıştırma analizindeki etkinliği hatalı sınıflandırma yüzdesi kriterine göre yapılmıştır. Yöntemlerin karma ayrıştırma analizindeki performansları %5, %10 ve %15 eksik gözlem oranlarında değerlendirilmiştir.

Cam verisinde eksik gözlem probleminin çözümünde kullanılan yöntemlerin karma ayrıştırma analizindeki etkinliği incelendiğinde EM atama yönteminin diğer yöntemlere göre daha başarılı olduğu ve tüm eksik gözlem oranlarında en başarısız yöntemin liste bazında silme yöntemi olduğu görülmüştür.

İris verisinde eksik gözlem probleminin çözümünde kullanılan yöntemlerin karma ayrıştırma analizindeki etkinliği incelendiğinde %5 ve %10'luk eksik gözlem oranında grup ortalamalarının atanması yöntemi, %15'lik eksik gözlem oranında EM

atama yöntemi en başarılı yöntemler olarak tespit edilmiştir. Tüm eksik gözlem oranlarında en başarısız yöntem ise liste bazında silme yöntemi olmuştur.

Eksik gözlem probleminin çözümünde kullanılan yöntemlerin şarap verinde karma ayrıştırma analizindeki etkinliğinin incelenmesinde liste bazında silme yöntemi dışındaki diğer yöntemler bir başka ifade ile atama yöntemlerinin başarılı olduğu gözlemlenmiştir. Yöntemlerin uygulanması sonrası ya tam veri durumunda elde edilen sonuçlar elde edilmiş ya da daha başarılı sınıflandırma sonuçları elde edilmiştir.

Cam, iris ve şarap verilerinden eksik gözlem oranlarına göre elde edilen hatalı sınıflandırma yüzdelerinin ortalamaları incelendiğinde %5 oranındaki eksik gözlem durumunda karma ayrıştırma analizinde en başarılı yöntem %5.188 ortalama hatalı sınıflandırma yüzdesi ile grup ortalamalarının atanması yöntemi olduğu görülmektedir. EM atama yöntemi %5.327 ortalama hatalı sınıflandırma yüzdesi ile karma ayrıştırma analizinde başarılı performans gösteren atama yöntemlerinden biri olmuştur.

Eksik gözlem oranı %10 olduğunda %3.105 ortalama hatalı sınıflandırma yüzdesi ile grup ortalamalarının atanması yöntemi karma ayrıştırma analizinde en başarılı yöntem olmuştur. Eksik gözlem oranı %10 olduğunda da ikinci en başarılı yöntemde %4.236 ile regresyon atama yöntemi olmuştur.

Eksik gözlem oranı %15 olduğunda ise karma ayrıştırma analizinde %5.188'lik ortalama hatalı sınıflandırma yüzdesi ile en başarılı yöntem olarak tespit edilmiştir. Eksik gözlem oranı %15 olduğunda karma ayrıştırma analizinde en başarılı ikinci yöntem %7.341'lik ortalama hatalı sınıflandırma yüzdesi ile grup ortalamalarının atanması yöntemi olmuştur.

Çalışmada ayrıca EM atama yönteminin karma ayrıştırma analizinde hatalı sınıflandırma yüzdesi bakımından eksik gözlem oranından diğer yöntemlere göre daha az etkilendiği tespit edilmiştir.

6.2. Öneriler

Çalışmadan elde edilen sonuçlar incelendiğinde karma ayrıştırma analizinde eksik veri probleminin çözümünde atama yöntemlerinin kullanılmasının daha uygun olduğu görülmüştür. Özellikle hesaplama basitliği de göz önüne alındığında eksik gözlemler yerine grup ortalamalarının atanması yönteminin kullanılması araştırmacılara daha faydalı olacaktır. İncelenen atama yöntemleri içersinde EM atama yöntemi de hem düşük hatalı sınıflandırma yüzdesi hem de eksik gözlem oranından en az etkilenen

yöntem olması dolayı diğer atama yöntemlerine göre karma ayrıştırma analizinde eksik veri probleminin çözümünde daha çok tercih edilmelidir.

Karma ayrıştırma analizi uygulamalarında eksik veri problemi ile karşılaşıldığında silme yöntemleri tercih edilmemelidir. Silme yöntemleri sonucunda örneklem hacminin düşmesi sonucu ya yüksek hatalı sınıflandırma yüzdesi ile karşılaşılmakta ya da birim sayısının değişken sayısından az olması problemi ile karşılaşılaşılmaktadır.

7. KAYNAKLAR

Acock, A. C., 2005, Working with missing values, *Journal of Marriage and Family*, 67, 1012-1028.

Acuna Edgar, Rodriguez Caroline, 2004, The treatment of missing values and its effect in the classifier accuracy.

Aik Lim Eng., Zarita Zainuddin, 2008, A Comparative Study of Missing Value Estimation Methods: Which Method Performs Better?, 2008 International Conference on Electronic Design. December 1-3,2008, Penang, Malaysia.

Alpar, R., 2011, Çok Değişkenli İstatistiksel Yöntemler, Detay Yayıncılık, Ankara.

Ashikaga, T., and Chang, P. C., 1981, Robustness Of Fisher's Linear Discriminant Function Under Two-Component Mixed Normal Models, *Journal Of The American Statistical Association*, 76, 676-680.

Bal Cengiz, 2003, Çok Gruplu Veri Setlerinde Eksik Gözlem Sorununun Çözümlenmesi ve Sağlık Alanında Bir Uygulama. Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilimdalı Doktora Tezi, Eskişehir.

Bashir, S. And Carter, E. M., 2005, Robust Reduced Rank Mixture Discriminant Analysis, *Communications in Statistics—Theory and Methods,* 34, 135–145.

Baum, L.E., Betrie, T., Solues, G., And Weiss, N., 1970, A maximization techniqueoccuring in the Statistical Analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics,* 41, 164-171.

Baygül Arzu, 2007, Kayıp Veri Analizinde Sıklıkla Kullanılan Etkin Yöntemlerin Değerlendirilmesi, İstanbul Üniversitesi Cerrahpaşa Tıp Fakültesi Biyoistatistik Bilim Dalı Yüksek Lisans Tezi, İstanbul.

Beale, E.M.L., and Little R.J.A., 1975, Missing Values in Multivariate Analysis, *Journal of the Royal Statistical Society. Series B (Methodological),* 37, 1, 129-145.

Belen Rahime, 2009, Detecting Disguised Missing Data. Ortadoğu Teknik Üniversitesi Bilişim Sistemleri Yüksek Lisans Tezi, Ankara.

Biernacki, C., and Govaert, G., 1999, Choosing Models İn Model-Based Clustering And Discriminant Analysis, Journal *Of Statistical Computation And Simulation,* 64, 49-71.

Buck S. F., 1996, A Method of Estimation of Missing Values in Multivariate Data suitable for use with an Electronic Computer, *Journal of the Royal Statistical Society. Series B (Methodological)*, 22 (2),. 302-306.

Craig K. Enders, 2010, Applied Missing Data Analysis, The Guilford Press, New York, United States of America.

Celeux, G. And Govaert, G., 1993, Comparison of the Mixture and the Classification Maximum likelihood in Cluster Analysis, *Journal of Statistical Computation and Simulation,* 47, 127-146.

Çalış, N., 2011, Karma Dağılım Modellerine Dayalı Ayrıştırma Analizi ve Sınıflandırma, Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı Doktora Tezi, Adana.

Dempster, A.P., Laırd, N.M., and Rubın, D.M., 1977, Maximum likelihood form incomplete data via the EM algorithm, with discussion, *Journal of the Royal Statistical Society B*, 39, 1-38.

Enders C.K. and Bandalos D.L. 2001, The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models, *Structural Equation Modeling: A Multidisciplinary Journal*, 8:3, 430-457.

Erişoğlu, Ü., 2011, Yaşam Süresi Analizinde Karma ve Karıştırılmış Dağılımların Kullanılması, Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı Doktora Tezi, Adana.

Erol, H., and Akdeniz, F. 1996, A multispectral classification algorithm for Classifying Parcels In An Agricultural Region, *International Journal Of Remote Sensing,* 17, 3357 3371.

Everıt, B. S., and Hand, D. J., 1981, Finite Mixture Distributions. London: Chapman&Hall.

Fraley, C. and Raftery, A.E., 2002, Model-Based Clustering, Discriminant Analysis, And Density Estimation, *Journal Of The American Statistical Association*, 97, 611-631.

Halbe, Z. and Aladjem, M., 2005, Model-Based Mixture Discriminant Analysis– An Experimental Study, *Pattern Recognition*, 38, Pp. 437–440.

Hastie, T. and Tibshirani, R., 1996, Discriminant Analysis by Gaussian Mixtures, *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, (1), 155-176.

Jackson E.C., 1968, Missing Values in Linear Multiple Discriminant Analysis. *Biometrics,* 24, (4) 835-844

Ju, J., Kolaczyk, E. D. and Gopal, S., 2003, Gaussian Mixture Discriminant Analysis And Sub-Pixel Land Cover Characterization in Remote Sensing, *Remote Sensing And Environment,* 84, 550-560.

Krzanowski, W. J., 1982, Mixtures of continuous and categorical variables in discriminant analysis: A hypothesis-testing approach, *Biometrics,* 38, 991-1002.

Li, J., 2005, Clustering based on a multi-layer mixture model, *Journal of Computational and Graphical Statistics,* 14, (3), 547–568.

Little, R. J. C., Rubin, D. B., 1987, Statistical Analysis with Missing Data, John Wiley&Sons, New York, United States of America.

Little, R. J. A., 1988, A test of missing completely at random for multivariate data with missing values, *Journal of the American Statistical Association*, 83, 1198–1202.

Murphy, P. M., And Aha, D. W., 1995, UCI Repository of Machine Learning Databases. Irvine, CA: University of California, Dept. of Information and Computer Science.

Olga Troyanskaya, Michael Cantor , Gavin Sherlock, Pat Brown, Trevor Hastie,, Robert Tibshirani, David Botstein and Russ B. Altman, 2001, Missing value estimation methods for DNA microarrays, *Bioinformatics*, 17, (6), 520-525.

Qiao Mu. and Li Jia, 2010, Two-way Gaussian Mixture Models for High Dimensional Classification, *Statistical Analysis and Data Mining*, 3, 259–271.

Orchard, T. And Woodbury, M.A., 1972, A missing information principle: theory and applications. In: Proceedings of the 6th Berkeley Symposium in Mathematical Statistics and Probability, Vol I, 697-715.

Pearson, K., 1894, Contributions to the theory of mathematical evolution. *Philosophical Transaction of the Royal Society of London A*, 185, 71-110.

Raftery, A.E., Dean, N., 2006, Variable Selection for Model-based Clustering, *Journal of the American Statistical Association,* 101, 168–178.

Roeder, K., 1990, Density estimation with confidence sets exemplified by superclusters and voids in the galaxies, *JASA*, 85 617-624.

Rubin, D. B., 1976, Inference and Missing Data, *Biometrika,* 63, 3, 581-592.

Satıcı Esra, 2009, Kayıp Gözlem Olması Durumunda Kitle Ortalaması Tahmini, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı Doktora Tezi, Ankara.

Schafer, J. L., 1997, Analysis of Incomplete Multivariate Data, Chapman & Hall/CRC, United States of America.

Servi, T., 2009, Çok Değişkenli Karma Dağılım Modeline Dayalı Kümeleme Analizi, Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı Doktora Tezi, Adana.

Singh, G.N., Priyanka, K., 2007, Effect of non-response on current occasion in search of good rotation patterns on successive occasions, *Statistics in Transition*, 8 (2), 273-292.

Singh, S., Horn, S., 2000, Compromised imputation in survey sampling, *Metrika,* 51, 267-276.

Sundberg, R., 1974, Maximum likelihood theory for incomplete data from an exponential family, *Scandinavian Journal of Statistics: Theory and Applications,* 1, 49-58.

Weldon, W. F. R., 1982, Certain corralated variations in Crangon vulgaris, *Proceeding of the Royal Society of London,* 51, 2-21.

Weldon, W. F. R., 1983, On certain corralated variations in Carcinus moenas. *Proceeding of the Royal Society of London,* 54, 318-329.

Wilks, S. S., 1932, Moments and distributions of estimates of population parameters from fragmentary samples, *The Annals of Mathematical Statistics*, 3, 163–195.

EKLER

Ek 1. Cam veri seti

Gözlem No	x1	x2	x3	x4	x5	x6	x7	Grup
1	1.521	13.64	4.49	1.1	71.78	0.06	8.75	1
2	1.5176	13.89	3.6	1.36	72.73	0.48	7.83	1
3	1.5162	13.53	3.55	1.54	72.99	0.39	7.78	1
4	1.5177	13.21	3.69	1.29	72.61	0.57	8.22	1
5	1.5174	13.27	3.62	1.24	73.08	0.55	8.07	1
6	1.516	12.79	3.61	1.62	72.97	0.64	8.07	1
7	1.5174	13.3	3.6	1.14	73.09	0.58	8.17	1
8	1.5176	13.15	3.61	1.05	73.24	0.57	8.24	1
9	1.5192	14.04	3.58	1.37	72.08	0.56	8.3	1
10	1.5176	13	3.6	1.36	72.99	0.57	8.4	1
11	1.5157	12.72	3.46	1.56	73.2	0.67	8.09	1
12	1.5176	12.8	3.66	1.27	73.01	0.6	8.56	1
13	1.5159	12.88	3.43	1.4	73.28	0.69	8.05	1
14	1.5175	12.86	3.56	1.27	73.21	0.54	8.38	1
15	1.5176	12.61	3.59	1.31	73.29	0.58	8.5	1
16	1.5176	12.81	3.54	1.23	73.24	0.58	8.39	1
17	1.5178	12.68	3.67	1.16	73.11	0.61	8.7	1
18	1.522	14.36	3.85	0.89	71.36	0.15	9.15	1
19	1.5191	13.9	3.73	1.18	72.12	0.06	8.89	1
20	1.5174	13.02	3.54	1.69	72.73	0.54	8.44	1
21	1.5175	12.82	3.55	1.49	72.75	0.54	8.52	1
22	1.5197	14.77	3.75	0.29	72.02	0.03	9	1
23	1.5221	14.21	3.82	0.47	71.77	0.11	9.57	1
24	1.5179	12.79	3.5	1.12	73.03	0.64	8.77	1
25	1.5176	12.71	3.42	1.2	73.2	0.59	8.64	1
26	1.5178	13.21	3.39	1.33	72.76	0.59	8.59	1
27	1.5221	13.73	3.84	0.72	71.76	0.17	9.74	1
28	1.5179	12.73	3.43	1.19	72.95	0.62	8.76	1
29	1.519	13.49	3.48	1.35	71.95	0.55	9	1
30	1.5187	13.19	3.37	1.18	72.72	0.57	8.83	1
31	1.5267	13.99	3.7	0.71	71.57	0.02	9.82	1
32	1.5222	13.21	3.77	0.79	71.99	0.13	10.02	1
33	1.519	13.58	3.35	1.23	72.08	0.59	8.91	1
34	1.5232	13.72	3.72	0.51	71.75	0.09	10.06	1
35	1.5193	13.2	3.33	1.28	72.36	0.6	9.14	1
36	1.5181	13.43	2.87	1.19	72.84	0.55	9.03	1
37	1.5184	13.14	2.84	1.28	72.85	0.55	9.07	1
38	1.5178	13.21	2.81	1.29	72.98	0.51	9.02	1
39	1.5177	12.45	2.71	1.29	73.7	0.56	9.06	1
40	1.5122	12.99	3.47	1.12	72.98	0.62	8.35	1
41	1.5182	12.87	3.48	1.29	72.95	0.6	8.43	1
42	1.5175	13.48	3.74	1.17	72.99	0.59	8.03	1
43	1.5175	13.39	3.66	1.19	72.79	0.57	8.27	1
44	1.5191	13.6	3.62	1.11	72.64	0.14	8.76	1
45	1.5198	13.81	3.58	1.32	71.72	0.12	8.67	1
46	1.5217	13.51	3.86	0.88	71.79	0.23	9.54	1
47	1.5223	14.17	3.81	0.78	71.35	0	9.69	1
48	1.5217	13.48	3.74	0.9	72.01	0.18	9.61	1

Cam veri seti (devam)

Gözlem No	x1	x2	x3	x4	x5	x6	x7	Grup
49	1.521	13.69	3.59	1.12	71.96	0.09	9.4	1
50	1.5215	13.05	3.65	0.87	72.22	0.19	9.85	1
51	1.5215	13.05	3.65	0.87	72.32	0.19	9.85	1
52	1.5215	13.12	3.58	0.9	72.2	0.23	9.82	1
53	1.523	13.31	3.58	0.82	71.99	0.12	10.17	1
54	1.5177	13.65	3.66	1.11	72.77	0.11	8.6	1
55	1.5161	13.33	3.53	1.34	72.67	0.56	8.33	1
56	1.5167	13.24	3.57	1.38	72.7	0.56	8.44	1
57	1.5164	12.16	3.52	1.35	72.89	0.57	8.53	1
58	1.5167	13.14	3.45	1.76	72.48	0.6	8.38	1
59	1.5213	14.32	3.9	0.83	71.5	0	9.49	1
60	1.5178	13.64	3.65	0.65	73	0.06	8.93	1
61	1.5161	13.42	3.4	1.22	72.69	0.59	8.32	1
62	1.5169	12.86	3.58	1.31	72.61	0.61	8.79	1
63	1.5165	13.04	3.4	1.26	73.01	0.52	8.58	1
64	1.5166	13.41	3.39	1.28	72.64	0.52	8.65	1
65	1.5212	14.03	3.76	0.58	71.79	0.11	9.65	1
66	1.5178	13.53	3.41	1.52	72.04	0.58	8.79	1
67	1.518	13.5	3.36	1.63	71.94	0.57	8.81	1
68	1.5183	13.33	3.34	1.54	72.14	0.56	8.99	1
69	1.5193	13.64	3.54	0.75	72.65	0.16	8.89	1
70	1.5221	14.19	3.78	0.91	71.36	0.23	9.14	1
71	1.5174	12.78	3.62	1.29	72.79	0.59	8.7	1
72	1.5175	12.81	3.57	1.35	73.02	0.62	8.59	1
73	1.5172	13.38	3.5	1.15	72.85	0.5	8.43	1
74	1.5176	12.98	3.54	1.21	73	0.65	8.53	1
75	1.5179	13.21	3.48	1.41	72.64	0.59	8.43	1
76	1.5172	12.87	3.48	1.33	73.04	0.56	8.43	1
77	1.5177	12.56	3.52	1.43	73.15	0.57	8.54	1
78	1.5178	13.08	3.49	1.28	72.86	0.6	8.49	1
79	1.5177	12.65	3.56	1.3	73.08	0.61	8.69	1
80	1.5175	12.84	3.5	1.14	73.27	0.56	8.55	1
81	1.5178	12.85	3.48	1.23	72.97	0.61	8.56	1
82	1.5175	12.57	3.47	1.38	73.39	0.6	8.55	1
83	1.5178	12.69	3.54	1.34	72.95	0.57	8.75	1
84	1.5157	13.29	3.45	1.21	72.74	0.56	8.57	1
85	1.5191	13.89	3.53	1.32	71.81	0.51	8.78	1
86	1.518	12.74	3.48	1.35	72.96	0.64	8.68	1
87	1.5221	14.21	3.82	0.47	71.77	0.11	9.57	1
88	1.5157	14.86	3.67	1.74	71.87	0.16	7.36	2
89	1.5185	13.64	3.87	1.27	71.96	0.54	8.32	2
90	1.5159	13.09	3.59	1.52	73.1	0.67	7.83	2
91	1.5163	13.34	3.57	1.57	72.87	0.61	7.89	2
92	1.516	13.02	3.56	1.54	73.11	0.72	7.9	2
93	1.5159	13.02	3.58	1.51	73.12	0.69	7.96	2
94	1.5165	13.44	3.61	1.54	72.39	0.66	8.03	2
95	1.5163	13	3.58	1.54	72.83	0.61	8.04	2
96	1.5161	13.92	3.52	1.25	72.88	0.37	7.94	2

Cam veri seti (devam)

Gözlem No	x1	x2	x3	x4	x5	x6	x7	Grup
97	1.5159	12.82	3.52	1.9	72.86	0.69	7.97	2
98	1.5159	12.86	3.52	2.12	72.66	0.69	7.97	2
99	1.5159	13.25	3.45	1.43	73.17	0.61	7.86	2
100	1.5165	13.41	3.55	1.25	72.81	0.68	8.1	2
101	1.5159	13.09	3.52	1.55	72.87	0.68	8.05	2
102	1.5141	14.25	3.09	2.08	72.28	1.1	7.08	2
103	1.5163	13.36	3.58	1.49	72.72	0.45	8.21	2
104	1.5157	13.24	3.49	1.47	73.25	0.38	8.03	2
105	1.5165	13.4	3.49	1.52	72.65	0.67	8.08	2
106	1.5162	13.01	3.5	1.48	72.89	0.6	8.12	2
107	1.5164	12.55	3.48	1.87	73.23	0.63	8.08	2
108	1.5184	12.93	3.74	1.11	72.28	0.64	8.96	2
109	1.516	12.9	3.44	1.45	73.06	0.44	8.27	2
110	1.5159	13.12	3.41	1.58	73.26	0.07	8.39	2
111	1.5159	13.24	3.34	1.47	73.1	0.39	8.22	2
112	1.5163	12.71	3.33	1.49	73.28	0.67	8.24	2
113	1.5186	13.36	3.43	1.43	72.26	0.51	8.6	2
114	1.5184	13.02	3.62	1.06	72.34	0.64	9.13	2
115	1.5174	12.2	3.25	1.16	73.55	0.62	8.9	2
116	1.5169	12.67	2.88	1.71	73.21	0.73	8.54	2
117	1.5181	12.96	2.96	1.43	72.92	0.6	8.79	2
118	1.5166	12.75	2.85	1.44	73.27	0.57	8.79	2
119	1.5173	12.35	2.72	1.63	72.87	0.7	9.23	2
120	1.5182	12.62	2.76	0.83	73.81	0.35	9.42	2
121	1.5273	13.8	3.15	0.66	70.57	0.08	11.64	2
122	1.5241	13.83	2.9	1.17	71.15	0.08	10.79	2
123	1.5248	11.45	0	1.88	72.19	0.81	13.24	2
124	1.5313	10.73	0	2.1	69.81	0.58	13.3	2
125	1.5339	12.3	0	1	70.16	0.12	16.19	2
126	1.5171	13.48	3.48	1.71	72.52	0.62	7.99	2
127	1.5218	13.2	3.68	1.15	72.75	0.54	8.52	2
128	1.5187	12.93	3.66	1.56	72.51	0.58	8.55	2
129	1.5167	12.94	3.61	1.26	72.75	0.56	8.6	2
130	1.5208	13.78	2.28	1.43	71.99	0.49	9.85	2
131	1.5207	13.55	2.09	1.67	72.18	0.53	9.57	2
132	1.5202	13.98	1.35	1.63	71.76	0.39	10.56	2
133	1.5218	13.75	1.01	1.36	72.19	0.33	11.14	2
134	1.5261	13.7	0	1.36	71.24	0.19	13.44	2
135	1.5181	13.43	3.98	1.18	72.49	0.58	8.15	2
136	1.518	13.71	3.93	1.54	71.81	0.54	8.21	2
137	1.5181	13.33	3.85	1.25	72.78	0.52	8.12	2
138	1.5179	13.19	3.9	1.3	72.33	0.55	8.44	2
139	1.5181	13	3.8	1.08	73.07	0.56	8.38	2
140	1.5171	12.89	3.62	1.57	72.96	0.61	8.11	2
141	1.5167	12.79	3.52	1.54	73.36	0.66	7.9	2
142	1.5167	12.87	3.56	1.64	73.14	0.65	7.99	2
143	1.5169	13.33	3.54	1.61	72.54	0.68	8.11	2
144	1.5185	13.2	3.63	1.07	72.83	0.57	8.41	2

Cam veri seti (devam)

Gözlem No	x1	x2	x3	x4	x5	x6	x7	Grup
145	1.5166	12.85	3.51	1.44	73.01	0.68	8.23	2
146	1.5171	13	3.47	1.79	72.72	0.66	8.18	2
147	1.5166	12.99	3.18	1.23	72.97	0.58	8.81	2
148	1.5184	12.85	3.67	1.24	72.57	0.62	8.68	2
149	1.5222	14.43	0	1	72.67	0.1	11.52	2
150	1.5182	13.72	0	0.56	74.45	0	10.99	2
151	1.5266	11.23	0	0.77	73.21	0	14.68	2
152	1.5274	11.02	0	0.75	73.08	0	14.96	2
153	1.5278	12.64	0	0.67	72.02	0.06	14.4	2
154	1.5189	13.46	3.83	1.26	72.55	0.57	8.21	2
155	1.5185	13.1	3.97	1.19	72.44	0.6	8.43	2
156	1.5185	13.41	3.89	1.33	72.38	0.51	8.28	2
157	1.5183	13.24	3.9	1.41	72.33	0.55	8.31	2
158	1.5171	13.72	3.68	1.81	72.06	0.64	7.88	2
159	1.5167	13.3	3.64	1.53	72.53	0.65	8.03	2
160	1.5165	13.56	3.57	1.47	72.45	0.64	7.96	2
161	1.5184	13.25	3.76	1.32	72.4	0.58	8.42	2
162	1.5166	12.93	3.54	1.62	72.96	0.64	8.03	2
163	1.5169	13.23	3.54	1.48	72.84	0.56	8.1	2

Ek 2. İris veri seti

Gözlem No	Çanak Uzunluğu	Çanak Genişliği	Taç Uzunluğu	Taç Genişliği	Tür
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3	1.4	0.1	setosa
14	4.3	3	1.1	0.1	setosa
15	5.8	4	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa
23	4.6	3.6	1	0.2	setosa
24	5.1	3.3	1.7	0.5	setosa
25	4.8	3.4	1.9	0.2	setosa
26	5	3	1.6	0.2	setosa
27	5	3.4	1.6	0.4	setosa
28	5.2	3.5	1.5	0.2	setosa
29	5.2	3.4	1.4	0.2	setosa
30	4.7	3.2	1.6	0.2	setosa
31	4.8	3.1	1.6	0.2	setosa
32	5.4	3.4	1.5	0.4	setosa
33	5.2	4.1	1.5	0.1	setosa
34	5.5	4.2	1.4	0.2	setosa
35	4.9	3.1	1.5	0.2	setosa
36	5	3.2	1.2	0.2	setosa
37	5.5	3.5	1.3	0.2	setosa
38	4.9	3.6	1.4	0.1	setosa
39	4.4	3	1.3	0.2	setosa
40	5.1	3.4	1.5	0.2	setosa
41	5	3.5	1.3	0.3	setosa
42	4.5	2.3	1.3	0.3	setosa
43	4.4	3.2	1.3	0.2	setosa
44	5	3.5	1.6	0.6	setosa
45	5.1	3.8	1.9	0.4	setosa
46	4.8	3	1.4	0.3	setosa
47	5.1	3.8	1.6	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
49	5.3	3.7	1.5	0.2	setosa
50	5	3.3	1.4	0.2	setosa

İris veri seti (devam)

Gözlem No	Çanak Uzunluğu	Çanak Genişliği	Taç Uzunluğu	Taç Genişliği	Tür
51	7	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4	1.3	versicolor
55	6.5	2.8	4.6	1.5	versicolor
56	5.7	2.8	4.5	1.3	versicolor
57	6.3	3.3	4.7	1.6	versicolor
58	4.9	2.4	3.3	1	versicolor
59	6.6	2.9	4.6	1.3	versicolor
60	5.2	2.7	3.9	1.4	versicolor
61	5	2	3.5	1	versicolor
62	5.9	3	4.2	1.5	versicolor
63	6	2.2	4	1	versicolor
64	6.1	2.9	4.7	1.4	versicolor
65	5.6	2.9	3.6	1.3	versicolor
66	6.7	3.1	4.4	1.4	versicolor
67	5.6	3	4.5	1.5	versicolor
68	5.8	2.7	4.1	1	versicolor
69	6.2	2.2	4.5	1.5	versicolor
70	5.6	2.5	3.9	1.1	versicolor
71	5.9	3.2	4.8	1.8	versicolor
72	6.1	2.8	4	1.3	versicolor
73	6.3	2.5	4.9	1.5	versicolor
74	6.1	2.8	4.7	1.2	versicolor
75	6.4	2.9	4.3	1.3	versicolor
76	6.6	3	4.4	1.4	versicolor
77	6.8	2.8	4.8	1.4	versicolor
78	6.7	3	5	1.7	versicolor
79	6	2.9	4.5	1.5	versicolor
80	5.7	2.6	3.5	1	versicolor
81	5.5	2.4	3.8	1.1	versicolor
82	5.5	2.4	3.7	1	versicolor
83	5.8	2.7	3.9	1.2	versicolor
84	6	2.7	5.1	1.6	versicolor
85	5.4	3	4.5	1.5	versicolor
86	6	3.4	4.5	1.6	versicolor
87	6.7	3.1	4.7	1.5	versicolor
88	6.3	2.3	4.4	1.3	versicolor
89	5.6	3	4.1	1.3	versicolor
90	5.5	2.5	4	1.3	versicolor
91	5.5	2.6	4.4	1.2	versicolor
92	6.1	3	4.6	1.4	versicolor
93	5.8	2.6	4	1.2	versicolor
94	5	2.3	3.3	1	versicolor
95	5.6	2.7	4.2	1.3	versicolor
96	5.7	3	4.2	1.2	versicolor
97	5.7	2.9	4.2	1.3	versicolor
98	6.2	2.9	4.3	1.3	versicolor
99	5.1	2.5	3	1.1	versicolor
100	5.7	2.8	4.1	1.3	versicolor

İris veri seti (devam)

Gözlem No	Çanak Uzunluğu	Çanak Genişliği	Taç Uzunluğu	Taç Genişliği	Tür
101	6.3	3.3	6	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3	5.9	2.1	virginica
104	6.3	2.9	5.6	1.8	virginica
105	6.5	3	5.8	2.2	virginica
106	7.6	3	6.6	2.1	virginica
107	4.9	2.5	4.5	1.7	virginica
108	7.3	2.9	6.3	1.8	virginica
109	6.7	2.5	5.8	1.8	virginica
110	7.2	3.6	6.1	2.5	virginica
111	6.5	3.2	5.1	2	virginica
112	6.4	2.7	5.3	1.9	virginica
113	6.8	3	5.5	2.1	virginica
114	5.7	2.5	5	2	virginica
115	5.8	2.8	5.1	2.4	virginica
116	6.4	3.2	5.3	2.3	virginica
117	6.5	3	5.5	1.8	virginica
118	7.7	3.8	6.7	2.2	virginica
119	7.7	2.6	6.9	2.3	virginica
120	6	2.2	5	1.5	virginica
121	6.9	3.2	5.7	2.3	virginica
122	5.6	2.8	4.9	2	virginica
123	7.7	2.8	6.7	2	virginica
124	6.3	2.7	4.9	1.8	virginica
125	6.7	3.3	5.7	2.1	virginica
126	7.2	3.2	6	1.8	virginica
127	6.2	2.8	4.8	1.8	virginica
128	6.1	3	4.9	1.8	virginica
129	6.4	2.8	5.6	2.1	virginica
130	7.2	3	5.8	1.6	virginica
131	7.4	2.8	6.1	1.9	virginica
132	7.9	3.8	6.4	2	virginica
133	6.4	2.8	5.6	2.2	virginica
134	6.3	2.8	5.1	1.5	virginica
135	6.1	2.6	5.6	1.4	virginica
136	7.7	3	6.1	2.3	virginica
137	6.3	3.4	5.6	2.4	virginica
138	6.4	3.1	5.5	1.8	virginica
139	6	3	4.8	1.8	virginica
140	6.9	3.1	5.4	2.1	virginica
141	6.7	3.1	5.6	2.4	virginica
142	6.9	3.1	5.1	2.3	virginica
143	5.8	2.7	5.1	1.9	virginica
144	6.8	3.2	5.9	2.3	virginica
145	6.7	3.3	5.7	2.5	virginica
146	6.7	3	5.2	2.3	virginica
147	6.3	2.5	5	1.9	virginica
148	6.5	3	5.2	2	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3	5.1	1.8	virginica

Ek 3. Şarap verisi (Standartlaştırılmış veri)

Gözlem No	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	Bölge
1	0.923	-0.543	0.158	-1.047	-0.752	0.487	0.732	-0.577	0.383	0.234	0.842	0.406	1.820	1
2	2.155	-0.543	0.086	-2.424	-0.612	1.286	1.663	0.548	2.130	0.147	1.280	0.167	1.280	1
3	1.699	-0.417	0.049	-2.244	0.158	1.606	1.613	-0.577	2.392	1.053	1.061	0.547	2.541	1
4	0.775	-0.471	1.216	-0.687	0.858	0.887	0.882	-0.497	-0.229	0.967	1.411	0.378	1.788	1
5	1.601	-0.373	1.288	0.151	1.418	0.807	1.112	-0.256	0.662	0.493	0.492	0.054	1.693	1
6	1.022	-0.686	0.924	0.151	1.068	1.046	1.372	0.307	0.226	0.665	0.755	-0.059	1.217	1
7	1.465	-0.668	0.414	-0.897	0.578	1.606	1.903	-0.336	0.470	1.571	1.192	0.293	2.963	1
8	0.788	0.684	0.705	-1.286	1.138	0.647	1.002	-1.542	0.121	0.018	0.011	1.054	0.312	1
9	1.305	-0.632	-0.315	-1.047	1.839	1.126	1.142	-0.979	0.889	0.255	0.580	1.547	0.105	1
10	-0.087	1.310	1.033	-0.268	0.158	0.184	0.381	-0.899	0.680	-0.241	0.317	1.279	0.073	1
11	0.874	-0.426	-0.024	-0.867	0.088	0.503	0.852	-0.738	0.173	-0.543	0.667	1.955	0.915	1
12	-0.186	-0.659	0.559	-0.508	-0.332	0.295	0.341	-0.818	-0.229	-0.487	0.580	1.434	0.851	1
13	0.615	-0.471	0.888	0.151	-0.262	0.375	0.581	-0.658	0.121	-0.663	0.711	1.702	0.312	1
14	0.061	-0.256	3.111	1.648	1.698	0.535	0.651	0.869	0.575	-0.638	0.755	0.829	0.264	1
15	0.480	-0.507	0.924	-1.017	-0.472	0.887	0.912	-0.176	-0.246	-0.111	-0.164	0.857	1.423	1
16	0.369	-0.552	-0.826	-0.747	-0.402	0.168	0.161	-0.738	-0.421	-0.478	0.274	0.223	1.709	1
17	1.071	-0.391	1.580	-0.028	0.508	1.046	0.942	0.065	0.295	-0.241	1.280	1.110	0.534	1
18	1.256	-0.588	-0.571	-1.047	-0.262	0.567	0.301	-0.818	0.680	-0.154	0.361	1.378	0.915	1
19	0.898	-0.749	1.216	0.900	0.088	1.126	1.222	-0.577	1.379	0.277	1.017	0.138	1.709	1
20	0.714	-0.605	-0.024	-0.118	0.438	0.903	1.162	-1.140	0.627	0.795	0.580	0.378	2.439	1
21	0.837	-0.453	-0.024	-0.687	0.298	0.200	0.661	0.467	0.662	-0.525	1.192	0.364	0.772	1
22	0.935	-0.722	1.216	0.002	2.259	1.046	0.712	1.110	-0.421	0.147	1.280	0.547	1.550	1
23	0.627	-0.480	1.033	-0.148	0.718	0.088	0.501	-0.577	-0.089	-0.370	0.624	0.364	1.105	1
24	0.590	-0.471	0.158	0.301	0.018	0.647	0.952	-0.818	0.470	0.018	0.361	1.209	0.550	1
25	0.344	-0.623	1.726	-1.196	0.718	0.487	0.651	-0.176	-0.403	-0.198	0.580	0.237	0.423	1
26	0.061	-0.614	0.669	-0.448	-0.122	0.247	0.401	-0.577	-0.264	-0.349	0.711	-0.143	1.137	1
27	0.085	-0.749	-0.971	-1.196	-0.122	0.168	0.611	-0.658	-0.386	-0.586	0.974	0.110	0.867	1
28	1.502	1.480	0.523	-1.885	1.979	1.126	1.012	-1.301	0.855	0.018	-0.295	1.293	0.042	1
29	0.689	-0.561	-0.206	-0.987	1.208	1.366	1.262	-0.176	1.309	0.462	-0.033	1.082	0.153	1
30	0.504	1.346	-0.899	-0.208	-0.682	0.247	0.651	-0.738	-0.194	-0.336	-0.208	0.547	0.915	1

Şarap verisi (devam)

Gözlem No	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	Bölge
31	1.083	-0.400	0.815	-1.346	0.088	1.526	1.532	-1.542	0.191	0.160	-0.339	1.336	1.105	1
32	0.295	1.471	-0.279	-0.597	0.228	0.551	0.601	-0.336	0.121	-0.301	-0.601	0.547	-0.212	1
33	0.061	-0.507	-0.971	-0.747	0.508	1.126	0.972	-0.658	0.767	-0.008	-0.339	1.040	0.439	1
34	1.490	1.525	0.268	-0.178	0.788	0.887	0.621	-0.497	-0.596	0.078	-0.383	1.012	1.058	1
35	1.699	1.122	-0.315	-1.047	0.158	1.526	1.142	-0.738	1.047	-0.068	0.361	1.167	1.010	1
36	1.108	-0.588	-0.899	-1.047	0.088	1.286	1.362	-1.220	0.959	0.449	-0.208	1.012	0.756	1
37	1.354	-0.283	0.122	-0.208	0.228	0.727	0.892	-0.336	1.379	0.493	0.492	0.195	0.994	1
38	1.157	-0.543	-0.352	-0.627	0.578	0.935	1.512	-0.336	0.855	1.657	0.711	0.688	1.629	1
39	0.061	-0.543	-1.190	-2.125	-0.542	0.679	1.242	-1.542	2.305	0.924	0.711	0.420	1.280	1
40	1.022	-0.614	0.851	-0.687	-0.402	0.247	0.962	-1.140	1.221	0.234	1.236	1.068	1.645	1
41	1.009	-0.525	0.195	-1.645	0.788	2.532	1.713	-0.336	0.488	0.859	0.230	0.913	1.407	1
42	0.948	-0.391	1.143	-0.717	1.068	1.126	0.762	0.226	0.156	0.536	0.755	0.448	1.995	1
43	0.911	-0.596	-0.425	-0.927	1.278	0.487	0.872	-1.220	0.051	0.342	-0.164	0.829	0.994	1
44	0.689	-0.543	0.341	0.301	1.138	1.062	0.752	-1.301	1.501	0.514	0.099	0.589	1.185	1
45	1.502	-0.570	-0.242	-0.957	1.278	1.446	0.972	-0.818	0.767	0.570	-0.076	0.984	0.708	1
46	0.356	-0.328	1.143	-0.807	0.158	1.126	1.202	-0.417	0.121	0.406	0.492	0.322	1.661	1
47	0.886	-0.811	0.487	-0.837	0.578	1.765	1.643	-1.381	0.785	0.751	-0.295	0.364	1.709	1
48	1.514	-0.561	0.231	-1.166	1.909	0.807	1.032	-0.658	1.221	0.251	0.361	1.843	1.010	1
49	0.246	-0.498	-0.826	-2.484	0.018	0.567	0.732	-0.818	-0.543	-0.292	0.405	1.110	0.963	1
50	0.196	0.021	1.106	-0.268	0.088	0.807	1.212	-0.497	2.130	0.268	0.317	0.786	1.391	1
51	1.687	-0.346	0.487	-0.807	0.928	2.484	1.462	-0.979	1.029	1.183	-0.426	1.181	2.328	1
52	0.295	0.227	1.835	0.451	1.278	0.807	0.661	0.226	0.400	-0.318	0.361	0.448	-0.038	1
53	1.477	-0.516	0.304	-1.286	0.858	1.558	1.362	-0.176	0.662	0.730	0.405	0.336	2.233	1
54	1.711	-0.417	0.304	-1.466	-0.262	0.327	0.491	-0.497	0.680	0.083	0.274	1.364	1.725	1
55	1.305	-0.167	0.888	-0.567	1.488	0.487	0.481	-0.417	-0.596	-0.003	0.449	1.364	1.741	1
56	2.253	-0.623	-0.716	-1.645	-0.192	0.807	0.952	-0.577	0.680	0.061	0.536	0.336	0.947	1
57	1.059	-0.883	-0.352	-1.047	-0.122	1.094	1.122	-1.140	0.453	0.933	0.230	1.322	0.947	1
58	1.354	-0.158	-0.242	-0.448	0.368	1.046	1.292	-1.140	1.379	0.298	1.280	0.786	2.423	1
59	1.379	-0.767	-0.170	-0.807	-0.332	-0.152	0.401	-0.818	-0.037	-0.025	0.930	0.293	1.693	1
60	-0.777	-1.250	-3.669	-2.664	-0.822	-0.503	-1.461	-0.658	-2.046	-1.341	0.405	-1.115	-0.721	2

Şarap verisi (devam)

Gözlem No	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	Bölge
61	-0.826	-1.107	-0.315	-1.047	0.088	-0.392	-0.940	2.155	-2.063	-0.771	1.280	-1.326	-0.212	2
62	-0.444	-0.874	-1.263	-0.807	0.018	-0.440	-0.620	1.351	-1.696	0.298	0.099	-1.439	-0.943	2
63	0.825	-0.972	-1.628	-0.448	-0.402	-0.312	-0.240	-0.336	-1.504	-0.543	1.192	-0.214	-0.371	2
64	-0.777	-1.080	-0.753	-0.148	-0.892	1.925	1.072	-1.381	0.488	-0.262	1.149	0.364	-1.038	2
65	-1.023	-0.793	0.596	-0.148	0.298	-0.647	-0.280	0.708	-0.980	-0.909	2.155	-0.538	-1.244	2
66	-0.777	-1.008	0.705	-0.418	-0.122	0.200	0.621	0.065	0.855	-0.198	1.017	-0.439	-0.219	2
67	0.135	-1.187	-2.429	-1.346	-1.522	1.094	1.152	-0.818	1.204	0.104	0.711	0.800	-0.778	2
68	-0.777	-1.044	-1.628	0.031	-1.522	-0.296	-0.029	-0.738	-0.963	-0.163	0.711	1.223	-0.752	2
69	0.418	-1.250	-0.024	-0.747	0.718	0.375	-0.730	1.512	-2.046	-0.814	0.274	-0.960	0.010	2
70	-0.974	-1.026	-2.247	-0.807	3.589	-0.711	-0.750	-1.783	1.588	-0.952	1.411	0.646	-0.092	2
71	-0.875	-0.650	-0.571	0.271	0.228	-1.910	-1.010	0.065	-0.229	-0.866	-0.225	-1.115	0.391	2
72	1.059	-0.740	1.106	1.648	-0.962	1.046	0.832	-1.220	0.488	-0.724	1.761	0.772	-1.070	2
73	0.603	-0.605	-0.461	1.349	-0.892	-0.663	-0.189	-0.738	-0.980	-0.569	0.099	0.237	-0.873	2
74	-0.013	-0.596	0.851	3.146	2.749	1.606	0.862	-1.220	0.645	-0.737	1.542	1.251	0.756	2
75	-1.282	-1.116	-0.242	0.451	0.088	1.733	0.111	-1.863	0.103	-0.797	0.142	0.730	0.442	2
76	-1.651	-0.408	-1.628	-1.047	-0.192	-1.095	-0.460	-0.176	-0.770	-0.543	1.192	-0.664	-1.013	2
77	0.036	-1.286	-2.393	-1.047	-0.962	-0.551	0.001	-0.979	-0.229	-0.198	1.017	-0.185	-1.127	2
78	-1.430	0.496	-0.498	-0.448	0.858	-0.919	-0.710	0.548	-1.120	-1.039	0.011	-0.129	-0.784	2
79	-0.826	-1.205	-1.518	-1.406	2.539	-0.631	-0.179	-0.095	2.043	-0.715	0.449	-0.425	0.010	2
80	-0.370	1.373	0.122	1.050	0.088	0.855	0.521	0.548	0.627	-1.073	1.017	0.730	-0.902	2
81	-1.233	-1.268	-1.336	-0.148	-0.962	0.200	0.231	-0.497	-0.281	-1.103	1.849	0.716	-1.489	2
82	-0.346	-0.471	-0.607	-0.208	-0.962	-0.152	0.501	-0.818	0.313	-0.500	0.886	0.744	-0.104	2
83	-1.134	-1.080	0.523	1.349	-1.522	-0.472	-0.450	0.307	-0.334	-1.233	1.542	0.153	-0.371	2
84	0.061	1.364	-0.170	0.900	-1.032	-1.031	-0.440	1.994	0.051	-0.111	-0.514	-0.847	-0.736	2
85	-1.430	-1.295	0.778	-0.448	-0.402	-0.152	0.181	-1.140	1.326	-0.866	-0.733	0.660	-0.721	2
86	-0.407	-1.214	-0.461	-0.448	-0.052	-0.152	-0.089	-0.497	-0.229	-1.052	1.192	0.772	-0.943	2
87	-1.035	-0.650	-0.206	0.990	-0.682	-0.823	-0.340	0.548	-0.054	-1.125	1.630	-0.495	-0.800	2
88	-1.664	-0.596	0.924	1.948	-0.822	-0.599	-0.420	0.307	-0.438	-1.060	1.761	0.843	-0.587	2
89	-1.676	-0.247	0.341	0.630	-1.102	-0.551	-0.340	0.949	-0.421	-0.974	0.186	0.195	-0.212	2
90	-1.134	-0.901	-0.242	1.229	-2.082	-0.152	-0.440	0.467	-0.368	-1.431	0.492	0.843	-0.387	2

Şarap verisi (devam)

Gözlem No	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	Bölge
91	-1.134	-0.453	-0.170	-0.298	-1.312	-1.111	-0.530	1.271	0.086	-1.147	0.536	-0.481	-0.848	2
92	-1.233	-0.740	0.195	0.750	-0.962	-1.350	-0.780	1.110	0.068	-0.629	0.405	0.054	-0.943	2
93	-0.383	-0.722	-0.388	0.361	-1.382	-1.462	-0.570	1.753	0.051	-0.866	0.011	-0.777	-0.800	2
94	-0.875	0.442	-0.534	-0.448	-0.822	0.247	0.221	-0.899	0.697	-1.254	0.842	0.969	-1.451	2
95	-1.701	-0.310	-0.315	-0.448	-0.122	1.158	0.231	-1.542	-0.421	-0.780	0.886	0.491	-1.276	2
96	-0.654	-0.731	-0.607	-0.148	4.359	0.327	0.241	-0.336	2.951	-1.060	0.886	0.026	0.604	2
97	-1.467	-0.194	1.361	0.600	2.399	-1.111	-1.040	-1.783	-0.054	-1.103	-0.033	-0.495	-0.387	2
98	-0.875	-0.829	-1.409	-1.047	-1.032	0.407	0.471	-0.577	0.313	-0.931	1.192	0.181	-1.013	2
99	-1.898	1.256	-1.992	0.002	0.508	1.414	0.551	-0.979	3.475	-0.931	-0.908	0.279	-0.587	2
100	-0.592	0.084	-0.716	0.451	-0.822	0.407	0.241	-0.818	-0.648	-1.319	-0.251	0.237	-1.340	2
101	-1.528	0.308	2.017	0.151	0.228	-0.871	0.001	1.914	-0.945	-0.543	1.192	-0.157	-0.444	2
102	-1.959	-1.429	0.487	0.451	-0.822	0.295	-0.019	0.467	-0.264	-0.853	0.624	-0.425	-0.994	2
103	-1.134	-0.847	0.487	0.900	-1.102	0.423	0.261	0.548	-0.963	-0.931	-0.120	0.815	-1.149	2
104	-2.427	-0.740	-0.607	0.600	-1.032	0.263	0.141	1.271	0.732	-1.362	3.292	0.364	-1.079	2
105	-1.454	-0.776	-1.372	0.391	-0.962	-0.503	-0.430	-0.497	-0.106	-1.341	-0.033	1.012	-0.800	2
106	-0.715	-0.650	-0.643	0.900	0.578	-0.472	0.061	-0.176	0.033	-1.293	0.449	0.491	-1.276	2
107	-0.284	0.979	-1.409	-1.047	-1.382	-1.063	-0.780	0.548	-1.329	-0.715	-1.126	-0.693	-1.190	2
108	-1.233	0.979	-1.336	-0.148	-0.892	-0.472	-0.390	0.065	0.488	-1.630	-0.120	0.617	-0.581	2
109	-1.910	0.057	0.195	0.151	-0.262	0.967	0.762	-0.336	0.418	-0.780	-0.689	1.096	-0.387	2
110	-1.775	-0.256	3.147	2.696	1.348	1.414	3.054	0.869	0.488	0.406	-0.120	1.519	-0.895	2
111	-0.715	1.874	1.325	2.098	0.158	-0.152	0.101	0.548	0.208	-1.285	-0.164	0.716	-1.213	2
112	0.061	3.100	-0.862	0.600	-0.962	0.519	0.621	-0.497	0.732	-1.060	-0.995	0.688	-1.165	2
113	-1.393	1.767	0.086	0.451	-1.242	0.903	1.002	-1.220	2.305	-0.974	-0.908	1.448	-1.165	2
114	-1.146	-0.158	-0.716	0.451	-1.032	0.487	0.621	0.065	-0.421	-0.991	-0.426	0.941	-1.171	2
115	-0.703	-0.722	-0.279	0.600	-0.962	0.711	1.122	0.226	0.313	-0.482	-1.170	0.322	-1.254	2
116	-1.491	-0.185	1.507	2.696	-0.542	-0.264	0.211	1.753	0.295	-0.888	0.055	-0.242	-0.892	2
117	-0.777	-0.632	-0.242	1.499	-0.822	-0.120	0.421	0.307	0.540	-1.267	-0.295	0.237	-1.286	2
118	-1.183	1.758	0.049	0.750	-1.382	-0.312	-0.280	0.467	-0.421	-1.060	-0.733	-0.059	-0.530	2
119	-0.777	-1.134	-0.971	-0.298	-0.822	1.957	1.723	-0.979	0.627	-0.241	0.361	0.223	-0.276	2
120	-0.875	0.746	-0.571	-0.448	-0.822	0.887	0.962	0.708	2.130	-1.190	2.024	0.307	-1.083	2

Şarap verisi (devam)

Gözlem No	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	Bölge
121	-1.134	-0.229	-2.429	-0.597	-0.192	-0.104	0.141	-0.818	-0.334	-0.758	1.367	0.491	-0.117	2
122	-0.493	-0.892	-1.700	-0.298	-0.822	-1.350	-0.670	-0.577	-0.421	-1.125	0.361	0.223	-0.587	2
123	-0.814	0.102	0.341	0.451	-0.122	0.423	0.081	-0.176	-0.491	-0.974	-0.689	1.082	-0.981	2
124	-1.454	-0.552	-1.773	0.002	-0.962	0.327	-0.390	0.065	-0.299	-1.293	-0.076	-0.242	-1.054	2
125	-0.604	-0.543	-1.409	0.301	-1.032	-0.152	-0.109	-0.336	-0.194	-0.914	0.361	1.350	-0.238	2
126	-0.715	0.191	-0.352	0.750	-0.682	-0.983	-0.189	2.396	-0.299	-1.017	-0.426	0.969	-1.371	2
127	-0.925	-0.543	-0.899	-0.148	-1.382	-1.031	0.001	0.065	0.068	-0.715	0.186	0.786	-0.752	2
128	-0.346	-0.525	-0.315	0.900	-1.102	-1.462	-0.270	0.949	0.068	-0.758	-0.339	-0.270	-0.822	2
129	-0.962	-0.937	-1.555	-0.148	-0.542	0.104	0.011	0.226	0.855	-1.017	-0.426	0.575	-1.381	2
130	-1.713	-0.883	1.216	0.151	-0.402	0.711	0.892	-0.577	1.571	-1.039	0.011	0.913	-0.212	2
131	-0.173	-0.883	-0.170	-0.448	1.558	-1.254	-0.780	-1.220	-1.137	-0.413	-0.864	-1.862	-0.371	3
132	-0.149	0.585	0.122	0.151	0.298	-1.590	-0.810	-0.979	-1.329	0.147	-0.951	-1.678	-0.689	3
133	-0.235	-0.024	0.122	1.349	-0.122	-1.830	-0.940	-0.738	-1.329	0.277	-1.301	-1.763	-0.593	3
134	-0.370	1.086	-0.024	0.600	0.438	-0.951	-0.830	-1.542	-1.312	-0.025	-0.776	-1.862	-0.466	3
135	-0.604	-0.981	-0.425	-0.597	-1.032	-0.472	-1.451	1.914	-0.596	0.169	-0.908	-1.552	-0.308	3
136	-0.493	0.111	-0.607	-0.298	-0.402	-1.079	-1.371	2.155	-1.137	0.881	-0.995	-1.453	-0.165	3
137	-0.925	2.134	0.632	0.451	-0.752	-1.462	-1.561	1.351	-1.382	-0.521	-0.908	-1.890	-0.085	3
138	-0.580	2.841	0.997	1.648	-0.262	-0.807	-1.431	2.155	-0.858	-0.025	-0.601	-1.298	-0.736	3
139	0.603	1.122	-0.643	0.002	-0.822	-1.079	-1.551	1.753	-1.242	0.277	-0.645	-1.115	-0.530	3
140	-0.198	0.558	0.888	1.349	0.088	0.040	-1.431	1.351	-1.364	-0.060	-0.295	-0.650	-0.498	3
141	-0.087	0.424	1.216	0.451	-0.262	-1.207	-1.531	1.351	-1.469	-0.198	-0.820	-0.425	-0.466	3
142	0.443	0.200	-0.060	0.151	-0.752	-1.430	-1.531	0.065	-1.661	0.234	-1.126	-0.200	0.105	3
143	0.640	0.746	1.288	1.199	-0.192	-1.191	-1.511	1.110	-1.819	-0.305	-0.295	-0.777	-0.721	3
144	0.763	2.340	-0.060	0.151	-0.542	-0.472	-1.231	0.869	-0.997	-0.284	-0.208	-0.791	-0.625	3
145	-0.925	1.382	-0.607	-0.298	0.858	-1.462	-1.251	-0.577	-0.788	1.360	-1.345	-0.862	0.343	3
146	0.196	1.104	-0.789	0.451	0.158	-1.270	-1.481	0.548	-0.508	-0.456	-1.564	-1.312	0.264	3
147	1.083	2.420	-0.498	0.151	-1.382	-2.101	-1.691	0.307	-1.591	-0.068	-1.651	-1.805	-1.054	3
148	-0.161	2.035	0.414	0.600	-0.962	-0.951	-1.381	0.869	-1.277	1.118	-1.826	-1.059	-0.387	3
149	0.393	0.809	0.049	0.600	-0.542	-0.583	-1.271	0.708	-0.596	1.450	-1.783	-1.397	-0.308	3
150	0.098	1.400	-0.024	0.600	0.928	-1.414	-0.640	-0.176	-0.788	1.873	-1.695	-1.805	-0.625	3

Şarap verisi (devam)

Gözlem No	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	Bölge
151	0.615	0.701	0.924	1.349	1.628	-1.430	-0.460	-1.140	-0.596	1.528	-1.608	-1.847	-0.784	3
152	-0.259	0.299	0.414	0.750	0.858	-1.302	-0.670	-0.979	-0.578	2.477	-2.089	-1.608	-0.848	3
153	0.135	-0.391	1.398	1.798	1.138	-0.152	-0.750	-0.818	-0.054	0.881	-1.520	-1.805	-1.022	3
154	0.283	0.863	-0.315	-0.298	-0.122	-0.791	-1.201	1.994	0.488	2.356	-1.739	-1.552	-0.228	3
155	-0.518	-0.937	-0.971	0.151	0.228	-1.302	-1.451	1.351	-0.334	1.096	-1.651	-1.495	-0.339	3
156	0.209	2.554	-0.170	0.750	-0.472	-0.887	-1.401	1.994	-0.071	1.226	-1.564	-1.594	-0.070	3
157	1.034	1.597	0.049	0.002	-0.752	-0.791	-1.201	0.949	-0.054	1.705	-1.695	-1.369	-0.848	3
158	-0.678	0.621	0.997	2.247	-0.192	-0.631	-1.451	2.155	-0.788	1.053	-1.258	-1.242	0.423	3
159	1.650	-0.588	1.216	1.648	-0.122	0.807	-0.720	1.351	1.938	3.426	-1.695	-0.918	-0.276	3
160	0.590	-0.596	0.997	0.900	-0.752	0.487	-0.930	1.271	1.221	2.887	-1.695	-1.171	-0.403	3
161	-0.789	1.337	0.049	0.451	-0.822	0.008	-1.111	1.110	-0.963	1.118	-1.739	-1.453	-0.721	3
162	0.849	0.827	0.632	0.151	0.508	-0.743	-1.471	1.110	-1.382	0.355	0.011	-1.115	-0.212	3
163	-0.186	0.836	0.778	0.750	0.438	-1.031	-1.431	1.914	-1.102	0.225	-0.383	-0.707	-0.562	3
164	-0.050	0.997	-0.060	-0.298	0.438	-1.446	-1.331	0.307	-1.137	0.096	-1.214	-1.214	-0.228	3
165	0.960	0.379	-0.242	0.750	-0.682	-1.510	-1.351	0.387	-0.980	1.951	-1.126	-1.312	-0.419	3
166	0.898	1.811	-0.388	0.900	-0.822	-1.622	-1.561	1.271	-0.770	0.674	-0.776	-1.214	-0.721	3
167	0.554	1.221	0.851	1.050	0.788	-0.951	-1.111	0.548	-0.229	2.425	-0.470	-1.481	-0.165	3
168	-0.222	0.925	-0.242	0.002	-0.822	-1.302	-1.371	0.307	-1.085	2.244	-1.039	-1.214	-0.197	3
169	0.714	0.218	1.179	1.499	0.368	-1.191	-1.191	0.226	-0.089	1.554	-0.951	-1.143	0.010	3
170	0.492	2.026	1.799	1.648	0.858	-0.503	-1.070	-0.738	-0.840	1.485	-1.258	-0.974	-0.371	3
171	-0.986	0.621	-0.170	-0.148	-0.262	-1.670	-1.541	0.307	-1.504	0.191	-1.301	-1.101	-0.752	3
172	-0.284	0.048	-0.315	0.002	-0.962	-1.446	-1.521	0.949	-1.661	2.089	-1.695	-1.383	-0.879	3
173	1.428	0.155	0.414	0.151	-0.612	-0.983	-1.331	0.628	-0.613	2.002	-1.476	-1.270	-0.276	3
174	0.874	2.966	0.304	0.301	-0.332	-0.983	-1.421	1.271	-0.928	1.140	-1.389	-1.228	-0.022	3
175	0.492	1.409	0.414	1.050	0.158	-0.791	-1.281	0.548	-0.316	0.967	-1.126	-1.481	0.010	3
176	0.332	1.740	-0.388	0.151	1.418	-1.127	-1.341	0.548	-0.421	2.218	-1.608	-1.481	0.280	3
177	0.209	0.227	0.013	0.151	1.418	-1.031	-1.351	1.351	-0.229	1.830	-1.564	-1.397	0.296	3
178	1.391	1.579	1.361	1.499	-0.262	-0.392	-1.271	1.592	-0.421	1.787	-1.520	-1.425	-0.593	3

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	**:**	İsmet Kürşat SARI
Uyruğu	**:**	T.C.
Doğum Yeri ve Tarihi	**:**	Ankara, 1979
Telefon	**:**	0 532 630 73 93
Faks	**:**	
e-mail	**:**	ismetkursatsari@gmail.com

EĞİTİM

Derece		**Adı, İlçe, İl**	**Bitirme Yılı**
Lise	:	Yahya Kemal Beyatlı, Yenimahalle, Ankara	1997
Üniversite	:	Selçuk Üniversitesi, Konya	2004
Yüksek Lisans	:	Selçuk Üniversitesi, Konya	2012

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
2007 - 2008	S.Ü. Fen Bilimleri Enstitüsü İstatistik Bölümü	Arş. Görevlisi
2008 -	BOTAŞ	Uzman Yardımcısı

UZMANLIK ALANI

Uygulamalı İstatistik

YABANCI DİLLER

İngilizce

BELİRTMEK İSTEĞİNİZ DİĞER ÖZELLİKLER

YAYINLAR*

Murat Erişoğlu and **İsmet Kürşat Sarı** (2012), "Comparisons of Imputation Methods for Missing Data in Mixture Discriminant Analysis", 1st International Eurasian Conference on Mathematical Sciences and Applications (IECMSA), Proceeding Book, 250, Prishtine, KOSOVA.