**T.C.**

**AYDIN ADNAN MENDERES ÜNİVERSİTESİ**

**SAĞLIK BİLİMLERİ ENSTİTÜSÜ**

**BİYOİSTATİSTİK**

**DOKTORA PROGRAMI**

# YÜKSEK BOYUTLU VERİLERDE EKSİK VERİ DEĞER ATAMA YÖNTEMLERİNİN SINIFLANDIRMA PERFORMANSINA ETKİSİNİN SİMÜLASYONLA KARŞILAŞTIRILMASI

**BUĞRA VAROL**

**DOKTORA TEZİ**

**DANIŞMAN**

**PROF. DR. İMRAN KURT ÖMÜRLÜ**

**AYDIN–2023**

# KABUL VE ONAY

T.C. Aydın Adnan Menderes Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Doktora Programı çerçevesinde Buğra VAROL tarafından hazırlanan "Yüksek Boyutlu Verilerde Eksik Veri Değer Atama Yöntemlerinin Sınıflandırma Performansına Etkisinin Simülasyonla Karşılaştırılması" başlıklı tez, aşağıdaki jüri tarafından Doktora Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 01 / 08 / 2023

| | | | |
|---|---|---|---|
| Üye (T.D.) | Prof. Dr. İmran KURT ÖMÜRLÜ | Aydın Adnan Menderes Üniversitesi | …… |
| Üye | Prof. Dr. Mevlüt TÜRE | Aydın Adnan Menderes Üniversitesi | …… |
| Üye | Prof. Dr. Gökay BOZKURT | Aydın Adnan Menderes Üniversitesi | …… |
| Üye | Prof. Dr. Ferhan Elmalı | İzmir Kâtip Çelebi Üniversitesi | …… |
| Üye | Dr. Öğr. Üyesi Büşra Emir | İzmir Kâtip Çelebi Üniversitesi | …… |

ONAY:

Bu tez Aydın Adnan Menderes Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun görülmüş ve Sağlık Bilimleri Enstitüsünün ……..……..… tarih ve …… sayılı oturumunda alınan …… nolu Yönetim Kurulu kararıyla kabul edilmiştir.

Prof. Dr. Süleyman AYPAK

Enstitü Müdürü V.

# TEŞEKKÜR

Öğrenim sürecim boyunca kıymetli zamanını ayırıp sabırla ve büyük bir ilgiyle bana faydalı olabilmek için elinden gelenden fazlasını sunan, her sorun yaşadığımda yanına çekinmeden gidebildiğim, desteğini hiçbir zaman esirgemeyen, kıymetli bilgi ve birikimlerini her zaman benimle paylaşan değerli danışman hocam Sayın Prof. Dr. İmran KURT ÖMÜRLÜ'ye;

Tez çalışmam boyunca değerli bilgilerinden yararlandığım, görüş ve önerileri ile yol gösteren değerli hocam Sayın Prof. Dr. Mevlüt TÜRE'ye;

Tez yazım aşamasında verdikleri moral ve ihtiyaç duyduğum zamanlardaki yardım ve destekleri ile tezimi bitirmeme yardımcı olan değerli arkadaşlarım Arş. Gör. Dr. Fulden CANTAŞ TÜRKİŞ ve Arş. Gör. Dr. Hakan ÖZTÜRK'e sonsuz teşekkürlerimi sunarım.

Bu tez çalışmasını, gösterdiği sabır, anlayış ve verdiği destek ile aldığım her nefeste sevgisini hissettiğim sevgili eşim Gülnur VAROL'a ve canım kızım Alya VAROL'a, beni yetiştirip bugünlere getiren, maddi-manevi desteklerini hiçbir zaman esirgemeyen ve haklarını ödeyemeyeceğim sevgili annem İsmet VAROL'a, babam Oğuz Sabri VAROL'a ve ablam Tuğçe VAROL KAYABAŞI'na ithaf ediyorum. İyi ki varsınız…

# İÇİNDEKİLER

KABUL VE ONAY ..... i
TEŞEKKÜR ..... ii
İÇİNDEKİLER ..... iii
SİMGELER VE KISALTMALAR DİZİNİ ..... vi
ŞEKİLLER DİZİNİ ..... ix
TABLOLAR DİZİNİ ..... xii
ÖZET ..... xiii
ABSTRACT ..... xv
1. GİRİŞ ..... 1
1.1. Tezin Amacı ..... 4
2. GENEL BİLGİLER ..... 5
2.1. Rastgele Eksik Veri Mekanizması ..... 5
2.2. Eksik Veri Değer Atama Yöntemleri ..... 6
2.2.1. Ortalama Değer Atama ..... 7
2.2.2. Medyan Değer Atama ..... 7
2.2.3. Rastgele Değer Atama ..... 7
2.2.4. K-en Yakın Komşu Değer Atama ..... 7
2.2.5. Rastgele Orman ile Değer Atama (I-RF) ..... 9
2.2.5.1. Rastgele Orman (RF) ..... 9
2.2.5.2. Rastgele Orman ile Değer Atama Algoritması ..... 10
2.2.6. Zincirleme Denklemlerle Çok Değişkenli Değer Atama (MICE) ..... 10
2.2.6.1. Sınıflandırma ve Regresyon Ağaçları Tabanlı Zincirleme Denklemlerle Çok Değişkenli Değer Atama (MICE-CART) ..... 13
2.2.6.1.1. Sınıflandırma ve Regresyon Ağaçları (CART) ..... 13

2.2.6.1.2. Sınıflandırma ve Regresyon Ağaçları Tabanlı Zincirleme Denklemlerle Çok Değişkenli Değer Atama Algoritması ..... 14
2.2.6.2. Yüksek Boyutlu Veriler için Geliştirilen Eksik Veri Değer Atama Yöntemleri ..... 15
2.2.6.2.1. Düzenlileştirilmiş Regresyonun Doğrudan Kullanımı (DURR) ..... 16
2.2.6.2.2. Düzenlileştirilmiş Regresyonun Dolaylı Kullanımı (IURR) ..... 17
2.3. Aşırı Öğrenme Makineleri (ELM) ..... 18
2.3.1. Düzleştirilmiş Doğrusal Birim (RELU) Aktivasyon Fonksiyonu ..... 21
2.4. Yöntemlerin Performanslarının Değerlendirmesinde Kullanılan Ölçütler ..... 21
2.4.1. Değer Atama Hata Kareler Ortalaması ..... 21
2.4.2. Dengeli Doğruluk Oranı ..... 22
2.4.3. ROC Eğrisi Altında Kalan Alan ..... 22
2.4.4. Cohen'in Kappa Katsayısı ..... 23
3. GEREÇ VE YÖNTEM ..... 24
3.1. Simülasyon Algoritmaları ..... 24
3.1.1. Rastgele Bir Değişken Setinin Doğrusal Kombinasyonundan Türetilen Eksik Verili Değişkenler için Simülasyon Algoritması ..... 24
3.1.2. Tamamen Rastgele Türetilen Veriler için Simülasyon Algoritması ..... 25
3.2. Değer Atama ve Sınıflandırma Modellerine İlişkin Parametreler ..... 26
3.3. Kullanılan Programlar ..... 27
4. BULGULAR ..... 28
4.1. Rastgele Bir Değişken Setinin Doğrusal Kombinasyonundan Türetilen Eksik Verili Değişkenler için Bulgular ..... 28
4.1.1. $-0{,}1 \leq r \leq 0{,}1$ Aralığına göre Rastgele Bir Değişken Setinin Doğrusal Kombinasyonundan Türetilen Eksik Verili Değişkenler için Bulgular ..... 28
4.1.2. $-0{,}5 \leq r \leq 0{,}5$ Aralığına göre Rastgele Bir Değişken Setinin Doğrusal Kombinasyonundan Türetilen Eksik Verili Değişkenler için Bulgular ..... 37
4.1.3. $-0{,}8 \leq r \leq 0{,}8$ Aralığına göre Rastgele Bir Değişken Setinin Doğrusal Kombinasyonundan Türetilen Eksik Verili Değişkenler için Bulgular ..... 46

4.2. Tamamen Rastgele Türetilen Veriler için Bulgular ..... 57
4.2.1. Tamamen Rastgele Türetilen Verilerde -0,1 ≤ r ≤ 0,1 Aralığı için Bulgular ..... 57
4.2.2. Tamamen Rastgele Türetilen Verilerde -0,5 ≤ r ≤ 0,5 Aralığı için Bulgular ..... 66
4.2.3. Tamamen Rastgele Türetilen Verilerde -0,8 ≤ r ≤ 0,8 Aralığı için Bulgular ..... 75
5. TARTIŞMA ..... 86
6. SONUÇ VE ÖNERİLER ..... 90
KAYNAKLAR ..... 93

# SİMGELER VE KISALTMALAR DİZİNİ

$\boldsymbol{\beta}$ : Regresyon katsayıları vektörü

$\boldsymbol{\theta}$ **:** Model parametreleri

$\boldsymbol{\gamma}$ **:** Rastgele orman ile değer atama yöntemi için durdurma kriteri

$\boldsymbol{\delta_{i,h,j}}$ : j değişkeninin i ve h birimleri arasındaki mesafeye katkısı

$\boldsymbol{\lambda}$ : Ceza terimi

$\boldsymbol{s_{i,h,j}}$ : i ve h birimlerinin j değişkenine göre benzerliğinin ölçüsü

$\mathbf{G}$ : Eksik değerler için gösterge matrisi

$\boldsymbol{\varphi}$ : $\mathbf{G}$'nin bilinmeyen parametre vektörü

$\mathbf{X}$ : Bağımsız değişkenler matrisi

$\acute{\mathbf{X}}$ : Atanan bağımsız değişkenler matrisi

$\acute{\mathbf{X}}^{(\mathbf{m})}$ : m. iterasyon sonucu atanan bağımsız değişkenler matrisi

$\mathbf{X}^{\mathrm{e}}$ : $\mathbf{X}$ matrisinin eksik değerli değişkenlerinden oluşan alt kümesi

$\mathbf{X}^{\mathrm{g}}$ : $\mathbf{X}$ matrisinin tüm değişkenlerinin eksiksiz olduğu alt kümesi

$\mathbf{X}_{\mathrm{eks}}$ : $\mathbf{X}$ matrisinin eksik birimleri

$\mathbf{X}_{\mathrm{goz}}$ : $\mathbf{X}$ matrisinin gözlenen birimleri

$\acute{\mathbf{X}}^{*}$ : $\acute{\mathbf{X}}$ matrisinden elde edilen bootstrap örnekleme matrisi

$\mathbf{X}_{-\mathrm{j}}$ : $\mathbf{X}$ matrisinin j. değişken dışındaki diğer değişkenlerinden oluşan alt kümesi

$\mathbf{X}_{\mathrm{goz_j},-\mathrm{j}}$ : $\mathbf{X}$ matrisinin j. değişken dışındaki diğer değişkenlerinden oluşan ve j. değişkenin gözlenen birimlerini içeren alt kümesi.

$\mathbf{X}_{\mathrm{eks_j},-\mathrm{j}}$ : $\mathbf{X}$ matrisinin j. değişken dışındaki diğer değişkenlerinden oluşan ve j. değişkenin eksik birimlerini içeren alt kümesi.

$\mathbf{Y}$ : İki kategorili bağımlı değişken

$\mathbf{x_i}$ : i. birimin bağımsız değişkenler vektörü

**n** : Birim sayısı

**p** : Bağımsız değişken sayısı

**argmin** : Fonksiyonu minimum yapan değerler

**AUC** : Eğri altında kalan alan

**b** : Bağımsız değişkene göre ikili bölünmede oluşan düğüm

**C** : ELM için çıktı ağırlıkları ile eğitim hatası arasındaki denge parametresi

**CART** : Sınıflandırma ve regresyon ağaçları

$\mathbf{d_{i,h}}$ : i ve h gözlemleri arasındaki Gower uzaklığı

**DURR** : Düzenlileştirilmiş regresyonun doğrudan kullanımı

**ELM** : Aşırı öğrenme makineleri

**GP** : Gerçek pozitif

**GN** : Gerçek negatif

**HKO** : Hata kareler ortalaması

$\mathbf{HKO_{DA}}$ : Değer atama hata kareler ortalaması

$\mathbf{h(x)}$ : Gizli katmanlar vektörü

**H** : Gizli katman çıkış ağırlıkları matrisi

$\mathbf{H^{\dagger}}$ : Moore-Penrose genelleştirilmiş tersi

**IURR** : Düzenlileştirilmiş regresyonun dolaylı kullanımı

**KNN** : K-en yakın komşu

**M** : Çoklu değer atama iterasyon sayısı

**I-RF** : Rastgele orman ile değer atama

**MAR** : Rastgele eksik mekanizması

**MICE** : Zincirleme denklemlerle çok değişkenli değer atama

$\mathbf{p_\lambda(\beta)}$ : Düzenlileştirme fonksiyonu

**ROC** : Alıcı işlem karakteristiği

$\mathbf{w_i}$ : Giriş ve çıkış katmanı arasındaki giriş ağırlıkları vektörü

**YP** : Yalancı pozitif

**YN** : Yalancı negatif

**Z** : RF algoritması için iterasyon sayısı

# ŞEKİLLER DİZİNİ

**Şekil 1**. **X** matrisinin bileşenleri ..... 6
**Şekil 2**. RF algoritmasının işlem aşamaları ..... 9
**Şekil 3.** Çoklu değer atama ..... 11
**Şekil 4.** ELM yapısı ..... 19
**Şekil 5.** $-0,1 \leq r \leq 0,1$ aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği ..... 29
**Şekil 6.** $-0,1 \leq r \leq 0,1$ aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranlarının orman grafiği ..... 33
**Şekil 7**. $-0,1 \leq r \leq 0,1$ aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin AUC değerlerinin orman grafiği ..... 34
**Şekil 8.** $-0,1 \leq r \leq 0,1$ aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin kappa değerlerinin orman grafiği ..... 35
**Şekil 9.** $-0,1 \leq r \leq 0,1$ aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için farklı eksik oranlarında yöntemlerin dendrogram grafikleri ..... 36
**Şekil 10.** $-0,5 \leq r \leq 0,5$ aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği ..... 38
**Şekil 11.** $-0,5 \leq r \leq 0,5$ aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranlarının orman grafiği ..... 42
**Şekil 12.** $-0,5 \leq r \leq 0,5$ aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin AUC değerlerinin orman grafiği ..... 43
**Şekil 13.** $-0,5 \leq r \leq 0,5$ aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin kappa değerlerinin orman grafiği ..... 44

**Şekil 14.** -0,5 ≤ r ≤ 0,5 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için farklı eksik oranlarında yöntemlerin dendrogram grafikleri ..... 45
**Şekil 15**. -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği ..... 47
**Şekil 16.** -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranlarının orman grafiği ..... 51
**Şekil 17**. -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin AUC değerlerinin orman grafiği ..... 52
**Şekil 18**. -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin kappa değerlerinin orman grafiği ..... 53
**Şekil 19.** -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için farklı eksik oranlarında yöntemlerin dendrogram grafikleri ..... 54
**Şekil 20**. Rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için farklı korelasyon düzeylerinde yöntemlerin dendrogram grafikleri ..... 56
**Şekil 21.** Tamamen rastgele türetilen verilerde -0,1 ≤ r ≤ 0,1 aralığı için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği ..... 58
**Şekil 22.** Tamamen rastgele türetilen verilerde -0,1 ≤ r ≤ 0,1 aralığı için yöntemlerin dengeli doğruluk oranlarının orman grafiği ..... 62
**Şekil 23.** Tamamen rastgele türetilen verilerde -0,1 ≤ r ≤ 0,1 aralığı için yöntemlerin AUC değerlerinin orman grafiği ..... 63
**Şekil 24.** Tamamen rastgele türetilen verilerde -0,1 ≤ r ≤ 0,1 aralığı için yöntemlerin kappa değerlerinin orman grafiği ..... 64
**Şekil 25.** Tamamen rastgele türetilen veri setlerinde -0,1 ≤ r ≤ 0,1 aralığı ve farklı eksik oranları için yöntemlerin dendrogram grafikleri ..... 65
**Şekil 26.** Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği ..... 67

**Şekil 27.** Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin dengeli doğruluk oranlarının orman grafiği ..... 71
**Şekil 28.** Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin AUC değerlerinin orman grafiği ..... 72
**Şekil 29.** Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin kappa değerlerinin orman grafiği ..... 73
**Şekil 30.** Tamamen rastgele türetilen veri setlerinde -0,5 ≤ r ≤ 0,5 aralığı ve farklı eksik oranları için yöntemlerin dendrogram grafikleri ..... 74
**Şekil 31**. Tamamen rastgele türetilen verilerde 0,8 ≤ r ≤ 0,8 aralığı için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği ..... 76
**Şekil 32.** Tamamen rastgele türetilen verilerde -0,8 ≤ r ≤ 0,8 aralığı için yöntemlerin dengeli doğruluk oranlarının orman grafiği ..... 80
**Şekil 33.** Tamamen rastgele türetilen verilerde -0,8 ≤ r ≤ 0,8 aralığı için yöntemlerin AUC değerlerinin orman grafiği ..... 81
**Şekil 34.** Tamamen rastgele türetilen verilerde -0,8 ≤ r ≤ 0,8 aralığı için yöntemlerin kappa değerlerinin orman grafiği ..... 82
**Şekil 35.** Tamamen rastgele türetilen veri setlerinde -0,8 ≤ r ≤ 0,8 aralığı ve farklı eksik oranları için yöntemlerin dendrogram grafikleri ..... 83
**Şekil 36.** Tamamen rastgele türetilen veri setlerinde farklı korelasyon düzeyleri için yöntemlerin dendrogram grafikleri ..... 85

# TABLOLAR DİZİNİ

**Tablo 1.** 2x2 sınıflandırma tablosu ..... 22
**Tablo 2**. -0,1 ≤ r ≤ 0,1 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerleri ..... 29
**Tablo 3.** -0,1 ≤ r ≤ 0,1 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri ..... 32
**Tablo 4.** -0,5 ≤ r ≤ 0,5 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerleri ..... 38
**Tablo 5**. -0,5 ≤ r ≤ 0,5 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri ..... 41
**Tablo 6.** -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerleri ..... 47
**Tablo 7**. -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri ..... 50
**Tablo 8.** Tamamen rastgele türetilen verilerde -0,1 ≤ r ≤ 0,1 aralığı için yöntemlerin $HKO_{DA}$ değerleri ..... 58
**Tablo 9.** Tamamen rastgele türetilen verilerde -0,1 ≤ r ≤ 0,1 aralığı için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri ..... 61
**Tablo 10**. Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin $HKO_{DA}$ değerleri ..... 67
**Tablo 11.** Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri ..... 70
**Tablo 12.** Tamamen rastgele türetilen verilerde -0,8 ≤ r ≤ 0,8 aralığı için yöntemlerin $HKO_{DA}$ değerleri ..... 76
**Tablo 13.** Tamamen rastgele türetilen verilerde -0,8 ≤ r ≤ 0,8 aralığı için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri ..... 79

# ÖZET


## YÜKSEK BOYUTLU VERİLERDE EKSİK VERİ DEĞER ATAMA YÖNTEMLERİNİN SINIFLANDIRMA PERFORMANSINA ETKİSİNİN SİMÜLASYONLA KARŞILAŞTIRILMASI

**Varol B. Aydın Adnan Menderes Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Programı, Doktora Tezi, Aydın, 2023.**

**Amaç:** Bu çalışmanın amacı, türetilmiş yüksek boyutlu verilerde farklı eksik veri değer atama yöntemlerinin eksik verileri en az hata ile tahmin etmeleri ve aşırı öğrenme makineleriyle (ELM) sınıflandırma performansına etkilerinin incelenmesidir.

**Gereç ve Yöntem:** Çalışmada farklı veri yapılarına, eksik veri oranlarına ve korelasyon düzeylerine göre n=150 gözlemden oluşan iki kategorili bağımlı değişken ve p=500 bağımsız değişkenden oluşan rastgele veriler türetilerek rastgele eksik (MAR) mekanizmalı eksik değerler oluşturuldu. Eksik veri değer atama yöntemlerinden; ortalama, medyan, rastgele, k-en yakın komşu (KNN), rastgele orman ile değer atama (I-RF), sınıflandırma ve regresyon ağaçları tabanlı zincirleme denklemlerle çok değişkenli değer atama (MICE-CART) yöntemlerinin yanı sıra yüksek boyutlu veriler için geliştirilen düzenlileştirilmiş regresyonun doğrudan kullanımı (DURR) ve düzenlileştirilmiş regresyonun dolaylı kullanımı (IURR) yöntemleri ile eksik değerler atandı. 1000 döngü ile yapılan simülasyonlar sonunda yöntemlerin, ELM ile sınıflandırma skorlarının referansa yakınlığına göre eksik değer tahmin performansları değerlendirildi.

**Bulgular:** Simülasyon bulguları incelendiğinde, uygulanan aşamalı kümeleme analizine göre, değişen eksik oranları ve korelasyon düzeyleri için birbirine yakın performans gösteren yöntemlerin aynı kümede yer aldıkları tespit edildi. Eksik verili değişkenlerin veri setindeki belirli bir değişken seti ile ilişkili olduğu algoritmada, tüm korelasyon düzeyleri için düşük eksik oranlarında I-RF, MICE-CART, DURR, IURR ve bunları takiben KNN yöntemlerinin; yüksek eksik oranlarında ise DURR ve IURR yöntemlerinin referansa yakın ve benzer performans gösterdiği belirlendi. Verilerin tamamen rastgele türetildiği ikinci simülasyon

algoritmasında ise tüm korelasyon düzeyleri ve eksik oranları için yöntemlerin performanslarının birbirine yakın olduğu görüldü.

**Sonuç:** Veriler tamamen rastgele türetildiğinde, çalışmamızda kullanılan yöntemlerin tahmin performansları değişkenler arasındaki ilişkiden ve eksik oranından etkilenmemektedir. Ancak eksik verili değişkenlerin veri setindeki belirli bir değişken seti ile ilişkili olduğu durumlarda, özellikle DURR ve IURR yöntemleri diğer yöntemlere kıyasla daha etkili olmaktadır. Bu yöntemler değişkenler arasındaki ilişkiden ve eksik veri oranındaki değişimden diğer yöntemlere göre daha az etkilenmektedir.

**Anahtar kelimeler:** Aşırı öğrenme makineleri, Eksik veri, Değer atama, Sınıflandırma, Simülasyon

# ABSTRACT


## COMPARISON THE EFFECTS OF MISSING DATA IMPUTATION METHODS ON CLASSIFICATION PERFORMANCE IN HIGH DIMENSIONAL DATA THROUGH SIMULATION

**Varol B. Aydın Adnan Menderes University, Health Sciences Institute, Biostatistics Program, Doctorate Thesis, Aydın, 2023.**

**Objective:** This study aims to examine the performance of different missing data imputation methods in accurately estimating missing data in derived high-dimensional datasets and their impact on classification performance using extreme learning machines (ELM).

**Materials and Methods:** In this study, random datasets were generated consisting of n=150 observations with binary dependent variables and p=500 independent variables, considering different data structures, missing data rates, and levels of correlation. Random missing values were created using the missing at random (MAR) mechanism. The missing data imputation methods used in the study included mean, median, random, k-nearest neighbors (KNN), missing value imputation with random forests (I-RF), multiple imputations by chained equations with classification and regression trees (MICE-CART), as well as the direct use of regularized regression (DURR) and the indirect use of regularized regression (IURR) methods developed explicitly for high-dimensional data. Missing values were imputed using these methods. After 1000 iterations of simulations, the performance of the methods in estimating missing values was evaluated based on their proximity of the classification scores obtained using ELM to the reference.

**Findings:** Upon examining the simulation results, according to the applied hierarchical clustering analysis, it was determined that the methods that perform close to each other according to the varying missing rates and correlation levels were in the same cluster. it was observed that in algorithm where variables were associated with a specific set of variables in the dataset, the I-RF, MICE-CART, DURR, IURR, followed by KNN methods exhibited better performance and close to each other and the reference at low missing rates, while the DURR and IURR methods stood out at high missing rates. In the second simulation algorithm, where

the data were completely randomly generated, the performances of all methods were found to be close to each other across different correlation levels and missing rates.

**Conclusion:** When the data are completely randomly generated, the prediction performance of the methods used in our study is not affected by the relationships between variables and the missing rates. However, in cases where missing variables are associated with a specific set of variables in the dataset, particularly the DURR and IURR methods prove more effective than the others. These methods were less affected by the relationship between the variables and the variation of the missing rates compared to other methods.

**Keywords:** Extreme Learning Machines, Missing Data, Imputation, Classification, Simulation

# 1. GİRİŞ

Bir araştırmanın veri toplama aşamasında; örneklemi oluşturan tüm birimlerden, üzerinde çalışılan değişkenler bakımından tam bilgi elde edilmesi durumunda veri seti tam veri seti olarak adlandırılır. Ancak en az bir birim için bir ya da birden fazla değişkenden veri elde edilememiş ise o veri seti eksik veri seti olarak ifade edilir (Roth, 1994; Rubin, 1988). Sağlık alanı başta olmak üzere birçok alanda karşılaşılan eksik veriler, çeşitli nedenlerle ortaya çıkabilir. Örneğin; sağlık alanındaki bir araştırmada hastanın ölüm, yan etki gibi nedenlerle takipten çıkması, ölçüm aletinin arızalanmasından kaynaklı mekanik hata nedeniyle ölçüm yapılamaması, doktorların bazı hastalar için belirli tetkikler istememesi, verileri dijital ortama kaydederken yapılan hatalar gibi pek çok nedenle eksik veriler görülebilir (Little ve Rubin, 2019; Myers, 2000; Pedersen ve diğerleri, 2017). Eksik veriler, istatistiksel analiz süreçlerini olumsuz etkilediği ve birçok analiz yönteminin teorik yapısı gereği tam veriler ile çalışmayı gerektirdiği için eksik veri problemini çözmeye yönelik yaklaşımlar önem kazanmıştır. Bu problemi çözmek için eksik değerli birimleri veri setinden çıkarmak bir çözüm olarak düşünülebilir. Ancak eksik birimleri veri setinden çıkarmak, eksik verinin miktarına göre farklı derecelerde bilgi kaybına neden olur. Ayrıca örneklem hacmindeki azalmaya bağlı olarak yapılan analizin istatistiksel gücü azalır (Fox-Wasylyshyn ve El-Masri, 2005; Horton ve Lipsitz, 2001; Roth, 1994; Service, 2009; R. Zhang ve diğerleri, 2019). Eksik verili birimlerden kısmi olarak bilgi alınabiliyorsa, eldeki bilgileri kullanarak eksik verilerin yerine değer ataması yapmak ve analizleri tam veri seti üzerinde gerçekleştirmek, eksik birimleri analiz dışı bırakmaktan daha uygun bir yaklaşımdır (Schafer ve Graham, 2002).

Teknolojik gelişmeler yüksek boyutlu veriye erişmemizi sağlarken, aynı zamanda eksik değerler içeren veri setlerinin de oluşmasına yol açmıştır (Dobbin ve Simon, 2007; Xing ve diğerleri, 2001). Klasik eksik veri değer atama yöntemlerinin birçoğu, teorik yapıları nedeniyle yüksek boyutlu verilere doğrudan uygulanamadığı için yüksek boyutlu verilerdeki eksik veri problemi daha kompleks bir araştırma konusu haline gelmiştir. Literatürde veri boyutunu dikkate almayan değer atama yöntemleri bulunmaktadır. Ortalama, medyan ve rastgele değer atama yöntemleri, eksik verileri tamamlamak için en sık kullanılan ve en eski yöntemler arasında yer almaktadır. Uygulama kolaylığı, hızı ve pratikliği nedeniyle bu yöntemler birçok araştırmacı tarafından tercih edilmektedir. Ayrıca bu yöntemler, sadece eksik verinin ait olduğu değişkendeki bilgiyi kullanarak eksik veri problemini çözdükleri için veri yapısından ve

boyutundan etkilenmemektedir. Ancak, eksik veriyi açıklayan bilgi diğer değişkenlerde mevcutsa bu bilgi ihmal edilmiş olur (Acuna ve Rodriguez, 2004; Kalton, 1986). Bu yöntemlerin yetersiz kaldığı durumlarda, diğer değişkenlerdeki bilgiyi de kullanarak eksik değerleri atayan, daha gelişmiş yöntemlere olan gereksinim artmıştır. Bu konuda yapılan çalışmalarda, özellikle yüksek boyutlu veri setleriyle çalışıldığında, eksik değerler için ağaç tabanlı ve model budama tekniklerine dayalı yöntemler ön plana çıkmıştır (Y. Deng ve diğerleri, 2016; Zhao ve Long, 2016). Daniel J Stekhoven (2011) tarafından önerilen rastgele orman ile değer atama yöntemi (missing value imputation using random forests, I-RF), en popüler ağaç tabanlı değer atama yöntemlerinden biridir ve yüksek boyutlu verilere de uygulanabilmektedir. Bunun yanı sıra, k-en yakın komşu (k-nearest neighbors, KNN) algoritmasından geliştirilen değer atama yöntemleri, yüksek boyutlu verilerdeki eksik veri problemini ele alan araştırmacıların sıklıkla tercih ettiği yöntemlerdir. Kowarik ve Templ (2016), eksik değerleri tahmin etmek için gözlemlerin birbirine olan uzaklıklarını kullanan ve KNN algoritmasını temel alan bir yöntem geliştirmiştir. Bunun yanında günümüzde klasik yöntemler ile gelişmiş yöntemleri birleştiren, eksik veriler yerine birden fazla değer atama yapılması ile çoklu değer atama prosedürüne dayalı yöntemler ön plana çıkmaktadır. Çoklu değer atama teorisine göre, eksik verilerin yerine atanacak olan değerlerin kesin olarak tahmin edilememesinden kaynaklı bir belirsizlik durumu söz konusudur ve bu belirsizlik eksik veri kavramının temelini oluşturur. Çoklu değer atama yöntemleri, bu belirsizliğin neden olduğu değişkenlik kaynağını da dikkate alan ve eksik verilerin tahmini dağılımı yardımıyla değer ataması yaparak eksik veri probleminin üstesinden gelen bir teorik yapıya sahiptir (Burgette ve Reiter, 2010; Jadhav ve diğerleri, 2019; White ve diğerleri, 2011; Yin ve diğerleri, 2016). Van Buuren ve Groothuis-Oudshoorn (2011) tarafından önerilen ve çoklu değer atama prosedürü uygulayan zincirleme denklemlerle çok değişkenli değer atama (multivariate imputation by chained equations, MICE) yöntemleri son yıllardaki araştırmalarda ön plana çıkmış ve bu yönde yapılan çalışmalar hız kazanmıştır. Klasik MICE yöntemlerinin birçoğu teorik yapıları nedeniyle yüksek boyutlu veriler için uygun değildir. Bu nedenle MICE teorisinden yeni yöntemler uyarlanmış ya da mevcut yöntemler geliştirilmiştir. Burgette ve Reiter (2010) tarafından önerilen sınıflandırma ve regresyon ağaçları tabanlı zincirleme denklemlerle çok değişkenli değer atama (multivariate imputation by chained equations with classification and regression trees, MICE-CART) ve direkt yüksek boyutlu verilerdeki eksik değer problemi için geliştirilen, eşzamanlı olarak parametre tahmini ve değişken seçimine izin veren düzenlileştirilmiş regresyona dayalı, düzenlileştirilmiş regresyonun doğrudan kullanımı (direct use of regularized regression, DURR) ve düzenlileştirilmiş regresyonun dolaylı kullanımı

(indirect use of regularized regression, IURR) yöntemleri MICE teorisinden uyarlanan yöntemler arasında yer almaktadır (Costantini ve diğerleri, 2022; Y. Deng ve diğerleri, 2016; Zahid ve diğerleri, 2021; Zhao ve Long, 2016).

İstatistiksel analiz süreçlerini etkileyen eksik veriler, sınıflandırma problemleri ile ilgilenen araştırmacılar için de büyük bir sorundur (Aleryani ve diğerleri, 2018; De Brevern ve diğerleri, 2004). Yaygın olarak kullanılan sınıflandırma yöntemlerinin birçoğu model eğitim sürecinde eğitim verilerindeki eksik değer problemini çözmek için uygun algoritmaya sahip değildir. Bu nedenle, sınıflandırıcının test verileri üzerindeki genel performansını iyileştirmek için eğitim verilerindeki eksik değerli gözlemleri uygun bir değer atama yöntemi ile tamamlamaya yönelik araştırmalar önem kazanmıştır. Literatürde, eksik veri değer atama yöntemlerinin sınıflandırma performansına etkilerinin incelendiği sınırlı sayıda çalışma bulunmaktadır. Yapılan çalışmalarda değer atama yöntemlerinin sınıflandırma performanslarına etkileri çeşitli performans ölçütleri kullanılarak değerlendirilmiştir. Choudhury ve Kosorok (2020), MICE, I-RF ve KNN değer atama yöntemlerinden uyarladıkları 4 farklı yöntemin performansını farklı eksik veri oranları için gerçek ve türetilmiş yüksek boyutlu olmayan veri setlerinde karşılaştırmıştır. Ozen ve Bal (2020) çalışmalarında, KNN değer atama yönteminin sınıflandırma performansına etkisini, farklı gözlem sayılarına ve korelasyon düzeylerine göre türettikleri yüksek boyutlu olmayan 12 farklı veri setinde incelemişlerdir. García-Laencina ve diğerleri (2009), eksik gözlemleri silme, tekli ve çoklu KNN değer atama yöntemlerinin sınıflandırma performansına etkisini hem gerçek hem de türetilmiş, yüksek boyutlu olmayan veri setlerinde incelemişlerdir. Literatürde yüksek boyutlu veriler için geliştirilen DURR ve IURR yöntemlerinin performanslarını regresyon problemi için ve çeşitli performans ölçütlerine göre değerlendiren çalışmalar yer almaktadır. Zhao ve Long (2016) çalışmalarında, en küçük mutlak küçülme ve seçim operatörü (least absolute shrinkage and selection operator, lasso) ve elastik net düzenlileştirme fonksiyonlarını kullanarak, farklı ceza terimlerine göre oluşturdukları DURR ve IURR yöntemlerinin performanslarını hem gerçek hem türetilmiş yüksek boyutlu veri setlerinde değerlendirmişlerdir. Y. Deng ve diğerleri (2016) çalışmalarında, DURR ve IURR yöntemlerinin performanslarını KNN algoritmasından geliştirilen 2 farklı yöntem ve RF tabanlı bir MICE yöntemi ile 2 farklı yüksek boyutlu türetilmiş veri setinde kıyaslamışlardır.

## 1.1. Tezin Amacı

Literatür taraması sonucunda klasik ve çoklu eksik veri değer atama yöntemlerinin yüksek boyutlu verilerde sınıflandırma performansına etkisinin simülasyonla karşılaştırıldığı bir çalışmaya rastlanmamıştır. Bu çalışmada farklı dağılım, korelasyon yapıları ve eksik veri oranlarında oluşturulan simülasyon algoritmaları ışığında;

1- Yüksek boyutlu tam veri seti (referans) ile ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR eksik değer atama yöntemleriyle tamamlanan veri setlerinin değer atama hata kareler ortalamalarının ($HKO_{DA}$) tahmin edilmesi,

2- Ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR eksik değer atama yöntemlerinin aşırı öğrenme makineleri (ELM) ile sınıflandırma performansına etkisinin dengeli doğruluk oranı, alıcı işlem karakteristiği (ROC) eğrisi altında kalan alan (AUC) ve Cohen'in kappa katsayısı kriterlerine göre değerlendirilmesi,

3- Aşamalı kümeleme analiziyle ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR eksik değer atama yöntemlerinin birbirlerine ve referansa benzerliklerinin incelenmesi

amaçlandı.

# 2. GENEL BİLGİLER

## 2.1. Rastgele Eksik Veri Mekanizması

Eksik veriler bir araştırmada bilgi yokluğunu temsil ederler. Eksik verileri tamamlamak için kullanılan yöntemler eksik verinin oluşma sürecine bağlı olarak değişmektedir. Bu nedenle bu süreci tanımlayarak eksik verilerin mekanizmasını belirlemek önemlidir. Eksik veri mekanizmaları ilk defa Rubin (1976) tarafından tanımlanmıştır. Literatürde eksik verileri tamamlamak için kullanılan çoğu yöntem eksik verilerin rastgele eksik (missing at random, MAR) mekanizmasına bağlı olarak ortaya çıkığını varsayar (McCleary, 2002; Schafer ve Olsen, 1998; Van Buuren, 2007).

n sayıda bağımsız gözlemden oluşan $x_j$ bağımsız değişkenlerini aşağıdaki gibi gösterelim:

$$x_j = \left(x_{1,j}, x_{2,j}, \dots, x_{n,j}\right)^T, \qquad j = 1,2, \dots, p \tag{1}$$

$\mathbf{X}$, eksik gözlemler içeren, $x_j$ bağımsız değişkenlerinden oluşan n gözlem ve p değişkenli ($n \times p$ boyutlu) bir matris olmak üzere $\mathbf{X}$ matrisindeki eksik değerler için gösterge matrisi $\mathbf{G}$ ile tanımlansın:

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \tag{2}$$

$\mathbf{G}$ matrisinin her bir elemanı $g_{ij}$ ile temsil edilsin $(i = 1,2, \dots, n; j = 1,2, \dots, p)$. Veri setindeki gözlenen veriler $g_{i,j} = 1$, eksik verileri $g_{i,j} = 0$ ile gösterilsin. $\mathbf{G}$'nin bilinmeyen parametre vektörü φ olmak üzere $P(\mathbf{G}|\mathbf{X}, \varphi)$ koşullu olasılığı ile eksik verinin mekanizması tanımlanabilmektedir (Rubin, 1976). Buna göre MAR mekanizması için eksik verilerin ortaya çıkması olasılığı $\mathbf{X}$ matrisinin gözlenen değerlerine bağlıdır (Enders, 2022; Little ve Rubin, 2019; Schafer, 1997):

$$P(\mathbf{G} = 0|\mathbf{X}_{goz}, \mathbf{X}_{eks}, \varphi) = P(\mathbf{G} = 0|\mathbf{X}_{goz}, \varphi) \quad (3)$$

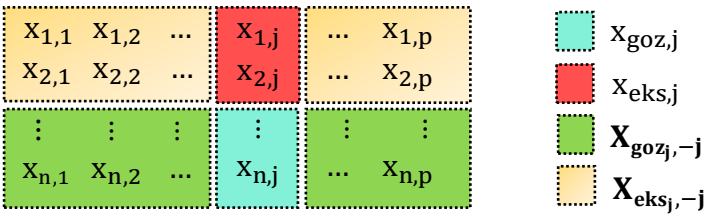
Burada $\mathbf{X}_{goz}$ ve $\mathbf{X}_{eks}$ sırasıyla $\mathbf{X}$ matrisinin gözlenen ve eksik bileşenlerini temsil etmektedir.

## 2.2. Eksik Veri Değer Atama Yöntemleri

Rastgelelik varsayımı altında $\mathbf{X}$ matrisinde ilk t değişkenin eksik değerler içerdiği varsayılsın $(t \leq p)$. $\mathbf{X}$'in en az bir gözlemi eksik değerler içeren değişkenlerden oluşan alt kümesi $n \times t$ boyutlu $\mathbf{X}^e$ matrisi; tüm değişkenlerin eksiksiz olduğu alt kümesi, $n \times (p - t)$ boyutlu $\mathbf{X}^g$ matrisi olarak tanımlanır:

$$\mathbf{X} = (\mathbf{X}^e, \mathbf{X}^g) = \big(\underbrace{x_1, \dots x_{j-1}, x_j, x_{j+1} \dots, x_t}_{\mathbf{X}^e}, \underbrace{x_{t+1}, \dots, x_p}_{\mathbf{X}^g}\big) \quad (4)$$

$\mathbf{X}^e$ değişkenler kümesindeki her $x_j$ değişkenindeki gözlenen ve eksik gözlemler sırasıyla $x_{goz,j}$ ve $x_{eks,j}$ ile; birim sayıları ise $n_{goz,j}$ ve $n_{eks,j}$ ile gösterilir. $\mathbf{X}$ matrisinin $x_j$ değişkeni dışındaki diğer $p - 1$ değişkeninden oluşan bileşimi $\mathbf{X}_{-j}$ ile; $\mathbf{X}_{-j}$ matrisinin $x_{goz,j}$ ve $x_{eks,j}$ gözlemlerine karşılık gelen bileşenleri sırasıyla $\mathbf{X}_{goz_j,-j}$ ve $\mathbf{X}_{eks_j,-j}$ ile gösterilir (Y. Deng ve diğerleri, 2016; D. J. Stekhoven ve Buhlmann, 2012). Bundan sonraki farklı matris indislerinin gösteriminde de aynı terminoloji kullanılacaktır. $\mathbf{X}$ matrisi için bu gösterimler Şekil 1'de belirtilmiştir (S. Zhang ve diğerleri, 2021).




**Şekil 1**. $\mathbf{X}$ matrisinin bileşenleri

### 2.2.1. Ortalama Değer Atama

Ortalama değer atama yönteminde, $j = 1,2, \dots, t$ için $x_j$ değişkenindeki $x_{eks,j}$ eksik verileri yerine $\bar{x}_{goz,j} = \sum_{goz=1}^{n_{goz,j}} x_{goz,j} / n_{goz,j}$ değerleri atanır (Acuna ve Rodriguez, 2004; Raymond, 1986).

### 2.2.2. Medyan Değer Atama

Medyan değer atama yönteminde, $j = 1,2, \dots, t$ için $x_j$ değişkenindeki $x_{eks,j}$ eksik verileri yerine $x_{goz,j}$ gözlemlerinin 50. persantil değeri atanır (Acuna ve Rodriguez, 2004).

### 2.2.3. Rastgele Değer Atama

Rastgele değer atama yönteminde, $j = 1,2, \dots, t$ için $x_j$ değişkenindeki $x_{eks,j}$ eksik verileri yerine $x_{goz,j}$ gözlemlerinden basit rastgele örnekleme ile seçilen rastgele gözlemler atanır (Kalton, 1986).

### 2.2.4. K-en Yakın Komşu Değer Atama

Bu yöntemin temelini KNN algoritması oluşturur. KNN algoritması Öklid, Minkowski, Gower gibi uzaklık ölçülerine dayanarak gözlemleri birbirlerine olan mesafelerine göre gruplara ayırır. Amaç, bir veri setinde yer alan gözlemlerin her birine en yakın k adet gözlemi belirlemektir. Eğitim ve test seti hataları arasındaki dengeyi sağlayacak optimum bir k sayısı belirlenerek örnekler sınıflandırılır (Weinberger ve Saul, 2009; Zhou, 2012).

**X** matrisinde en az bir değişkenin eksik değerler içerdiği gözlem vektörleri setini ve tüm değişkenlerin tam olduğu geriye kalan gözlem vektörleri setini sırasıyla $\mathbf{x}_{eks}$ ve $\mathbf{x}_{goz}$ ile; bu matrislerdeki birim sayılarını ise sırasıyla $n_{eks}$ ve $n_{goz}$ ile gösterelim. $i = 1,2, \dots, n_{eks}$ olmak üzere $\mathbf{x}_{eks}$ matrisindeki i. birimin bağımsız değişkenler vektörü $\mathbf{x_i} = (x_{i,1}, x_{i,2}, \dots, x_{i,p})$ olsun

(Acuna ve Rodriguez, 2004). Kowarik ve Templ (2016) tarafından tanımlanan KNN değer atama yöntemi gözlemler arasındaki mesafeyi farklı veri tipleri için ölçebilen Gower uzaklığına göre hesaplar. Buna göre j. değişkendeki en büyük değer ile en küçük değer arasındaki fark olarak tanımlanan dağılım aralığı $DA_j$ olmak üzere i ve h birimlerinin j değişkenine göre benzerliğinin ölçüsü $s_{i,h,j} \in [0,1]$ aşağıdaki gibi ifade edilir:

$$s_{i,h,j} = \left|x_{i,j} - x_{h,j}\right| / DA_j \tag{5}$$

$s_{i,h,j}$ 0'a yaklaştığında benzerlik artarken 1'e yaklaştığında benzerliğin azaldığı kabul edilir. i ve h gözlemleri arasındaki Gower uzaklığı ise aşağıdaki gibi hesaplanır (Gower, 1971):

$$d_{i,h} = \frac{\sum_{j=1}^{p} s_{i,h,j}\delta_{i,h,j}}{\sum_{j=1}^{p} \delta_{i,h,j}} \tag{6}$$

Burada $\delta_{i,h,j} \in \{0,1\}$, j değişkeninin i ve h birimleri arasındaki mesafeye katkısı olmak üzere h birimlerinden en az biri eksik değer içeriyorsa ya da bu birimler arasındaki mesafeyi hesaplamaya engel başka bir durum varsa $\delta_{i,h,j} = 0$, diğer durumlarda ise $\delta_{i,h,j} = 1$ olur.

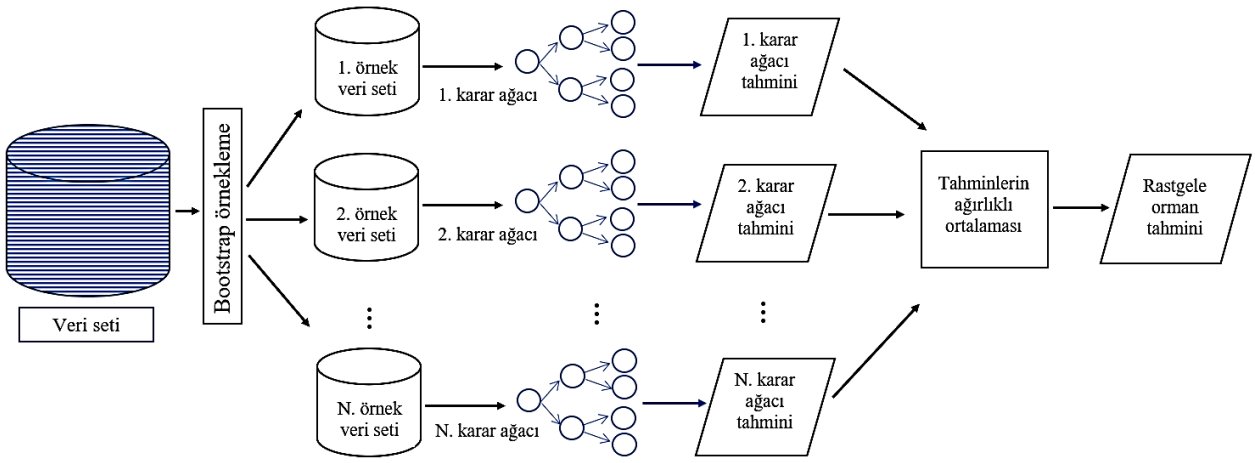
Eksik değer atama için KNN algoritmasının çalışma adımları aşağıda belirtilmiştir (Kowarik ve Templ, 2016):

1. $\mathbf{x}_i$ değişkenler vektöründeki gözlenen değişkenleri temsil eden $\mathbf{x}_{i,goz}$ vektörü ile $\mathbf{x}_{goz}$ bileşenlerinin bu değişkenlerle kesişen değerlerinden oluşan değişken vektörleri arasındaki Gower uzaklıkları hesaplanır.
2. $\mathbf{x}_i$ değişkenler vektörüne en yakın olan k adet komşu gözlem belirlenir.
3. $\mathbf{x}_i$ değişkenler vektöründeki eksik verinin olduğu sütunlara $(\mathbf{x}_{i,eks})$, k adet komşu gözlemin o sütunlardaki değerlerinin 50. persantil değeri atanır.
4. $1. - 3.$ adımlar $i = 1,2, \ldots, n_{eks}$ için tekrar edilir.
5. $\mathbf{X}$ matrisi eksik değerler yerine atanan değerler ile güncellenir ve atanmış $\hat{\mathbf{X}}$ matrisi elde edilir.

### 2.2.5. Rastgele Orman ile Değer Atama (I-RF)

#### 2.2.5.1. Rastgele Orman (RF)

RF, bir topluluk öğrenme yöntemidir. Karar ağaçları üretmek için CART algoritmasını kullanan RF, birbirinden bağımsız birden fazla karar ağacının ortak bir tahminde bulunması temeline dayanır (Breiman, 2001). RF algoritması; Breiman (1996) tarafından önerilen bagging ve Ho (1998) tarafından önerilen rastgele alt uzay tekniklerinin birleştirilmesi ile elde edilmiştir. Bagging yönteminde ağaçlar oluşturulurken gözlemler bootstrap örnekleme yöntemi ile rastgele seçilir. Rastgele alt uzay yönteminde ise öncelikle tüm değişkenler arasından basit rastgele örnekleme ile daha az sayıda değişken grubu seçilir (regresyon problemi için genellikle $p/3$). Dallara ayırıcı en iyi değişken, bu seçilen değişken grubu arasından belirlenir (Archer ve Kimes, 2008). Oluşturulması planlanan karar ağacı sayısı ve her düğüm ayrımında rasgele seçilecek olan bağımsız değişken sayısı parametreleri RF algoritmasının araştırmacı tarafından belirlenen 2 temel parametresidir. Karar ağaçları oluşturulurken veri setinin $2/3$'ü eğitim, kalan kısmı test veri seti olarak kullanılır. Eğitim verisi ile oluşturulan her ağaca test verisi için hesaplanan hata oranına göre bir ağırlık verilir. Daha sonra ağaç tahminlerinin ağırlıklı ortalaması alınarak genel bir tahmin elde edilir. Değişken ve gözlem seçimleri rastgele yapıldığı için bu yöntem aşırı uyum sorununa karşı dirençlidir (Breiman, 2001; Schonlau ve Zou, 2020). RF algoritmasının işlem aşamaları Şekil 2'de belirtilmiştir:




**Şekil 2**. RF algoritmasının işlem aşamaları (Doğaner, 2020)

### 2.2.5.2. Rastgele Orman ile Değer Atama Algoritması

Eksik değer atama için I-RF algoritmasının çalışma adımları aşağıda belirtilmiştir (D. J. Stekhoven ve Buhlmann, 2012; Tang ve Ishwaran, 2017; S. Zhang ve diğerleri, 2021):

1. **X** matrisindeki değişkenler eksik değer sayısı en az olandan en çok olana doğru sıralanır.
2. Tüm eksik değerler için ortalama değer atama ile ilk tahminler yapılır ve **X** matrisi güncellenir $(\acute{\mathbf{X}})$.
3. j. değişken bağımlı, diğer değişkenler bağımsız değişken olarak belirlenir.
4. RF modeli eğitilir $\left(x_{goz,j} \sim \acute{\mathbf{X}}_{goz_{j},-j}\right)$.
5. Eğitilen RF modeli ile $\acute{\mathbf{X}}_{eks_{j},-j}$ bağımsız değişkenleri için yeni $x_{eks,j}$ eksik değer tahminleri $\left(\acute{x}_{eks,j}\right)$ elde edilir.
6. $\acute{\mathbf{X}}$ matrisinin j. değişkeni yeni tahmin değerleri ile güncellenir $\left(\acute{\mathbf{X}} \leftarrow \acute{x}_{j}\right)$.
7. $3. - 6.$ adımlar $j = 1,2, \ldots, t$ için tekrar edilir.
8. m. iterasyon sonunda atanmış $\acute{\mathbf{X}}^{(m)}$ matrisi elde edilir $\left(\acute{\mathbf{X}}^{(m)} \leftarrow \acute{\mathbf{X}}\right)$.
9. Eşitlik 7'deki durdurma kriteri sağlanmazsa 3. adıma gidilir.
10. Z. iterasyon sonunda atanan $\acute{\mathbf{X}}^{(Z)}$ matrisi elde edilir.

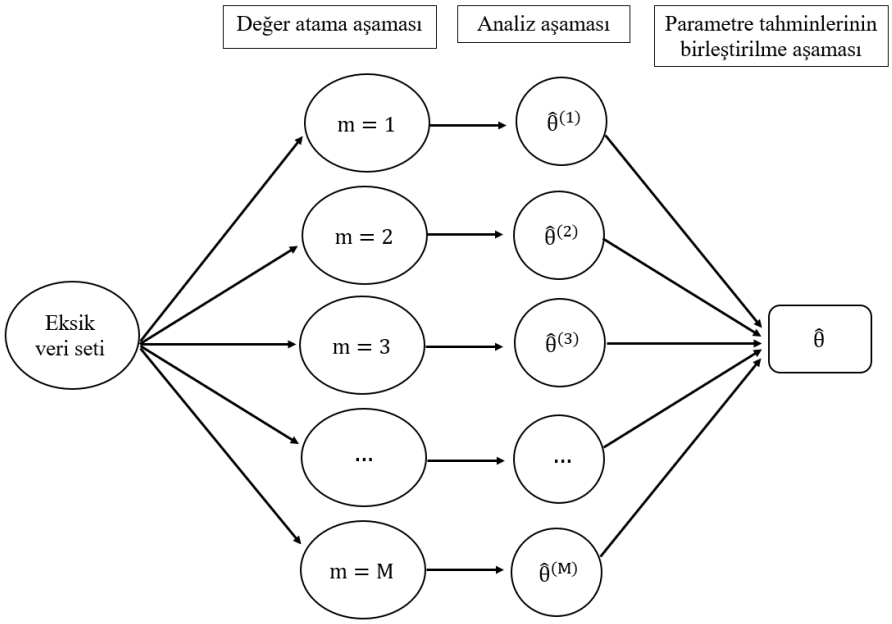
Eşitlik (7) ile gösterilen $\gamma$ değeri artmamaya başladığında durdurma kriteri sağlanmış olur. $\gamma$ aşağıdaki gibi hesaplanır:

$$\gamma = \frac{\sum_{j=1}^{t}\left(\acute{x}_{eks,j}^{(m)} - \acute{x}_{eks,j}^{(m-1)}\right)^2}{\sum_{j=1}^{t}\left(\acute{x}_{eks,j}^{(m)}\right)^2} \qquad (7)$$

### 2.2.6. Zincirleme Denklemlerle Çok Değişkenli Değer Atama (MICE)

Eksik verilerden kaynaklı bilinmezlik nedeniyle eksik veriye ait bir değişkenlik bulunmaktadır. Çoklu değer atama yaklaşımında eksik verilerin tahmini dağılımı ile bu değişkenlik tanımlanır (Rubin, 1988; Van Buuren ve Groothuis-Oudshoorn, 2011). Çoklu değer

atamaya dayalı yöntemlerde, eksik veriye $M \geq 2$ kez atama yapılır. Eksik verinin ait olduğu değişken için gözlenen veriler varlığında bir koşullu dağılım belirlenir ve bu dağılımdan rasgele seçilen değerler ile atama yapılır. Bu dağılım sürekli veriler için genelde çok değişkenli normal dağılımdır. Çoklu değer atama 3 temel aşamadan oluşur. İlk aşama değer atama aşamasıdır. Bu adımda veri setindeki eksik değerler yerine tahmin edilen değerler M kez atanır. İkinci aşamada M sayıda tamamlanmış veri seti araştırma kapsamında kullanılan yöntem ile analiz edilir. Üçüncü aşamada ise M sayıda tamamlanmış veri setinden elde edilen parametreler birleştirilir (Şekil 3) (Costantini ve diğerleri, 2022; Enders, 2022; Rubin, 1988; Yuan, 2010).




**Şekil 3.** Çoklu değer atama (Little ve Rubin, 2019)

MICE iteratif bir çoklu değer atama prosedürüdür. Her bir $x_j$ $(j = 1,2, \dots, t)$ değişkeni için M iterasyon olarak tekrarlanır (Z. Zhang, 2016). Buna göre öncelikle ortalama değer atama gibi basit bir değer atama yöntemi ile ilk eksik değer tahminleri elde edilir $\left(\acute{x}_1^{(0)}, \dots, \acute{x}_t^{(0)}\right)$. Daha sonra tamamlanmış veri seti kullanılarak iteratif bir şekilde regresyon modelleri yardımıyla yeni tahminler elde edilir. Veri setindeki eksik değer tahminleri her bir iterasyonda güncellenir (Van Buuren ve Groothuis-Oudshoorn, 2011; Zahid ve diğerleri, 2021).

**X'**in bilinmeyen bir parametre seti $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tarafından tanımlanan çok değişkenli bir dağılımdan alınan n rasgele örneğin sonucu olduğunu varsayalım $P(\mathbf{X}|\boldsymbol{\theta})$. MICE teorisinde veri setindeki eksik değişkenler için $P(x_1|\mathbf{X}_{-1}, \boldsymbol{\theta}_1), \dots, P(x_t|\mathbf{X}_{-t}, \boldsymbol{\theta}_t)$ koşullu dağılımları

tanımlanır ve bu koşullardan iteratif olarak örnekler çekerek **θ'**nın sonsal dağılımına ulaşılır. Böylece veri setine ait olduğu düşünülen çok değişkenli dağılıma yakınsanır (Y. Deng ve diğerleri, 2016; Zhao ve Long, 2016). Genel olarak 3 ile 10 arası iterasyonun yakınsama için yeterli olacağı görüşü hakimdir (Graham ve diğerleri, 2007; McCleary, 2002; Schafer ve Olsen, 1998; Zahid ve diğerleri, 2021). Burada $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_t$ parametreleri $x_1, \dots, x_t$ eksik değerli değişkenlerinin koşullu dağılımlarına özgü değer atama model parametreleridir.

$j = 1,2, \dots t$ için j. değişkenin m. iterasyondaki tahminleri ile güncellenen $\mathbf{X}$ matrisi $\acute{\mathbf{X}}_j^{(m)}$ ile gösterilmek üzere m. iterasyon için j. değişkendeki eksik değerler aşağıdaki 3 adımla tahmin edilir (Y. Deng ve diğerleri, 2016; Van Buuren, 2018):

1. j. değişkenin gözlenen birimleri kullanılarak bir doğrusal model oluşturulur:

$$x_{goz,j} = \beta_{0,j}\mathbf{1}_{n_{goz,j}} + \acute{\mathbf{X}}^{(m)}_{goz_j,-j}\boldsymbol{\beta}_j + \varepsilon_j, \qquad \varepsilon_j \sim N\left(0, \sigma_j^2 \mathbf{I}_{n_{goz,j}}\right) \tag{8}$$

   Burada $\boldsymbol{\theta}_j = \left(\beta_{0,j}, \boldsymbol{\beta_j}, \sigma_j^2\right)$ model parametreleri; $\boldsymbol{\beta_j} = \left(\beta_{1,j}, \beta_{2,j}, \dots, \beta_{p-1,j}\right)^{\mathbf{T}}$ model katsayıları; $\beta_{0,j}$ model katsayı sabiti; $\mathbf{I}_{n_{goz,j}}$, $n_{goz,j}$ elemanlı birim vektör; $\mathbf{1}_{n_{goz,j}}$ tüm değerleri 1'e eşit olan $n_{goz,j}$ elemanlı bir vektör olarak tanımlanır.

2. $\widehat{\boldsymbol{\theta}}_j^{(m)} = \left(\hat{\beta}_{0,j}^{(m)}, \widehat{\boldsymbol{\beta}}_j^{(m)}, \hat{\sigma}_j^{2(m)}\right)$ parametreleri standart bir parametre kestirim prosedürü uygulanarak tahmin edilir.
3. $x_{eks,j}$ eksik değerleri yerine aşağıdaki dağılımdan türetilen değerler atanır:

$$\acute{x}_{eks,j}^{(m)} \sim N\left(\hat{\beta}_{0,j}^{(m)}\mathbf{1}_{n_{eks,j}} + \acute{\mathbf{X}}^{(m)}_{eks_j,-j}\widehat{\boldsymbol{\beta}}_j^{(m)}, \hat{\sigma}_j^{2(m)}\mathbf{I}_{n_{eks,j}}\right) \tag{9}$$

MICE teorisine göre m. iterasyon için yukarıdaki 3 adım ile eksik değerler yerine atama yapılması aşağıdaki koşullu dağılımlardan art arda rastgele değerler çekilmesi ile gerçekleşen Gibbs örneklemesi ile eşdeğerdir:

$$\widehat{\boldsymbol{\theta}}_j^{(m)} \sim p\left(\boldsymbol{\theta_j} | x_{goz,j}, \acute{\mathbf{X}}^{(m)}_{goz_j,-j}\right) \tag{10}$$

$$\acute{x}_{eks,j}^{(m)} \sim p\left(x_{eks,j} | \acute{\mathbf{X}}^{(m)}_{eks_j,-j}, \widehat{\boldsymbol{\theta}}_j^{(m)}\right) \tag{11}$$

Bayesci yaklaşımda Formül (11), eksik gözlemlerin sonsal tahmini dağılımı olarak adlandırılır. m. iterasyon ile tamamlanmış $\acute{\mathbf{X}}_j^{(m)}$ matrisi Eşitlik (12) ile gösterilir. Burada $\acute{x}_j^{(m)}$, m. iterasyonda değer ataması yapılmış değişkendir:

$$\acute{\mathbf{X}}_j^{(m)} = \left(\acute{x}_j^{(m)}, \acute{\mathbf{X}}_{-j}^{(m)}\right) = \left(\acute{x}_1^{(m)}, \dots \acute{x}_{j-1}^{(m)}, \acute{x}_j^{(m)}, \acute{x}_{j+1}^{(m-1)} \dots, \acute{x}_t^{(m-1)}, x_{t+1}, \dots, x_p\right) \quad (12)$$

MICE prosedürü yakınsama sağlanana kadar iteratif bir şekilde devam eder. Yakınsama sağlandığında M adet tamamlanmış veri seti elde edilir. Daha sonra bu M tamamlanmış veri setinden elde edilen parametre kestirimlerinin aritmetik ortalaması alınarak sonuçlar birleştirilir (Y. Deng ve diğerleri, 2016; Zahid ve diğerleri, 2021).

#### 2.2.6.1. Sınıflandırma ve Regresyon Ağaçları Tabanlı Zincirleme Denklemlerle Çok Değişkenli Değer Atama (MICE-CART)

##### 2.2.6.1.1. Sınıflandırma ve Regresyon Ağaçları (CART)

CART, parametrik olmayan bir karar ağacı yöntemidir. Kategorik bağımlı değişkenler için sınıflandırma ağaçları; nicel bağımlı değişkenler için regresyon ağaçları üretir (Wilkinson, 2004). Her aşamada bağımsız değişkenlere göre bağımlı değişken ikili alt gruplara ayrılır. Ağacın başlama noktası, veri setindeki tüm birimleri içeren bir kök düğümdür. Daha sonra tekrarlı ikili bölünme süreci ile veri seti sağ ve sol düğüm olarak iki alt düğüme ayrılır. Bu işlemler her bir alt grubun yaprak düğüm olarak değerlendirilmesi sürecine kadar tekrar eder. CART algoritmasında bölünmenin başlayacağı değişken seçimi önemlidir (Belli ve Vantini, 2022; Breiman, 2017). Çeşitli safsızlık ölçüleri kullanılarak en iyi bölünmeyi sağlayan bağımsız değişken seçilir. Amaç, bağımlı değişkene ilişkin mümkün olan en homojen alt grupları üretmektir (Kurt Omurlu ve diğerleri, 2014). CART algoritması nicel bağımlı değişkenlere uygulandığında bölünme kuralı olarak gerçek değerler ile tahmin edilen değerler arasındaki farkı ölçen hata HKO uygulanır.

$\{(\mathbf{x}_i, y_i) \in b | \mathbf{x}_i \in R^p, y_i \in R\}$ herhangi bir b düğümü için eğitim setini temsil etsin. $n_b$, b düğümündeki birim sayısı olmak üzere b düğümündeki HKO değeri aşağıdaki gibi hesaplanır:

$$\mathrm{HKO(b)} = \frac{1}{n_b}\sum_{i=1}^{n_b}(y_i - \hat{y}_b)^2 \tag{13}$$

Burada b düğümündeki $\hat{y}_b$ tahmin değeri aşağıdaki gibi ifade edilir:

$$\hat{y}_b = \frac{1}{n_b}\sum_{i=1}^{n_b} y_i \tag{14}$$

Eğitim setindeki bağımsız değişkenlere göre b düğümü $b_{sağ}$ ve $b_{sol}$ şeklinde ikiye bölünürse, b düğümündeki bölünmede ortaya çıkan hata sağ ve sol alt düğümlerin hatalarının ağırlıklı ortalaması şeklinde aşağıdaki gibi hesaplanır:

$$\mathrm{HKO}_{\mathbf{x}_i}(b) = \frac{n_{b_{sağ}}}{n_b}\mathrm{HKO}\left(b_{sağ}\right) + \frac{n_{b_{sol}}}{n_b}\mathrm{HKO}(b_{sol}) \tag{15}$$

Her bir bağımsız değişken için olası ikili bölünmelerin her biri dikkate alınır. Sürekli bağımsız değişkenler için değişkenin her noktası kesim noktası olarak ele alınır ve her noktaya göre hatalar hesaplanır. İkili bir bölünmeden sonra safsızlık ölçüsündeki azalma aşağıdaki gibi hesaplanır:

$$\Delta\mathrm{HKO}(\mathbf{x}_i) = \mathrm{HKO(b)} - \mathrm{HKO}_{\mathbf{x}_i}(b) \tag{16}$$

En küçük hataya sahip olan, bir başka ifade ile safsızlıktaki azalmayı maksimum yapan bağımsız değişken, bölme değişkeni olarak seçilir.

#### 2.2.6.1.2. Sınıflandırma ve Regresyon Ağaçları Tabanlı Zincirleme Denklemlerle Çok Değişkenli Değer Atama Algoritması

MICE-CART eksik değer atama algoritmasının çalışma adımları aşağıda belirtilmiştir (Burgette ve Reiter, 2010; Doove ve diğerleri, 2014):

1. Tüm eksik değerler için ortalama değer atama ile ilk tahminler yapılır ve $\mathbf{X}$ matrisi güncellenir $(\acute{\mathbf{X}})$.
2. j. değişken bağımlı, diğer değişkenler bağımsız değişken olarak belirlenir.
3. CART modeli eğitilir $\left(x_{goz,j} \sim \acute{\mathbf{X}}_{goz_j,-j}\right)$ ve bölünme kuralları oluşturulur. Böylece yaprak düğümlerde $x_{goz,j}$ kümesinin elemanlarından oluşan aday havuzları oluşur.
4. Bölünme kuralları $\acute{\mathbf{X}}_{eks_j,-j}$ matrisi için uygulanır ve $x_{eks,j}$ eksik gözlemlerinin hangi yaprak düğümde yer alacağı belirlenir.
5. $x_{eks,j}$ eksik gözlemleri ait oldukları yaprak düğümlerdeki aday havuzlardan basit rastgele örnekleme ile alınan rastgele örneklerle güncellenir $(\acute{x}_{eks,j})$.
6. $\acute{\mathbf{X}}$ matrisinin j. değişkeni yeni tahmin değerleri ile güncellenir $(\acute{\mathbf{X}} \leftarrow \acute{x}_j)$.
7. $2. - 6.$ adımlar $j = 1,2, \dots, t$ için tekrar edilir.
8. m. iterasyon sonunda atanmış $\acute{\mathbf{X}}^{(m)}$ matrisi elde edilir $\left(\acute{\mathbf{X}}^{(m)} \leftarrow \acute{\mathbf{X}}\right)$.
9. $m \leq M$ ise 2. adıma gidilir. M sayıda atanmış veri seti elde edilene kadar iterasyon devam eder.

### 2.2.6.2. Yüksek Boyutlu Veriler için Geliştirilen Eksik Veri Değer Atama Yöntemleri

Çoklu atama yöntemleri için varsayılan eksik veri mekanizması MAR, test edilemeyen bir varsayım olmasına rağmen, değer atama modeline yeterli sayıda değişken dahil edilirse, MAR varsayımının sağlanacağı kabul edilir (Rubin, 1996). Ancak yüksek boyutlu verilerde tüm değişkenleri değer atama modeline dahil etmek mümkün değildir. Bu nedenle $p > n$ olması durumunda bazı değişken seçim prosedürleri uygulayarak eksik verilerin tahmininde en açıklayıcı değişkenleri belirleyen, böylece orijinal veri sayısından daha az sayıda değişkenle eksik değerleri tahmin eden MICE yöntemleri geliştirilmiştir (Yin ve diğerleri, 2016; Zou ve Hastie, 2005). Yüksek boyutlu veriler için geliştirilen yöntemlerdeki model katsayıları bir düzenlileştirilmiş regresyon yöntemi olan lasso regresyon ile elde edilir.

Aşağıdaki gibi bir lineer model düşünelim:

$$y = \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{17}$$

Burada $y = (y_1, y_2, \dots, y_n)^T$ sonuç değişkeni, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ regresyon katsayı vektörü, $\epsilon = (e_1, e_2, \dots, e_n)^T$ hata terimleridir.

$p < n$ olması durumunda en küçük kareler yöntemine göre hata kareler toplamını minimum yapan $\widehat{\boldsymbol{\beta}}$ katsayıları elde edilir. $p > n$ olması durumunda ise en küçük kareler yöntemi kullanılamaz bunun yerine düzenlileştirilmiş en küçük kareler yönteminden faydalanılır. Bu yönteme göre $\widehat{\boldsymbol{\beta}}$ katsayı tahminleri aşağıdaki gibi elde edilir (Zhao ve Long, 2016; Zou ve Hastie, 2005):

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (y - \mathbf{X}\boldsymbol{\beta})^T (y - \mathbf{X}\boldsymbol{\beta}) + p_\lambda(\boldsymbol{\beta}) \quad (18)$$

$\lambda$ ceza terimi olmak üzere $p_\lambda(\boldsymbol{\beta})$, bazı parametre tahmin değerlerini sıfıra yaklaştıran ya da sıfır yapan düzenlileştirme fonksiyonu olarak tanımlanır. Böylece veri setindeki bazı bağımsız değişkenlerin modeldeki etkisi azaltılır ya da tamamen ortadan kaldırılır. Bu tez çalışmasında lasso düzenlileştirme fonksiyonu kullanılmıştır. Buna göre $p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\boldsymbol{\beta}|$ olarak ifade edilir (Zou ve Hastie, 2005).

#### 2.2.6.2.1. Düzenlileştirilmiş Regresyonun Doğrudan Kullanımı (DURR)

DURR yöntemi eksik veriler yerine değer ataması yaparken hem ilgili değişkenlerin seçimi hem de oluşturulan modelin parametre tahminleri için düzenlileştirilmiş regresyonu kullanır. Çapraz geçerlilik yöntemi ile belirlenen $\lambda$ değerine bağlı olarak ilgisiz değişkenler modelden çıkarılır. Böylece yalnızca ilgili değişkenler ile model kurularak eksik değerler yerine atama yapılır. DURR algoritmasının çalışma adımları aşağıda belirtilmiştir (Costantini ve diğerleri, 2022; Y. Deng ve diğerleri, 2016; Zhao ve Long, 2016):

1. Tüm eksik değerler için ortalama değer atama ile ilk tahminler yapılır ve $\mathbf{X}$ matrisi güncellenir $(\acute{\mathbf{X}})$.
2. Yeni tahminler ile güncellenen $\acute{\mathbf{X}}$ matrisinden $n \times p$ boyutlarında bir bootstrap örnekleme matrisi oluşturulur $(\acute{\mathbf{X}}^*)$.
3. j. değişken bağımlı, diğer değişkenler bağımsız değişken olarak belirlenir.

4. Doğrusal model kurulur $\left( x^{*}_{goz,j} \sim \acute{\mathbf{X}}^{*}_{goz_j,-j} \right)$.
5. Lasso regresyon yöntemi ile model parametre tahminleri $\left(\widehat{\boldsymbol{\theta}}_j\right)$ elde edilir. Burada $\widehat{\boldsymbol{\theta}}_j$ tahminleri $\boldsymbol{\theta}_j$ 'nın koşullu dağılımdan çekilen rastgele örnekler ile yapılır: $\widehat{\boldsymbol{\theta}}_j \sim p\left(\boldsymbol{\theta}_j | x^{*}_{goz,j}, \acute{\mathbf{X}}^{*}_{goz_j,-j}\right)$.
6. $x_{eks,j}$ eksik değerlerinin tahmini dağılımından çekilen rastgele örnekler ile yeni eksik veri tahminleri hesaplanır: $\acute{\mathbf{x}}_{eks,j} \sim p\left(x_{eks,j} | \acute{\mathbf{X}}^{*}_{eks_j,-j}, \widehat{\boldsymbol{\theta}}_j\right)$.
7. $\acute{\mathbf{X}}$ matrisinin j. değişkeni yeni tahmin değerleri ile güncellenir $\left(\acute{\mathbf{X}} \leftarrow \acute{x}_j\right)$.
8. $2.-7.$ adımlar $j = 1,2,\ldots,t$ için tekrar edilir.
9. m. iterasyon sonunda atanmış $\acute{\mathbf{X}}^{(m)}$ matrisi elde edilir $\left(\acute{\mathbf{X}}^{(m)} \leftarrow \acute{\mathbf{X}}\right)$.
10. $m \leq M$ ise 2. adıma gidilir. M sayıda atanmış veri seti elde edilene kadar iterasyon devam eder.

#### 2.2.6.2.2. Düzenlileştirilmiş Regresyonun Dolaylı Kullanımı (IURR)

IURR yöntemi eksik veriler yerine değer ataması yaparken sadece bağımsız değişkenlerin seçimi için düzenlileştirilmiş regresyonu kullanır (Costantini ve diğerleri, 2022). Modelin parametre tahminleri düzenlileştirilmiş regresyon ile seçilen bağımsız değişkenler matrisi kullanılarak en çok olabilirlik yöntemi ile yapılır. En çok olabilirlik yöntemi $\boldsymbol{\theta}$'yı elde etme olasılığını maksimum yapan parametre tahminlerini verir. IURR algoritmasının çalışma adımları aşağıda belirtilmiştir (Costantini ve diğerleri, 2022; Y. Deng ve diğerleri, 2016; Zahid ve diğerleri, 2021; Zhao ve Long, 2016):

1. Tüm eksik değerler için ortalama değer atama ile ilk tahminler yapılır ve $\mathbf{X}$ matrisi güncellenir $\left(\acute{\mathbf{X}}\right)$.
2. j. değişken bağımlı, diğer değişkenler bağımsız değişken olarak belirlenir.
3. Doğrusal model kurulur $\left(x_{goz,j} \sim \acute{\mathbf{X}}_{goz_j,-j}\right)$ ve lasso regresyon yöntemi ile parametre tahminleri elde edilir.
4. Regresyon katsayıları sıfırdan büyük olan değişkenler seçilir $\left(\acute{\mathbf{X}}^{s}_{-j}\right)$
5. Seçilen değişkenler ile yeni bir doğrusal model kurulur $\left(x_{goz,j} \sim \acute{\mathbf{X}}^{s}_{goz_j,-j}\right)$.

6. En çok olabilirlik yöntemi ile model parametre tahminleri $\left(\widehat{\boldsymbol{\theta}}_{\mathrm{j}}\right)$ elde edilir. Burada $\widehat{\boldsymbol{\theta}}_{\mathrm{j}}$ tahminleri $\boldsymbol{\theta}_{\mathrm{j}}$ 'nın koşullu dağılımdan çekilen rastgele örnekler ile yapılır: $\widehat{\boldsymbol{\theta}}_{\mathrm{j}} \sim \mathrm{p}\left(\boldsymbol{\theta}_{\mathrm{j}} | \mathrm{x}_{\mathrm{goz,j}}, \acute{\mathbf{X}}^{\mathrm{s}}_{\mathrm{goz_j,-j}}\right)$.
7. $\mathbf{x}_{\mathrm{eks,j}}$ eksik değerlerinin tahmini dağılımından çekilen rastgele örnekler ile yeni eksik değer tahminleri hesaplanır: $\acute{\mathbf{x}}_{\mathrm{eks,j}} \sim \mathrm{p}\left(\mathrm{x}_{\mathrm{eks,j}} | \acute{\mathbf{X}}^{\mathrm{s}}_{\mathrm{eks_j,-j}}, \widehat{\boldsymbol{\theta}}_{\mathrm{j}}\right)$.
8. $\acute{\mathbf{X}}$ matrisinin j. değişkeni yeni tahmin değerleri ile güncellenir $\left(\acute{\mathbf{X}} \leftarrow \acute{\mathrm{x}}_{\mathrm{j}}\right)$.
9. $2.-8.$ adımlar $\mathrm{j} = 1,2,\ldots,\mathrm{t}$ için tekrar edilir.
10. m. iterasyon sonunda atanmış $\acute{\mathbf{X}}^{(\mathrm{m})}$ matrisi elde edilir $\left(\acute{\mathbf{X}}^{(\mathrm{m})} \leftarrow \acute{\mathbf{X}}\right)$.
11. $\mathrm{m} \leq \mathrm{M}$ ise 2. adıma gidilir. M sayıda atanmış veri seti elde edilene kadar iterasyon devam eder.

### 2.3. Aşırı Öğrenme Makineleri (ELM)

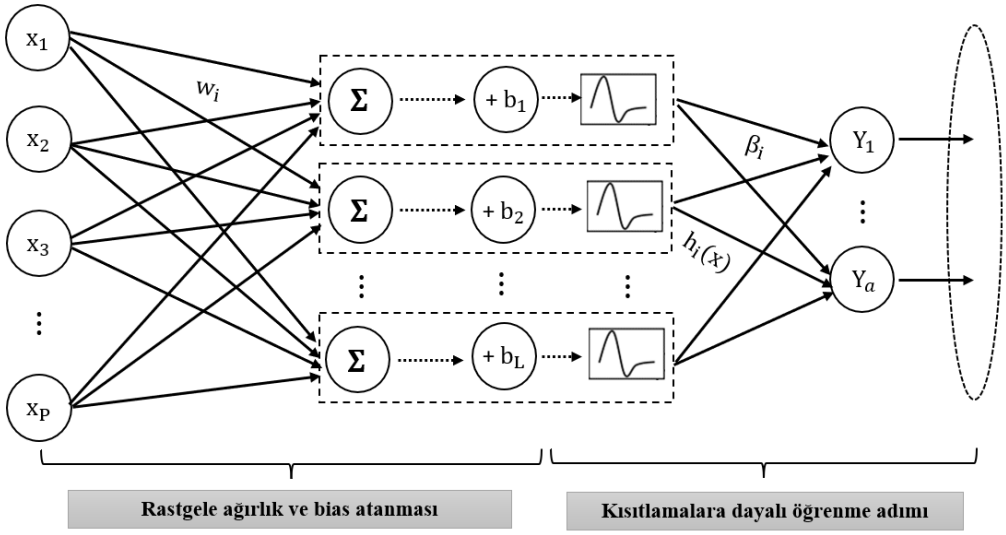
ELM, genelleştirilmiş tek gizli katmanlı ve ileri beslemeli yapay sinir ağının eğitimi için kullanılan bir makine öğrenme yöntemidir (Chen ve diğerleri, 2014; G.-B. Huang ve diğerleri, 2004; G. Huang ve diğerleri, 2015; Kasun ve diğerleri, 2016). ELM teorisinde eğitim setinden bağımsız olarak $\mathrm{w_i}$, gizli katman ağırlıkları ve $\mathrm{b_i}$, bias değerleri rastgele türetilir ve bu parametrelerin yeniden ayarlanmasına gerek yoktur. Çıktı ağırlık değerleri de bu parametrelere bağlı olarak hesaplanmaktadır. Böylece sadece eğitim hatasını minimum yapan geri yayılımlı öğrenme algoritmasından farklı olarak, ELM algoritması hem eğitim hatasını hem de çıktı ağırlıklarını minimum yapmaktadır (Chen ve diğerleri, 2014; Kasun ve diğerleri, 2016).

n gözlem p değişkenden oluşan bir eğitim seti $\{(\mathbf{x}_{\mathrm{i}}, \mathrm{y}_{\mathrm{i}}) | \mathbf{x}_{\mathrm{i}} \in \mathrm{R}^{\mathrm{p}}, \mathrm{y}_{\mathrm{i}} \in \mathrm{R}^{a}\}_{\mathrm{i}=1}^{\mathrm{n}}$ ($a$ =1 ise regresyon problemi, $a \geq 2$ ise sınıflandırma problemi) olmak üzere L sayıda gizli düğüme sahip bir ileri beslemeli yapay sinir ağının yapısı aşağıdaki eşitlik ile gösterilir (Şekil 4) (Cantaş Türkiş ve diğerleri, 2024; Kasun ve diğerleri, 2016; Wang ve Li, 2019):

$$\mathrm{f_L}(\mathrm{x}) = \sum_{\mathrm{i}=1}^{\mathrm{L}} \mathrm{g}(\mathbf{x}, \mathrm{w_i}, \mathrm{b_i})\, \beta_{\mathrm{i}} = \mathbf{h}(\mathbf{x})\boldsymbol{\beta} \tag{19}$$

Bu eşitlikte g(.) aktivasyon fonksiyonu; $\mathbf{x}$ bağımsız değişkenler vektörü, $\mathrm{w_i} \in \mathrm{R}^{\mathrm{p}}$, giriş ve çıkış katmanı arasındaki giriş ağırlıkları vektörü; $\mathrm{b_i}$, i. gizli katmana ilişkin bias; $\boldsymbol{\beta} =$

$[\beta_1, \ldots, \beta_i, \ldots, \beta_L]^T$ çıkış ağırlıkları vektörü; $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}),\ h_2(\mathbf{x}), \ldots, h_L(\mathbf{x})]^T$ ise gizli katmanlardır.




**Şekil 4.** ELM yapısı (Cui ve diğerleri, 2019)

Gizli katman çıkış ağırlıkları matrisi **H** aşağıdaki gibi gösterilebilir:

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}(x_1) \\ \vdots \\ \mathbf{h}(x_n) \end{bmatrix} = \begin{bmatrix} g(w_1, b_1, x_1) & \cdots & g(w_L, b_L, x_1) \\ \vdots & \ddots & \vdots \\ g(w_1, b_1, x_n) & \cdots & g(w_L, b_L, x_n \end{bmatrix}_{nxL} \tag{20}$$

Sonuç değişkeni matrisi $\mathbf{Y} = [y_1, \ldots, y_i, \ldots, y_n]^T$ olmak üzere Eşitlik (19) aşağıdaki gibi genel formda tekrar yazılabilir (Lu ve diğerleri, 2019):

$$\mathbf{Y} = \mathbf{H}\boldsymbol{\beta} \tag{21}$$

Böylece eğitim hatasını minimum yapacak şekilde $\boldsymbol{\beta}$ çıktı ağırlıklarının çözümü hesaplanır:

$$\hat{\beta} = \text{argmin}\|\mathbf{Y} - \mathbf{H}\boldsymbol{\beta}\| \tag{22}$$

Çıktı ağırlıkları için optimal tahmin Moore-Penrose genelleştirilmiş ters matrisi ($\mathbf{H}^{\dagger}$) ile elde edilir (Cantaş Türkiş ve diğerleri, 2024; Liang ve diğerleri, 2006):

$$\widehat{\boldsymbol{\beta}} = \mathbf{H}^{\dagger}\mathbf{Y} \tag{23}$$

Burada $\mathbf{H}^{\dagger} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}$ eşitliği ile elde edilir ve $\mathbf{H}$ matrisinin dik izdüşümüdür. Daha kararlı ve genellenebilir sonuçlar elde edilebilmesi için ridge regresyon teorisine göre $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ matrisinin köşegenlerine pozitif bir değer eklenir. Böylece $\mathbf{I}_{\mathrm{nxn}}$ birim matris, C çıktı ağırlıkları ile eğitim hatası arasındaki denge parametresi olmak üzere çıktı ağırlıklarının kapalı formdaki çözümü aşağıdaki eşitlikle gösterilir (W. Deng ve diğerleri, 2009; G.-B. Huang ve diğerleri, 2006):

$$\widehat{\boldsymbol{\beta}} = \left(\frac{\mathbf{I}}{\mathrm{C}} + \mathbf{H}^{\mathrm{Y}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{Y}}\mathbf{Y} \tag{24}$$

Gizli katmanda yer alan $\mathrm{h}(\cdot)$ fonksiyonu bilinmiyorsa, pozitif tanımlı bir $\mathrm{K}(\cdot,\cdot)$ çekirdek fonksiyonu belirlenerek gizli katmanlı ve ileri beslemeli yapay sinir ağının yapısı aşağıdaki şekle dönüşür (G.-B. Huang ve diğerleri, 2006; Wang ve Li, 2019):

$$\mathrm{K}(\mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{k}}) = \mathrm{h}(\mathbf{x}_{\mathrm{i}}) * \mathrm{h}(\mathbf{x}_{\mathrm{k}}) \tag{25}$$

Böylece çekirdek ELM ya da çekirdek fonksiyonlu, gizli katmanlı ve ileri beslemeli yapay sinir ağı aşağıdaki şekilde modellenir:

$$\mathrm{f}_{\mathrm{L}}(\mathbf{x}_{\mathbf{i}}) = \sum_{\mathrm{k}=1}^{\mathrm{L}} \mathrm{K}(\mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{k}})\beta_{\mathrm{k}} \ , \qquad \mathrm{i} = 1,2, \dots, \mathrm{n} \tag{26}$$

Bu model aşağıdaki eşitlikle yeniden ifade edilebilir (Wang ve Li, 2019):

$$\mathrm{f}_{\mathrm{L}}(\mathbf{x}) = \mathbf{K}_{\mathrm{nxL}}\boldsymbol{\beta} \tag{27}$$

### 2.3.1. Düzleştirilmiş Doğrusal Birim (RELU) Aktivasyon Fonksiyonu

Aktivasyon fonksiyonları, bir yapay sinir ağındaki gizli nöronlarda, ağa doğrusal olmayan modelleme yeteneği sağlamak için kullanılan fonksiyonlardır. Bu fonksiyonlar bir katmandaki düğümlerin çıkışını bir sonraki katmana iletir. Negatif girdiler için 0 değerini alan, pozitif girdiler için ise girdi değerini değiştirmeden çıktı olarak veren düzleştirilmiş doğrusal birim (rectified linear unit, RELU) fonksiyonu sıklıkla kullanılan aktivasyon fonksiyonlarından biridir. ELM için RELU fonksiyonu aşağıdaki eşitlik ile tanımlanır (Agarap, 2018; G. Huang ve diğerleri, 2015; Khan ve diğerleri, 2018):

$$f(x) = \text{maks}(0, \mathbf{wx} + b) = \begin{cases} \mathbf{wx} + b, & \mathbf{wx} + b > 0 \\ 0, & \mathbf{wx} + b \leq 0 \end{cases} \quad (28)$$

## 2.4. Yöntemlerin Performanslarının Değerlendirmesinde Kullanılan Ölçütler

Çalışmamızdaki eksik veri değer atama yöntemlerinin orijinal gözlem değerlerini tahmin etme performansları $HKO_{DA}$ile; sınıflandırma performansına etkileri dengeli doğruluk oranı, AUC ve kappa ölçütleri ile değerlendirildi.

### 2.4.1. Değer Atama Hata Kareler Ortalaması

Eksik verilerin yerine atanan değerlerin orijinal gözlem değerlerini tahmin etme performansları $HKO_{DA}$ ile değerlendirilir. $N_{eks}$ veri setindeki toplam eksik gözlem sayısı; $\acute{\mathbf{x}}_{i,eks,,m}$ i. birimdeki eksik veriler için m. iterasyon sonucu atanan değerler, $\mathbf{x}_{i,orj}$ i. birimdeki eksik verilere karşılık gelen tam veri setindeki orijinal değerler olmak üzere; $HKO_{DA}$, tekli değer atama yöntemleri için Eşitlik (29) ile; çoklu değer atama yöntemleri için Eşitlik (30) ile hesaplanır (Van Buuren, 2018; Zahid ve diğerleri, 2021):

$$HKO_{DA} = \frac{1}{N_{eks}} \sum_{i=1}^{n_{eks}} \left(\mathbf{x}_{i,orj} - \acute{\mathbf{x}}_{i,eks}\right)^2 \quad (29)$$

$$\mathrm{HKO_{DA}} = \frac{1}{\mathrm{M}} \sum_{\mathrm{m=1}}^{\mathrm{M}} \frac{1}{\mathrm{N_{eks}}} \sum_{\mathrm{i=1}}^{\mathrm{n_{eks}}} \left(\mathbf{x}_{\mathrm{i,orj}} - \acute{\mathbf{x}}_{\mathrm{i,eks,,m}}\right)^2 \tag{30}$$

### 2.4.2. Dengeli Doğruluk Oranı

Dengeli doğruluk oranı değerlerinin hesaplanması Tablo 1'de verilen sınıflandırma tablosundaki frekanslar kullanılarak gerçekleştirildi.

**Tablo 1.** 2x2 sınıflandırma tablosu

| | | **GERÇEK DURUM** | | |
|---|---|---|---|---|
| | | **Pozitif** | **Negatif** | **Toplam** |
| **TAHMİN** | **Pozitif** | Gerçek Pozitif (GP) | Yalancı Pozitif (YP) | GP+YP |
| | **Negatif** | Yalancı Negatif (YN) | Gerçek Negatif (GN) | YN+GN |
| | **Toplam** | DP+YN | YP+DN | N |

Dengeli doğruluk oranı, sınıflandırma modelinin doğru tahmin ettiği gerçek pozitiflerin oranı olarak tanımlanan duyarlılık oranı ile sınıflandırma modelinin doğru tahmin ettiği gerçek negatiflerin oranı olarak tanımlanan özgüllük oranlarının aritmetik ortalaması alınarak hesaplanır (Brodersen ve diğerleri, 2010):

$$\text{Dengeli doğruluk} = \frac{1}{2}\left(\frac{\mathrm{GP}}{\mathrm{GP + YN}} + \frac{\mathrm{GN}}{\mathrm{GN + YP}}\right) \tag{31}$$

### 2.4.3. ROC Eğrisi Altında Kalan Alan

Bağımsız iki grup, sonucu nicel verilerden elde edilen bir test ile ayırt edilmek istendiğinde ROC eğrilerinden yararlanılır (Alpar, 2010). AUC, farklı kesim noktaları için yapılan sınıflandırma tahminlerinden elde edilen duyarlılık ve 1-özgüllük oranı değerlerine göre oluşturulan ROC eğrisi altında kalan alan olarak tanımlanır (Fawcett, 2006; Hanley ve McNeil, 1982). AUC $(0,1)$ arasında değer alır ve 1'e yaklaştıkça sınıflandırma modelinin tahmin performansı artar.

$$\text{AUC} = \int_0^1 \text{ROCeğrisi(x)dx} \tag{32}$$

### 2.4.4. Cohen'in Kappa Katsayısı

Kappa katsayısı, iki sınıflı isimsel ölçekli veriler için gözlenen ve beklenen değerler arasındaki uyumu ölçen bir fonksiyondur. Kappa katsayısı [0,1] arasında değer alır. Sınıflandırma modeli tarafından tahmin edilen sınıf değerleri ile gerçek sınıf değerleri arasındaki uyum mükemmel olduğunda kappa katsayısı 1'e eşit olurken; arada uyum olmadığında kappa katsayısı 0'a eşit olur (Cohen, 1960; Warrens, 2015). Gerçek sınıf değerleri ile tahmin edilen sınıf değerlerinin aynı sonucu verme olasılığı $\rho_0$ ile; arada şansa bağlı bir uyumun olması olasılığı $\rho_c$ ile ifade edilmek üzere kappa katsayısı aşağıdaki gibi hesaplanır (Cohen, 1960):

$$\text{Kappa} = \frac{\rho_0 - \rho_c}{1 - \rho_c} \tag{33}$$

# 3. GEREÇ VE YÖNTEM

Bu çalışmanın uygulama bölümünde, yüksek boyutlu verilerde 8 farklı eksik veri değer atama yönteminin (ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR) orijinal gözlem değerlerini tahmin etme performansları ve bu yöntemler ile tamamlanan veri setlerinin ELM ile sınıflandırılması sonucunda sınıflandırma performanslarının nasıl etkilendiği incelendi. Bu amaçla hazırlanan simülasyon algoritmaları, n=150 birim ve p=500 değişken için, farklı korelasyon düzeyleri ve eksik veri oranlarına göre oluşturuldu. 1000 döngü ile gerçekleştirilen simülasyonlar ile elde edilen sonuçlara göre modeller karşılaştırıldı.

## 3.1. Simülasyon Algoritmaları

Bu çalışmada 2 farklı simülasyon algoritması oluşturuldu. İlk algoritmada eksik değer içeren değişkenler veri setindeki belirli bir değişken setine bağlı olarak türetildi. İkinci algoritmada ise tüm değişkenler tamamen rastgele türetildi. Böylece farklı veri yapıları için yöntemlerin performanslarının incelenmesi amaçlandı. Bu amaçla yazılan simülasyon algoritmaları aşağıda açıklanmıştır.

### 3.1.1. Rastgele Bir Değişken Setinin Doğrusal Kombinasyonundan Türetilen Eksik Verili Değişkenler için Simülasyon Algoritması

1. Örneklem büyüklüğü (n=150) ve bağımsız değişken sayısı ($p_1$=490) olarak belirlendi.
2. Korelasyon düzeyleri birinci durumda $-0{,}1 \leq r \leq 0{,}1$; ikinci durumda $-0{,}5 \leq r \leq 0{,}5$; üçüncü durumda $-0{,}8 \leq r \leq 0{,}8$ değerleri arasında rastgele değişen bir yapıda olmak üzere çok değişkenli standart normal dağılımdan $\mathbf{X_1} \sim N_{p_1}(\mathbf{0}, \mathbf{\Sigma})$ bağımsız değişkenleri türetildi.
3. $\mathbf{X_1}$ matrisinden basit rastgele örnekleme ile 50 değişken seçildi. Model katsayıları sabit bir değer (0,45) olacak şekilde 50 değişkenin doğrusal kombinasyonu alınarak bir model oluşturuldu. Oluşturulan modelden elde edilen ortalama değerleri ($\boldsymbol{\mu_2}$) alınarak ve

standart sapma $\sigma_2 = 1{,}22$ olacak şekilde normal dağılımdan $p_2 = 10$ tane eksik değer oluşturulacak bağımsız değişkenler ($\mathbf{X_2} \sim N_{p_2}(\boldsymbol{\mu_2}, \sigma_2^2)$) türetildi.

4. $\mathbf{X_1}$ ve $\mathbf{X_2}$ matrisleri birleştirilerek bağımsız değişken sayısı (p=500) olan $\mathbf{X}$ matrisi elde edildi.
5. Model katsayıları sabit bir değer (0,5) olacak şekilde $\mathbf{X}$ matrisinin doğrusal kombinasyonu alınarak bir model oluşturuldu. Oluşturulan modelden elde edilen ortalama değerleri ($\boldsymbol{\mu}$) alınarak ve standart sapma $\sigma = 1{,}41$ olacak şekilde normal dağılımdan $N_p(\boldsymbol{\mu}, \sigma^2)$ değerleri türetildi. Daha sonra bu değerler 50. persantil değerine göre iki gruba ayrıldı ve iki sınıflı Y bağımlı değişkeni oluşturuldu.
6. Tüm veri setlerinin, $\mathbf{X_2}$ matrisinin elemanları olan $p_2 = 10$ değişkeninde sırasıyla %10, %20, %30, %40 ve %50 oranlarında MAR mekanizmalı eksik değerler oluşturuldu.
7. Eksik değerli veri setleri ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR eksik veri değer atama yöntemleri ile tamamlandı ve yöntemlerin $HKO_{DA}$ değerleri hesaplandı.
8. Tam ve atanmış tüm veri setleri, 70:30 oranında eğitim-test setleri olarak rasgele ayrıldı.
9. Eğitim setlerinde ELM yöntemi ile tahmin modelleri oluşturuldu.
10. Oluşturulan tahmin modelleri ile test setlerinde dengeli doğruluk oranı, AUC ve kappa değerleri hesaplandı.
11. 1. ve 10. adımlar 1000 kez tekrar edildi.
12. Elde edilen sonuçların tanımlayıcı istatistikleri hesaplandı, performans metriklerinin dağılımına uygun orman grafikleri çizildi. Yöntemlerin $HKO_{DA}$ değerlerine göre orijinal gözlem değerlerini tahmin performansı değerlendirildi.
13. Elde edilen sonuçlara aşamalı kümeleme analizi uygulanarak dendrogram grafikleri çizildi. Böylece referans veri setlerinden elde edilen değerlere ve birbirine yakın performans gösteren yöntemler belirlendi.

### 3.1.2. Tamamen Rastgele Türetilen Veriler için Simülasyon Algoritması

1. Örneklem büyüklüğü (n=150) ve bağımsız değişken sayısı (p=500) için sabit bir değer atandı.
2. Korelasyon düzeyleri birinci durumda $-0{,}1 \leq r \leq 0{,}1$; ikinci durumda $-0{,}5 \leq r \leq 0{,}5$; üçüncü durumda $-0{,}8 \leq r \leq 0{,}8$ değerleri arasında rastgele değişen bir yapıda olmak

üzere çok değişkenli standart normal dağılımdan $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$ bağımsız değişkenleri türetildi.

3. Model katsayıları standart normal dağılımdan elde edilen **X** matrisinin doğrusal kombinasyonu alınarak bir model oluşturuldu. Oluşturulan modelden elde edilen ortalama değerleri ($\boldsymbol{\mu}$) alınarak ve standart sapma $\sigma = 1{,}41$ olacak şekilde normal dağılımdan $N_p(\boldsymbol{\mu}, \sigma^2)$ değerleri türetildi. Daha sonra bu değerler 50. persantil değerine göre iki gruba ayrıldı ve iki sınıflı Y bağımlı değişkeni oluşturuldu.
4. Tüm veri setlerinin rastgele seçilen 10 bağımsız değişkeninde sırasıyla %10, %20, %30, %40 ve %50 oranlarında MAR mekanizmalı eksik değerler oluşturuldu.
5. Eksik değerli veri setleri ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR eksik veri değer atama yöntemleri ile tamamlandı ve yöntemlerin $HKO_{DA}$ değerleri hesaplandı.
6. Tam ve atanmış tüm veri setleri, 70:30 oranında eğitim-test setleri olarak rasgele ayrıldı.
7. Eğitim setlerinde ELM yöntemi ile tahmin modelleri oluşturuldu.
8. Oluşturulan tahmin modelleri ile test setlerinde dengeli doğruluk oranı, AUC ve kappa değerleri hesaplandı.
9. 1. ve 8. adımlar 1000 kez tekrar edildi.
10. Elde edilen sonuçların tanımlayıcı istatistikleri hesaplandı, performans metriklerinin dağılımına uygun orman grafikleri çizildi. Yöntemlerin $HKO_{DA}$ değerlerine göre orijinal gözlem değerlerini tahmin performansı değerlendirildi.
11. Elde edilen sonuçlara aşamalı kümeleme analizi uygulanarak dendrogram grafikleri çizildi. Böylece referans veri setlerinden elde edilen değerlere ve birbirine yakın performans gösteren yöntemler belirlendi.

### 3.2. Değer Atama ve Sınıflandırma Modellerine İlişkin Parametreler

Çalışmada kullanılan eksik veri değer atama ve ELM yöntemlerinden faydalanan önceki çalışmalardaki en sık kullanılan parametre değerleri ile denemeler yapılarak, yöntemlerimiz için en yüksek performansı veren parametreler belirlendi. Buna göre, KNN değer atama yöntemi için komşuluk sayısı k=5 olarak atandı. I-RF yöntemi için oluşturulması planlanan karar ağacı sayısı 100, her düğüm ayrımında rasgele seçilecek olan bağımsız değişken sayısı $p/3 \cong 167$, maksimum iterasyon sayısı 10 olarak belirlendi. MICE-CART yönteminde 10 kat

çapraz geçerlilik yöntemi ile CART modeli eğitildi. Bölünmenin gerçekleşebilmesi için her bir düğümde olması gereken minimum gözlem sayısı 20, her terminal düğümde olması gereken minimum gözlem sayısı 7, oluşturulan karar ağacı için maksimum derinlik sayısı 30, model yanlılığı ve varyansı arasındaki optimum noktayı belirlemek için karmaşıklık katsayısı 0,01 olarak belirlendi. MICE tabanlı değer atama yöntemleri MICE-CART, DURR ve IURR için iterasyon sayısı $M = 5$ olarak belirlendi. DURR ve IURR yöntemlerinde model eğitim aşamasında kullanılan lasso regresyon için $\lambda$ ceza terimi 10 kat çapraz geçerlilik yöntemi ile belirlendi.

Bu çalışmada yapılan tüm sınıflandırmalar ELM yöntemi ile gerçekleştirildi. Model oluşturma aşamasında rastgele ağırlık ataması standart normal dağılımdan yapıldı. Aktivasyon fonksiyonu olarak RELU fonksiyonu kullanıldı. Modellerdeki gizli katman sayısı 20 olarak belirlendi.

### 3.3. Kullanılan Programlar

Bu çalışmada yer alan uygulamalar R programlama dilinin 4.2.3 versiyonu kullanılarak gerçekleştirildi. Veri türetimi, analizi ve sonuçların kaydedilmesi için R programlama dilinde rpart, data.table, MASS, mvtnorm, Metrics, dplyr, stats, plyr, dplyr, VIM, glmnet, e101, caret, pROC, openxlsx, SimDesign, norm, sigmoid, tidyverse, stringr, FactoMineR, MBESS, evolqg, lqmm, matrixcalc, InformationValue, mice, missForest paketleri kullanıldı. Zincirleme denklemler ile çok değişkenli değer atama yöntemleri için mice paketi genişletilerek oluşturulan impute_MICE_CART, impute_DURR ve impute_IURR fonksiyonları kullanıldı.

# 4. BULGULAR

Ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR eksik veri değer atama yöntemlerinin ELM ile sınıflandırma performansına etkilerinin karşılaştırılması için hesaplanan dengeli doğruluk oranı, AUC ve kappa değerlerinin normal dağılıma uygunluğu Kolmogorov-Smirnov analizi ile test edildi. Performans ölçütlerinin dağılımı normal dağılıma uygunluk göstermediği için modellerin performanslarına ilişkin tanımlayıcı istatistikler medyan (25. – 75. persantil) şeklinde verildi. Yöntemler arasındaki ilişkilerin belirlenebilmesi amacıyla değişen korelasyon düzeylerine ve eksik oranlarına göre yöntemlerin dengeli doğruluk, AUC ve kappa sonuçları kullanılarak aşamalı kümeleme analizi uygulandı.

## 4.1. Rastgele Bir Değişken Setinin Doğrusal Kombinasyonundan Türetilen Eksik Verili Değişkenler için Bulgular

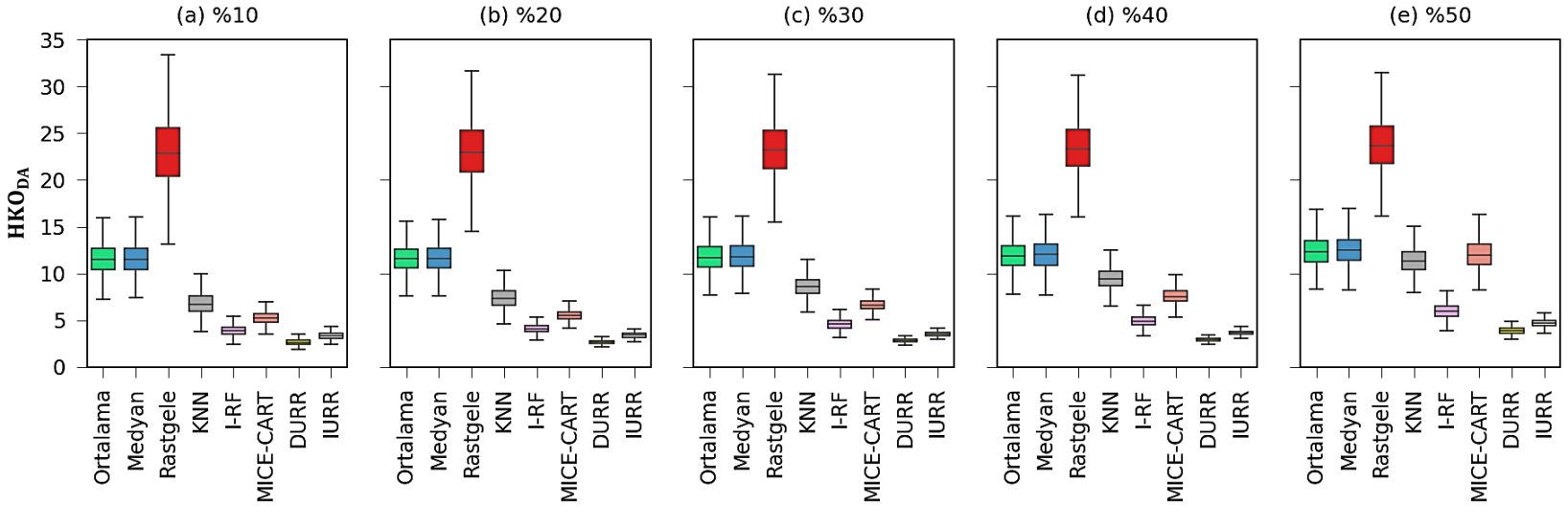
### 4.1.1. $-0{,}1 \leq r \leq 0{,}1$ Aralığına göre Rastgele Bir Değişken Setinin Doğrusal Kombinasyonundan Türetilen Eksik Verili Değişkenler için Bulgular

Eksik oranı %10, %20, %30, %40 ve %50 için $HKO_{DA}$ değerlerinin medyan değişim aralığı sırasıyla 2,629-22,886; 2,680-22,933; 2,804-23,215; 2,932-23,344; 3,892-23,679'dur. Değer atama yöntemlerinin $HKO_{DA}$ değerlerine ilişkin medyan değişim aralığı ise ortalama için 11,462-12,355; medyan için 11,542-12,511; rastgele için 22,886-23,679; KNN için 6,703-11,309; I-RF için 3,856-5,930; MICE-CART için 5,208-12,003; DURR için 2,629-3,892; IURR için 3,309-4,686'dır (Tablo 2).

Tüm değer atama yöntemlerinin $HKO_{DA}$ değerleri değişen eksik oranına göre incelendiğinde genel olarak eksik oranı arttıkça yöntemlerin $HKO_{DA}$ değerlerinin artış eğiliminde olduğu görülmüştür. DURR ve IURR yöntemlerinin tüm eksik oranları için diğer yöntemlere göre daha düşük $HKO_{DA}$ değerine sahip olduğu ve bu yöntemleri I-RF yönteminin takip ettiği; en yüksek $HKO_{DA}$ değerinin rastgele değer atama yönteminden elde edildiği belirlenmiştir (Tablo 2 ve Şekil 5).

**Tablo 2**. -0,1 ≤ r ≤ 0,1 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerleri

| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| **$HKO_{DA}$** | **Ortalama** | 11,462 (10,381 - 12,662) | 11,566 (10,569 - 12,587) | 11,698 (10,734 - 12,862) | 11,903 (10,842 - 12,996) | 12,355 (11,276 - 13,529) |
| | **Medyan** | 11,542 (10,417 - 12,729) | 11,580 (10,616 - 12,695) | 11,804 (10,813 - 12,971) | 12,028 (10,904 - 13,108) | 12,511 (11,373 - 13,638) |
| | **Rastgele** | 22,886 (20,366 - 25,603) | 22,933 (20,876 - 25,300) | 23,215 (21,219 - 25,283) | 23,344 (21,513 - 25,423) | 23,679 (21,807 - 25,735) |
| | **KNN** | 6,703 (5,944 - 7,581) | 7,354 (6,600 - 8,111) | 8,567 (7,881 - 9,344) | 9,447 (8,656 - 10,219) | 11,309 (10,392 - 12,316) |
| | **I-RF** | 3,856 (3,493 - 4,263) | 4,081 (3,761 - 4,411) | 4,580 (4,185 - 4,991) | 4,914 (4,475 - 5,373) | 5,930 (5,397 - 6,488) |
| | **MICE-CART** | 5,208 (4,813 - 5,699) | 5,536 (5,182 - 5,921) | 6,567 (6,205 - 7,058) | 7,542 (7,053 - 8,188) | 12,003 (10,945 - 13,134) |
| | **DURR** | 2,629 (2,414 - 2,842) | 2,680 (2,528 - 2,822) | 2,804 (2,688 - 2,937) | 2,932 (2,805 - 3,067) | 3,892 (3,648 - 4,171) |
| | **IURR** | 3,309 (3,061 - 3,564) | 3,384 (3,196 - 3,572) | 3,533 (3,376 - 3,688) | 3,665 (3,518 - 3,830) | 4,686 (4,423 - 4,978) |




**Şekil 5.** -0,1 ≤ r ≤ 0,1 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği

Referans veri setinden elde edilen dengeli doğruluk oranı değerlerinin medyanı 0,750 olarak elde edilmiştir. Atanmış veri setinden elde edilen sonuçlara göre eksik oranı %10, %20, %30, %40 ve %50 için dengeli doğruluk oranlarının medyan değişim aralığı sırasıyla 0,736-0,750; 0,729-0,747; 0,695-0,745; 0,680-0,741; 0,640-0,727'dir. Değer atama yöntemlerinin dengeli doğruluk oranlarına ilişkin medyan değişim aralığı ise ortalama için 0,646-0,738; medyan için 0,647-0,736; rastgele için 0,640-0,737; KNN için 0,669-0,744; I-RF için 0,710-0,747; MICE-CART için 0,702-0,750; DURR için 0,725-0,747; IURR için 0,727-0,746'dır (Tablo 3 ve Şekil 6).

Referans veri setinden elde edilen AUC değerlerinin medyanı 0,776 olarak elde edilmiştir. Atanmış veri setinden elde edilen sonuçlara göre AUC değerlerinin medyan değişim aralığı eksik oranı %10 için 0,769-0,778; %20 için 0,749-0,777; %30 için 0,704-0,775; %40 için 0,682-0,771; %50 için 0,624-0,752'dir. Ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR yöntemlerinin AUC değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,638-0,769; 0,636-0,770; 0,624-0,769; 0,665-0,775; 0,723-0,776; 0,713-0,778; 0,750-0,777; 0,752-0,778'dir (Tablo 3 ve Şekil 7).

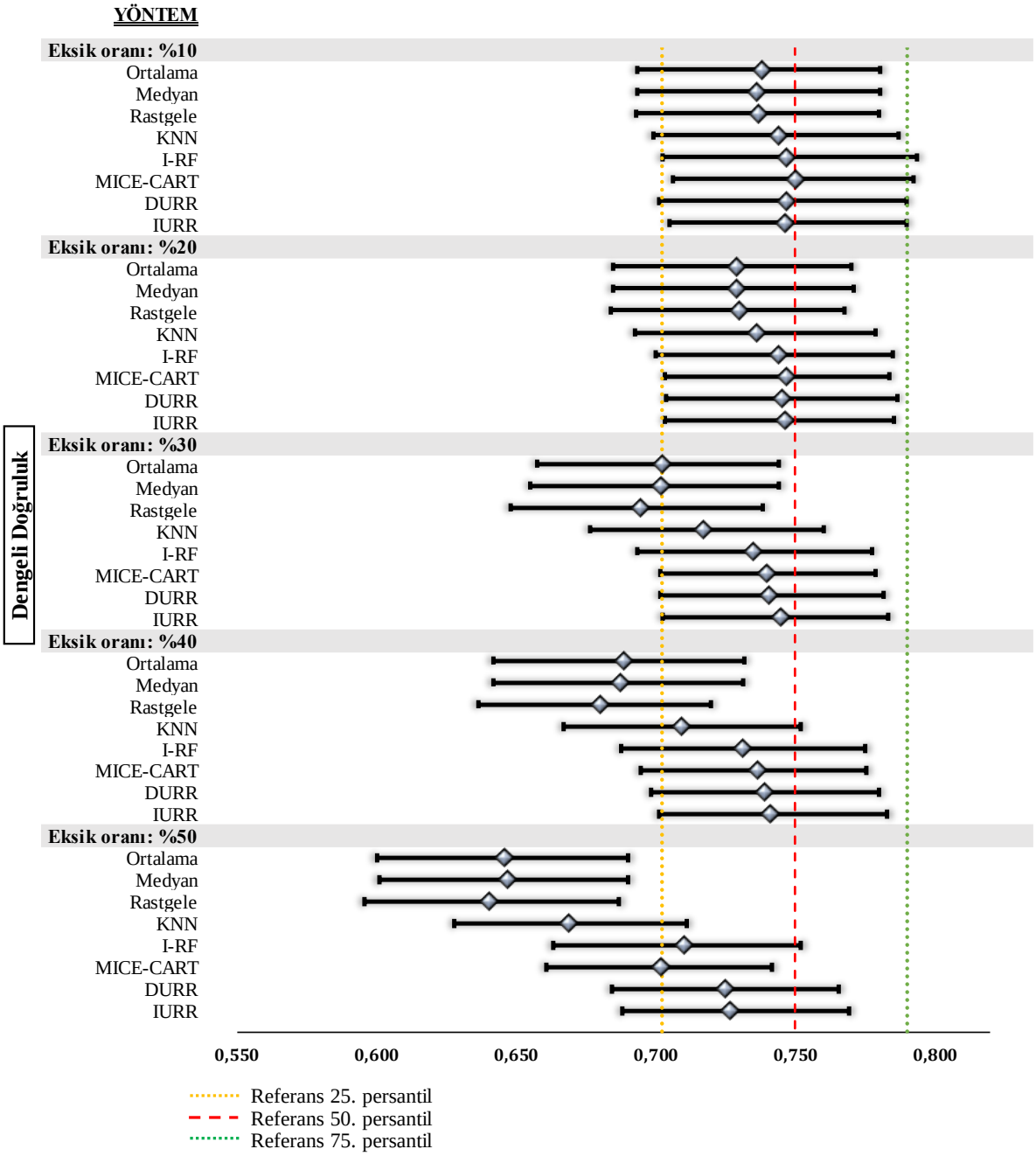
Referans veri setinden elde edilen kappa değerlerinin medyanı 0,500 olarak elde edilmiştir. Atanmış veri setinden elde edilen sonuçlara göre kappa değerlerinin medyan değişim aralığı eksik oranı %10 için 0,472-0,502; %20 için 0,461-0,496; %30 için 0,390-0,493; %40 için 0,368-0,484; %50 için 0,286-0,456'dır. Değer atama yöntemlerinin kappa değerlerine ilişkin medyan değişim aralığı ise ortalama için 0,295-0,474; medyan için 0,296-0,472; rastgele için 0,286-0,476; KNN için 0,339-0,492; I-RF için 0,420-0,498; MICE-CART için 0,407-0,502; DURR için 0,452-0,496; IURR için 0,456-0,496'dır (Tablo 3 ve Şekil 8).

Tüm değer atama yöntemlerinin dengeli doğruluk oranı, AUC ve kappa değeri performansları değişen eksik oranına göre incelendiğinde %10 eksik oranında ortalama, medyan ve rastgele yöntemlerinin referans ile orta düzeyde yakın; diğer yöntemlerin ise referans ile daha yakın tahminlerde bulunduğu; %20 eksik oranından itibaren ortalama, medyan, rastgele, KNN ve I-RF yöntemlerinin performanslarının azalmaya başladığı, diğer yöntemlerin ise %30 eksik oranına kadar iyi performanslarını koruduğu görülmüştür. %30 ve %40 eksik oranlarında bir miktar performans kaybına rağmen DURR ve IURR yöntemleri ve bunları takiben MICE-CART ve I-RF yöntemlerinin performanslarının diğer yöntemlere göre daha iyi olduğu; %50 eksik oranında ise DURR ve IURR yöntemlerinin en iyi performans gösteren yöntemler olduğu belirlenmiştir.

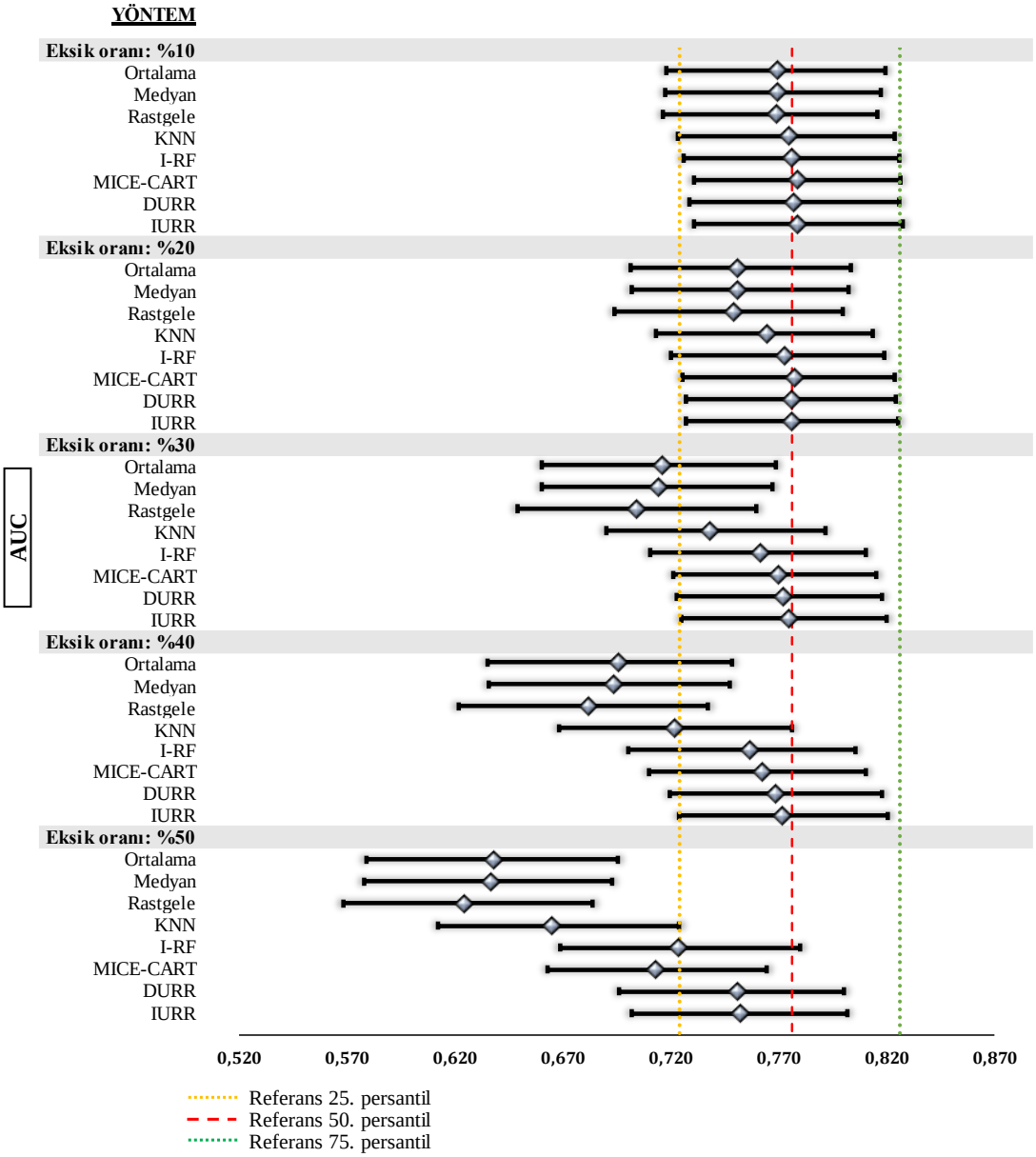
Dengeli doğruluk oranı, AUC ve kappa sonuçlarına göre uygulanan aşamalı kümeleme analizi ile elde edilen dendrogram grafikleri Şekil 9'da verilmiştir. Eksik oranı %10 için referans ile MICE-CART, KNN, I-RF, DURR ve IURR yöntemlerinin benzer performans göstererek aynı kümede yer aldıkları; ortalama, rastgele ve medyan yöntemlerinin de birbirine yakın performans göstererek ayrı bir küme oluşturdukları belirlenmiştir. Eksik oranı %20 için MICE-CART, IURR ve DURR yöntemlerinin benzer performans gösterdiği; bunun dışında iki ayrı küme oluşturan yöntemlerin ortalama, medyan ve rastgele; KNN ve I-RF olduğu belirlenmiştir. Eksik oranı %30, %40 ve %50 için I-RF, MICE-CART, DURR ve IURR yöntemlerinin referans ile aynı kümede yer aldıkları; ikinci kümeyi oluşturan yöntemlerin ortalama, medyan, rastgele ve KNN olduğu; buna ek olarak %40 ve %50 eksik oranları için DURR ve IURR yöntemlerinin referansa en yakın yöntemler olduğu belirlenmiştir (Şekil 9).

**Tablo 3.** -0,1 ≤ r ≤ 0,1 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri

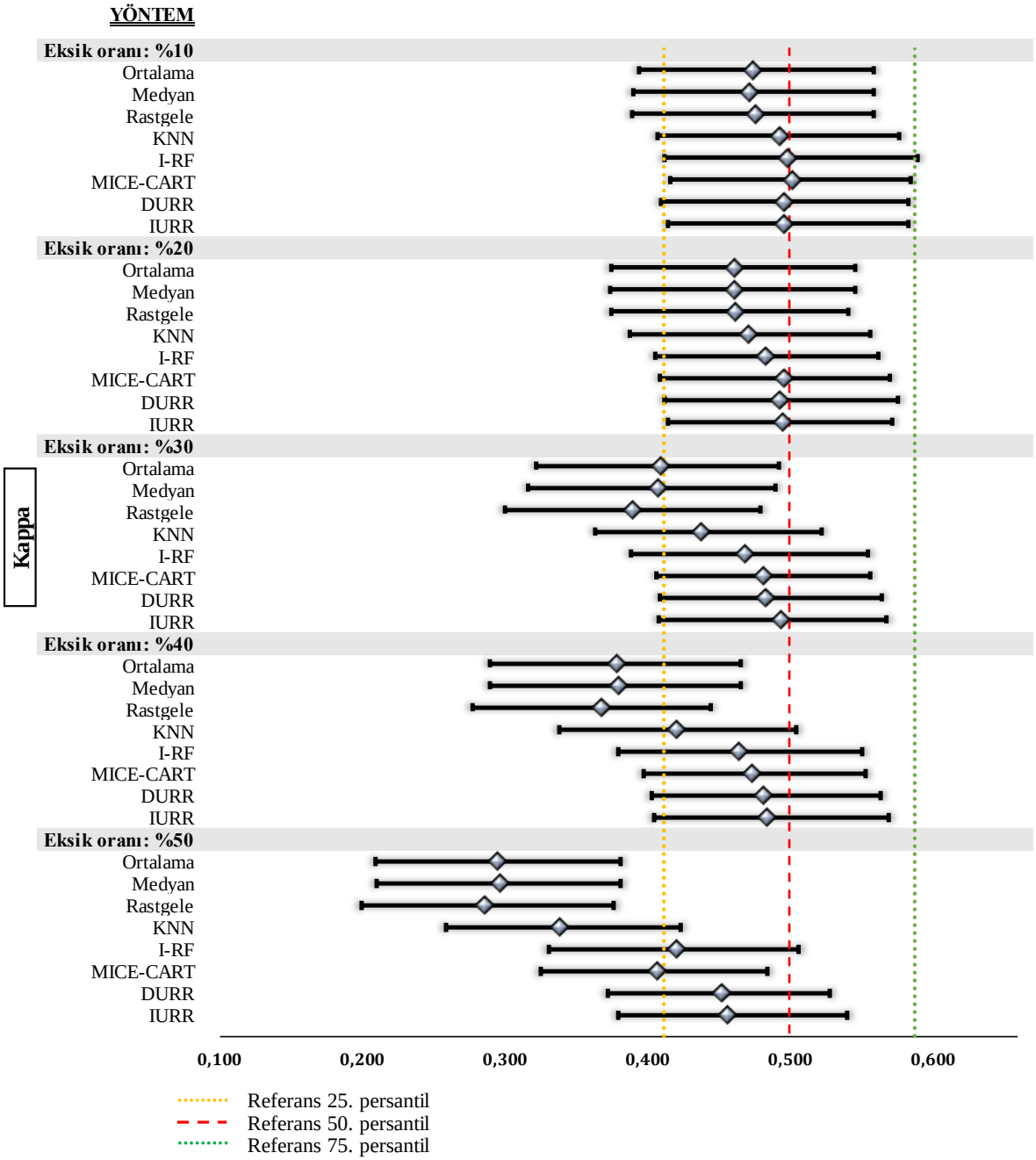
| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| **Dengeli Doğruluk** | **Referans** | 0,750 (0,702 - 0,790) | 0,750 (0,702 - 0,790) | 0,750 (0,702 - 0,790) | 0,750 (0,702 - 0,790) | 0,750 (0,702 - 0,790) |
| | **Ortalama** | 0,738 (0,693 - 0,781) | 0,729 (0,685 - 0,770) | 0,702 (0,658 - 0,744) | 0,689 (0,642 - 0,732) | 0,646 (0,600 - 0,690) |
| | **Medyan** | 0,736 (0,693 - 0,781) | 0,729 (0,685 - 0,771) | 0,702 (0,655 - 0,744) | 0,688 (0,642 - 0,731) | 0,647 (0,601 - 0,690) |
| | **Rastgele** | 0,737 (0,693 - 0,780) | 0,730 (0,684 - 0,768) | 0,695 (0,648 - 0,738) | 0,680 (0,636 - 0,720) | 0,640 (0,596 - 0,687) |
| | **KNN** | 0,744 (0,699 - 0,787) | 0,736 (0,692 - 0,779) | 0,717 (0,676 - 0,760) | 0,709 (0,667 - 0,752) | 0,669 (0,628 - 0,711) |
| | **I-RF** | 0,747 (0,702 - 0,793) | 0,744 (0,700 - 0,785) | 0,735 (0,693 - 0,778) | 0,731 (0,688 - 0,775) | 0,710 (0,664 - 0,752) |
| | **MICE-CART** | 0,750 (0,706 - 0,792) | 0,747 (0,703 - 0,784) | 0,740 (0,702 - 0,779) | 0,737 (0,695 - 0,776) | 0,702 (0,661 - 0,742) |
| | **DURR** | 0,747 (0,701 - 0,790) | 0,745 (0,704 - 0,787) | 0,741 (0,702 - 0,782) | 0,739 (0,698 - 0,780) | 0,725 (0,684 - 0,765) |
| | **IURR** | 0,746 (0,705 - 0,790) | 0,746 (0,704 - 0,785) | 0,745 (0,702 - 0,783) | 0,741 (0,701 - 0,783) | 0,727 (0,688 - 0,769) |
| **AUC** | **Referans** | 0,776 (0,724 - 0,826) | 0,776 (0,724 - 0,826) | 0,776 (0,724 - 0,826) | 0,776 (0,724 - 0,826) | 0,776 (0,724 - 0,826) |
| | **Ortalama** | 0,769 (0,718 - 0,819) | 0,750 (0,701 - 0,804) | 0,716 (0,660 - 0,769) | 0,696 (0,635 - 0,748) | 0,638 (0,579 - 0,696) |
| | **Medyan** | 0,770 (0,717 - 0,817) | 0,751 (0,702 - 0,802) | 0,714 (0,660 - 0,767) | 0,694 (0,635 - 0,747) | 0,636 (0,578 - 0,693) |
| | **Rastgele** | 0,769 (0,716 - 0,815) | 0,749 (0,694 - 0,800) | 0,704 (0,649 - 0,759) | 0,682 (0,622 - 0,737) | 0,624 (0,568 - 0,684) |
| | **KNN** | 0,775 (0,723 - 0,824) | 0,765 (0,713 - 0,814) | 0,738 (0,690 - 0,792) | 0,722 (0,668 - 0,776) | 0,665 (0,612 - 0,724) |
| | **I-RF** | 0,776 (0,726 - 0,826) | 0,773 (0,720 - 0,819) | 0,762 (0,710 - 0,810) | 0,756 (0,700 - 0,806) | 0,723 (0,669 - 0,780) |
| | **MICE-CART** | 0,778 (0,731 - 0,826) | 0,777 (0,726 - 0,824) | 0,770 (0,721 - 0,815) | 0,762 (0,710 - 0,810) | 0,713 (0,663 - 0,765) |
| | **DURR** | 0,777 (0,729 - 0,826) | 0,776 (0,727 - 0,824) | 0,772 (0,723 - 0,818) | 0,769 (0,719 - 0,818) | 0,750 (0,696 - 0,800) |
| | **IURR** | 0,778 (0,731 - 0,827) | 0,776 (0,727 - 0,825) | 0,775 (0,725 - 0,820) | 0,771 (0,724 - 0,820) | 0,752 (0,702 - 0,802) |
| **Kappa** | **Referans** | 0,500 (0,412 - 0,587) | 0,500 (0,412 - 0,587) | 0,500 (0,412 - 0,587) | 0,500 (0,412 - 0,587) | 0,500 (0,412 - 0,587) |
| | **Ortalama** | 0,474 (0,394 - 0,559) | 0,461 (0,374 - 0,545) | 0,410 (0,322 - 0,492) | 0,379 (0,290 - 0,465) | 0,295 (0,209 - 0,381) |
| | **Medyan** | 0,472 (0,391 - 0,559) | 0,461 (0,374 - 0,545) | 0,408 (0,316 - 0,490) | 0,380 (0,290 - 0,465) | 0,296 (0,211 - 0,381) |
| | **Rastgele** | 0,476 (0,389 - 0,559) | 0,462 (0,375 - 0,541) | 0,390 (0,301 - 0,479) | 0,368 (0,277 - 0,444) | 0,286 (0,200 - 0,376) |
| | **KNN** | 0,492 (0,407 - 0,576) | 0,471 (0,387 - 0,556) | 0,438 (0,364 - 0,522) | 0,420 (0,339 - 0,505) | 0,339 (0,259 - 0,423) |
| | **I-RF** | 0,498 (0,412 - 0,589) | 0,483 (0,405 - 0,562) | 0,469 (0,388 - 0,555) | 0,464 (0,380 - 0,551) | 0,420 (0,331 - 0,506) |
| | **MICE-CART** | 0,502 (0,416 - 0,585) | 0,496 (0,409 - 0,570) | 0,482 (0,407 - 0,556) | 0,473 (0,397 - 0,553) | 0,407 (0,326 - 0,485) |
| | **DURR** | 0,496 (0,410 - 0,583) | 0,493 (0,412 - 0,576) | 0,483 (0,409 - 0,565) | 0,481 (0,403 - 0,564) | 0,452 (0,372 - 0,528) |
| | **IURR** | 0,496 (0,415 - 0,583) | 0,495 (0,414 - 0,572) | 0,493 (0,408 - 0,568) | 0,484 (0,405 - 0,569) | 0,456 (0,380 - 0,540) |



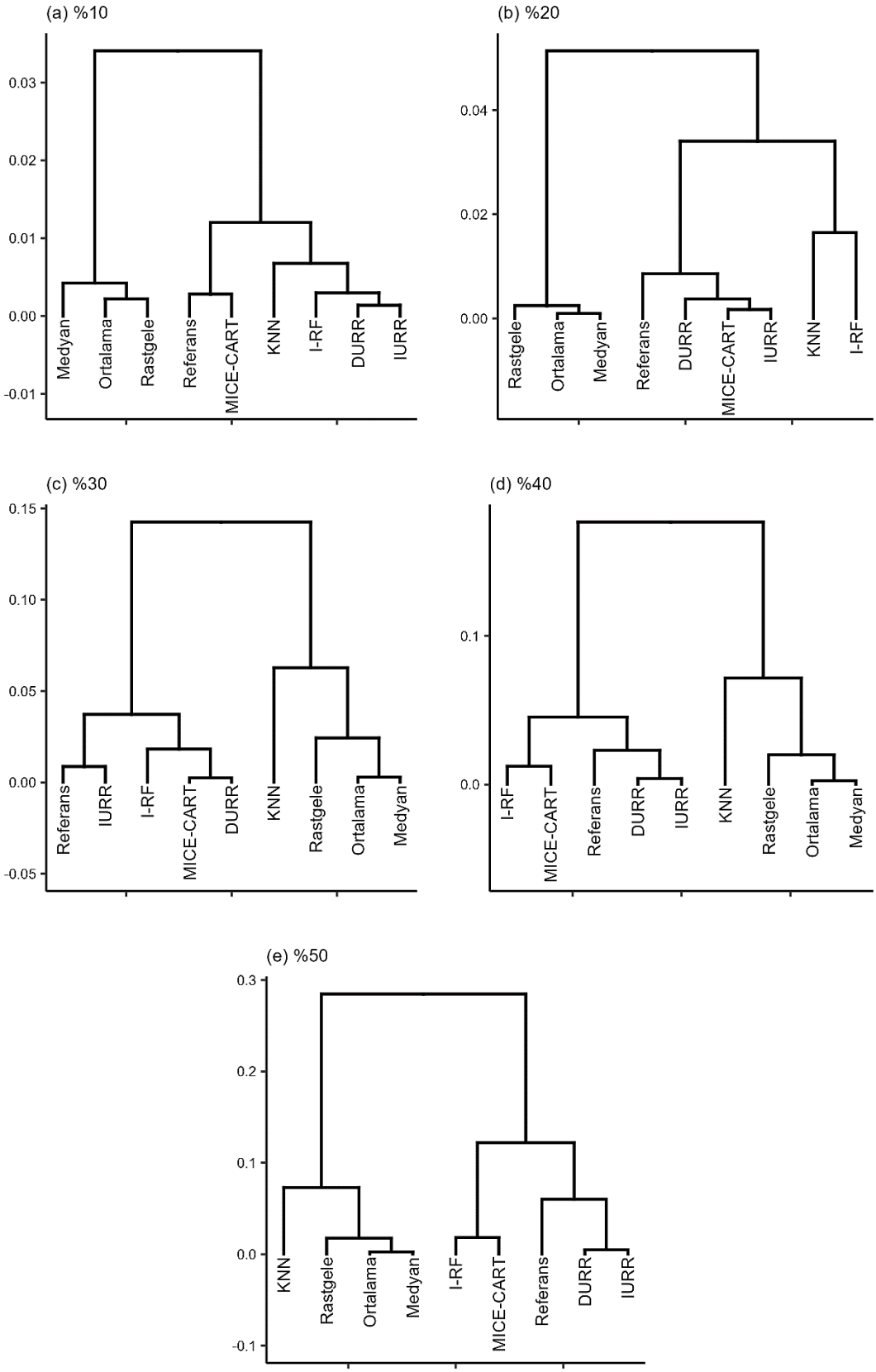

**Şekil 6.** -0,1 ≤ r ≤ 0,1 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranlarının orman grafiği




**Şekil 7**. -0,1 ≤ r ≤ 0,1 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin AUC değerlerinin orman grafiği




**Şekil 8.** -0,1 ≤ r ≤ 0,1 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin kappa değerlerinin orman grafiği




**Şekil 9.** -0,1 ≤ r ≤ 0,1 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için farklı eksik oranlarında yöntemlerin dendrogram grafikleri

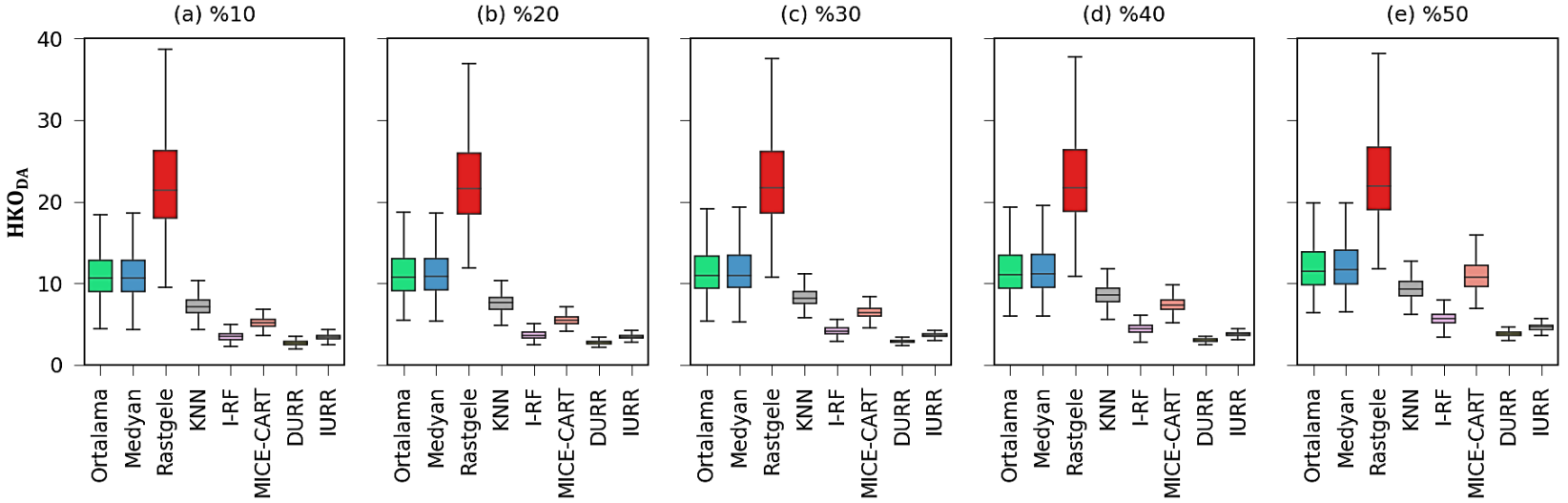
### 4.1.2. -0,5 ≤ r ≤ 0,5 Aralığına göre Rastgele Bir Değişken Setinin Doğrusal Kombinasyonundan Türetilen Eksik Verili Değişkenler için Bulgular

Eksik oranı %10, %20, %30, %40 ve %50 için $HKO_{DA}$ değerlerinin medyan değişim aralığı sırasıyla 2,701-21,392; 2,741-21,582; 2,887-21,702; 2,999-21,743; 3,777-21,943'dür. Ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR yöntemlerinin $HKO_{DA}$ değerlerinin ilişkin ilişkin medyan değişim aralığı ise sırasıyla 10,618-11,493; 10,683-11,652; 21,392-21,943; 7,159-9,259; 3,479-5,626; 5,166-10,709; 2,701-3,777; 3,384-4,596'dır (Tablo 4).

Tüm değer atama yöntemlerinin $HKO_{DA}$ değerleri değişen eksik oranına göre incelendiğinde eksik oranındaki artışın genel olarak yöntemlerin $HKO_{DA}$ değerlerini arttırdığı; I-RF, DURR ve IURR yöntemlerinin tüm eksik oranlarında diğer yöntemlere göre daha düşük $HKO_{DA}$ değerine sahip olduğu; en yüksek $HKO_{DA}$ değerine sahip olan yöntemin ise rastgele değer atama yöntemi olduğu belirlenmiştir (Tablo 4 ve Şekil 10).

**Tablo 4.** -0,5 ≤ r ≤ 0,5 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerleri

| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| **$HKO_{DA}$** | **Ortalama** | 10,618 (09,024 - 12,821) | 10,740 (09,124 - 12,985) | 10,930 (09,414 - 13,310) | 11,056 (09,398 - 13,399) | 11,493 (09,821 - 13,854) |
| | **Medyan** | 10,683 (09,009 - 12,853) | 10,798 (09,183 - 12,999) | 10,991 (09,447 - 13,440) | 11,177 (09,501 - 13,538) | 11,652 (09,945 - 13,998) |
| | **Rastgele** | 21,392 (18,012 - 26,307) | 21,582 (18,472 - 25,929) | 21,702 (18,591 - 26,167) | 21,743 (18,807 - 26,412) | 21,943 (18,969 - 26,648) |
| | **KNN** | 7,159 (6,394 - 7,986) | 7,582 (6,803 - 8,248) | 8,172 (7,498 - 8,975) | 8,565 (7,780 - 9,394) | 9,259 (8,442 - 10,184) |
| | **I-RF** | 3,479 (3,098 - 3,846) | 3,635 (3,328 - 4,011) | 4,131 (3,794 - 4,520) | 4,462 (4,043 - 4,886) | 5,626 (5,094 - 6,226) |
| | **MICE-CART** | 5,166 (4,753 - 5,585) | 5,453 (5,059 - 5,896) | 6,403 (5,933 - 6,922) | 7,299 (6,754 - 7,980) | 10,709 (9,564 - 12,150) |
| | **DURR** | 2,701 (2,491 - 2,903) | 2,741 (2,584 - 2,904) | 2,887 (2,751 - 3,016) | 2,999 (2,872 - 3,142) | 3,777 (3,557 - 3,990) |
| | **IURR** | 3,384 (3,132 - 3,636) | 3,431 (3,245 - 3,623) | 3,593 (3,449 - 3,773) | 3,752 (3,588 - 3,912) | 4,596 (4,346 - 4,863) |




**Şekil 10.** -0,5 ≤ r ≤ 0,5 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği

Referans veri setinden elde edilen dengeli doğruluk oranı değerlerinin medyanı 0,774 olarak elde edilmiştir. Atanmış veri setinden elde edilen sonuçlara göre dengeli doğruluk oranı değerlerinin medyan değişim aralığı eksik oranı %10 için 0,765-0,775; %20 için 0,755-0,773; %30 için 0,735-0,771; %40 için 0,715-0,770; %50 için 0,690-0,757'dir. Değer atama yöntemlerinin dengeli doğruluk oranı değerlerine ilişkin medyan değişim aralığı ise ortalama için 0,696-0,771; medyan için 0,695-0,770; rastgele için 0,690-0,765; KNN için 0,714-0,772; I-RF için 0,738-0,775; MICE-CART için 0,738-0,774; DURR için 0,755-0,775; IURR için 0,757-0,773'dür (Tablo 5 ve Şekil 11).

Referans veri setinden elde edilen AUC değerlerinin medyanı 0,813 olarak elde edilmiştir. Atanmış veri setinden elde edilen sonuçlara göre eksik oranı %10, %20, %30, %40 ve %50 için AUC değerlerinin medyan değişim aralığı sırasıyla 0,807-0,816; 0,794-0,815; 0,760-0,810; 0,740-0,808; 0,699-0,793'dür. Değer atama yöntemlerinin AUC değerlerine ilişkin medyan değişim aralığı ise ortalama için 0,706-0,808; medyan için 0,708-0,807; rastgele için 0,699-0,807; KNN için 0,732-0,810; I-RF için 0,767-0,813; MICE-CART için 0,766-0,816; DURR için 0,792-0,814; IURR için 0,793-0,814'dür (Tablo 5 ve Şekil 12).

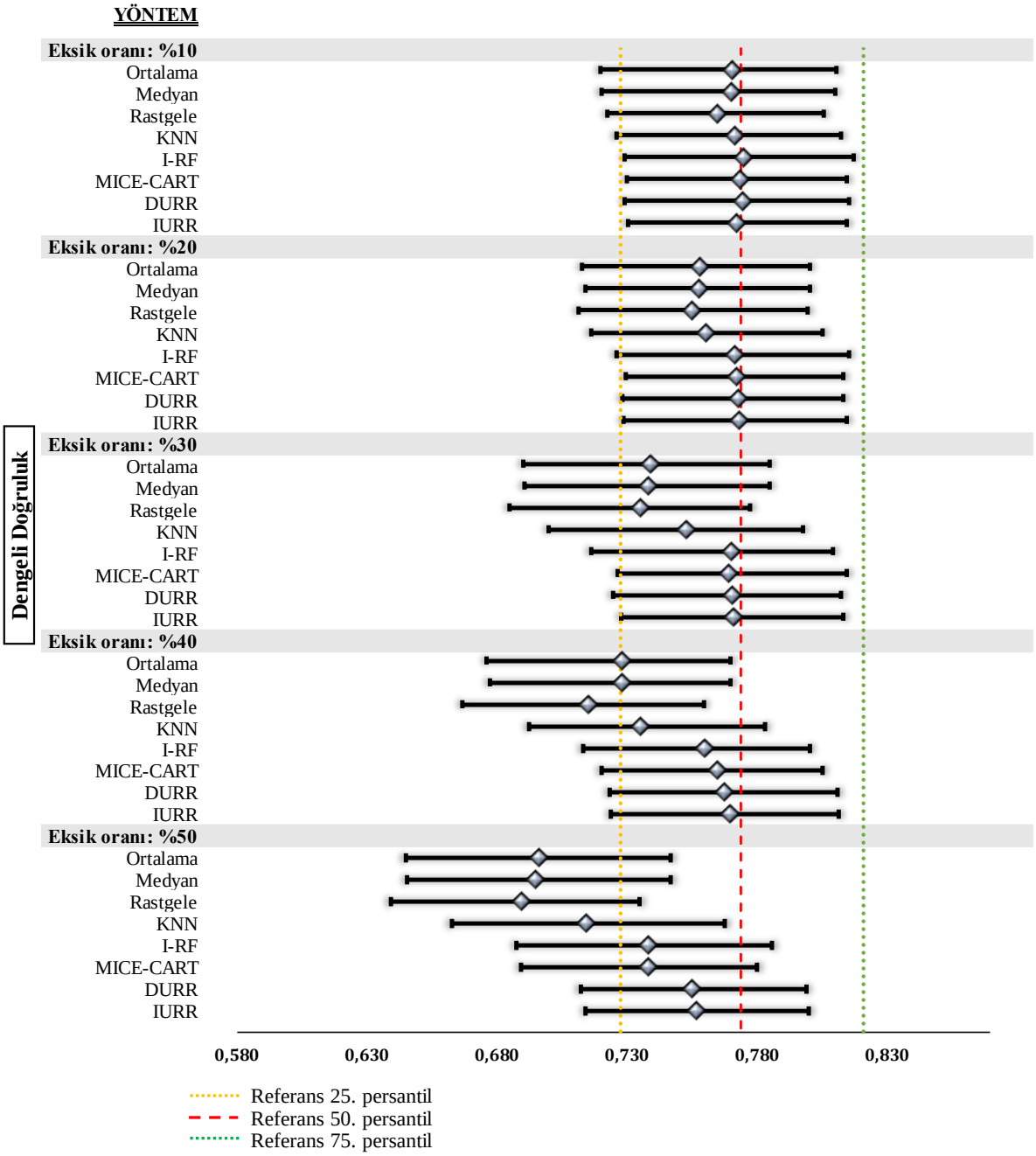
Referans veri setinden elde edilen kappa değerlerinin medyanı 0,551 olarak elde edilmiştir. Atanmış veri setinden elde edilen sonuçlara göre eksik oranı %10, %20, %30, %40 ve %50 için kappa değerlerinin medyan değişim aralığı sırasıyla 0,537-0,553; 0,510-0,551; 0,471-0,546; 0,436-0,542; 0,380-0,516'dır. Değer atama yöntemlerinin kappa değerlerine ilişkin medyan değişim aralığı ise ortalama için 0,394-0,548; medyan için 0,396-0,545; rastgele için 0,380-0,537; KNN için 0,426-0,548; I-RF için 0,474-0,552; MICE-CART için 0,480-0,553; DURR için 0,512-0,551; IURR için 0,516-0,550'dir (Tablo 5 ve Şekil 13).

Tüm değer atama yöntemlerinin dengeli doğruluk oranı, AUC ve kappa değeri performansları değişen eksik oranına göre incelendiğinde %10 eksik oranında I-RF, MICE-CART, DURR, IURR ve bunları takiben KNN yöntemlerinin diğer yöntemlere göre referans ile daha yakın tahminlerde bulunduğu; %20 eksik oranından itibaren yöntemler arası performans farklılıklarının belirginleştiği ve ortalama, medyan, rastgele, KNN yöntemlerinin performansının düşmeye başladığı, diğer yöntemlerin ise istikrarını koruduğu görülmüştür. %30 ve %40 eksik oranlarında I-RF, MICE-CART, DURR ve IURR yöntemlerinin iyi performans göstermeye devam ettiği; %50 eksik oranında ise tüm yöntemlerin performansları azalırken DURR ve IURR yöntemlerinin referansı tahmin uyumları en yüksek olan yöntemler olduğu belirlenmiştir.

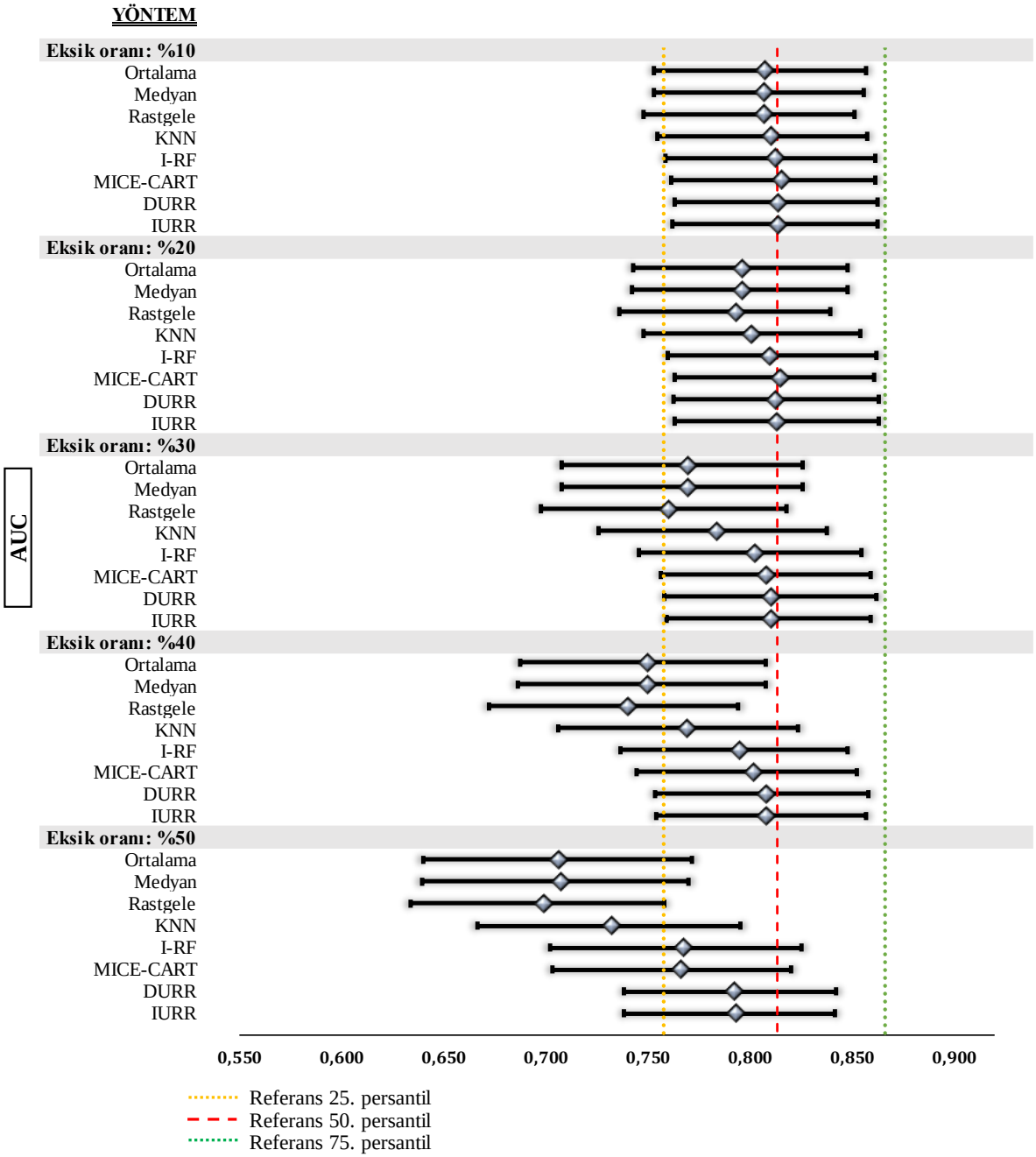
Dengeli doğruluk oranı, AUC ve kappa sonuçlarına göre uygulanan aşamalı kümeleme analizi ile elde edilen dendrogram grafikleri Şekil 14'de verilmiştir. Buna göre tüm eksik oranları için referans ile I-RF, MICE-CART, DURR ve IURR yöntemlerinin benzer performans göstererek aynı kümede yer aldıkları görülmüştür. Ayrıca eksik oranı %10 için ikinci kümeyi oluşturan yöntemlerin ortalama, KNN ve medyan yöntemleri olduğu; rastgele yönteminin ise tüm yöntemlerden ayrılarak üçüncü kümeyi oluşturduğu; diğer eksik oranlarında ise ortalama, medyan, rastgele ve bu yöntemleri takiben KNN yönteminin birbirine yakın performans gösterip aynı kümede yer alan yöntemler olduğu belirlenmiştir. Buna ek olarak %50 eksik oranı için DURR ve IURR yöntemlerinin referansa en yakın yöntemler olduğu bulgusuna ulaşılmıştır (Şekil 14).

**Tablo 5**. -0,5 ≤ r ≤ 0,5 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri

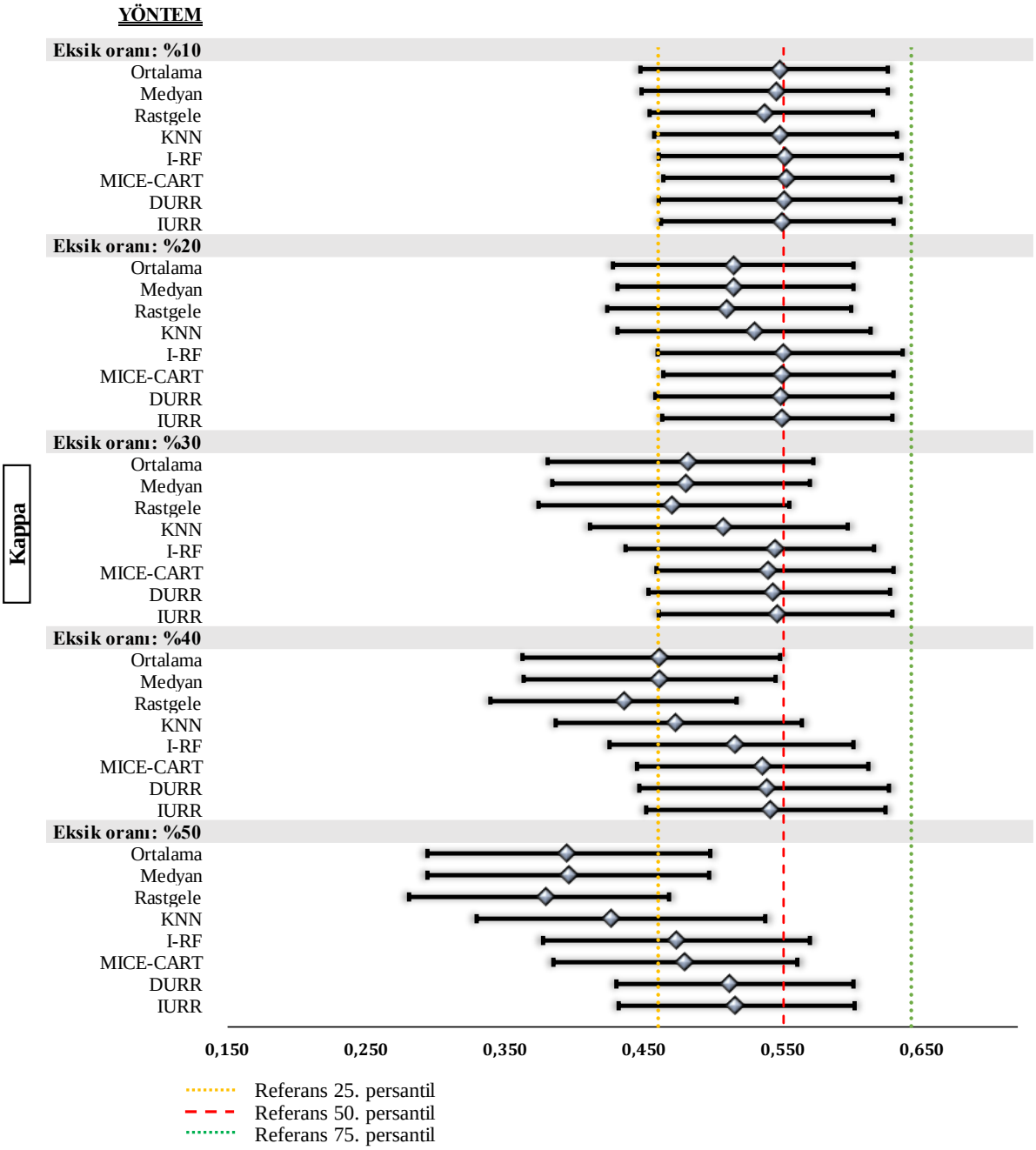
| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| **Dengeli Doğruluk** | **Referans** | 0,774 (0,728 - 0,821) | 0,774 (0,728 - 0,821) | 0,774 (0,728 - 0,821) | 0,774 (0,728 - 0,821) | 0,774 (0,728 - 0,821) |
| | **Ortalama** | 0,771 (0,720 - 0,811) | 0,758 (0,713 - 0,801) | 0,739 (0,690 - 0,785) | 0,728 (0,676 - 0,770) | 0,696 (0,645 - 0,747) |
| | **Medyan** | 0,770 (0,720 - 0,810) | 0,758 (0,714 - 0,801) | 0,738 (0,691 - 0,785) | 0,728 (0,677 - 0,770) | 0,695 (0,645 - 0,747) |
| | **Rastgele** | 0,765 (0,723 - 0,806) | 0,755 (0,711 - 0,800) | 0,735 (0,685 - 0,778) | 0,715 (0,667 - 0,760) | 0,690 (0,639 - 0,735) |
| | **KNN** | 0,772 (0,726 - 0,813) | 0,760 (0,716 - 0,806) | 0,753 (0,700 - 0,798) | 0,735 (0,692 - 0,783) | 0,714 (0,663 - 0,768) |
| | **I-RF** | 0,775 (0,729 - 0,817) | 0,772 (0,726 - 0,816) | 0,770 (0,716 - 0,810) | 0,760 (0,713 - 0,801) | 0,738 (0,688 - 0,786) |
| | **MICE-CART** | 0,774 (0,730 - 0,815) | 0,773 (0,730 - 0,813) | 0,769 (0,727 - 0,815) | 0,765 (0,720 - 0,805) | 0,738 (0,689 - 0,780) |
| | **DURR** | 0,775 (0,729 - 0,816) | 0,773 (0,728 - 0,813) | 0,771 (0,725 - 0,812) | 0,768 (0,723 - 0,811) | 0,755 (0,712 - 0,799) |
| | **IURR** | 0,772 (0,730 - 0,815) | 0,773 (0,729 - 0,814) | 0,771 (0,728 - 0,813) | 0,770 (0,724 - 0,812) | 0,757 (0,714 - 0,800) |
| **AUC** | **Referans** | 0,813 (0,758 - 0,866) | 0,813 (0,758 - 0,866) | 0,813 (0,758 - 0,866) | 0,813 (0,758 - 0,866) | 0,813 (0,758 - 0,866) |
| | **Ortalama** | 0,808 (0,753 - 0,857) | 0,796 (0,743 - 0,848) | 0,770 (0,708 - 0,826) | 0,750 (0,688 - 0,808) | 0,706 (0,640 - 0,772) |
| | **Medyan** | 0,807 (0,753 - 0,856) | 0,796 (0,742 - 0,848) | 0,770 (0,708 - 0,826) | 0,750 (0,687 - 0,808) | 0,708 (0,639 - 0,770) |
| | **Rastgele** | 0,807 (0,748 - 0,851) | 0,794 (0,736 - 0,839) | 0,760 (0,698 - 0,818) | 0,740 (0,672 - 0,794) | 0,699 (0,634 - 0,758) |
| | **KNN** | 0,810 (0,754 - 0,858) | 0,801 (0,748 - 0,854) | 0,784 (0,726 - 0,838) | 0,769 (0,706 - 0,823) | 0,732 (0,667 - 0,796) |
| | **I-RF** | 0,813 (0,759 - 0,862) | 0,810 (0,760 - 0,862) | 0,802 (0,746 - 0,854) | 0,795 (0,737 - 0,848) | 0,767 (0,702 - 0,825) |
| | **MICE-CART** | 0,816 (0,761 - 0,861) | 0,815 (0,763 - 0,861) | 0,808 (0,757 - 0,859) | 0,802 (0,745 - 0,853) | 0,766 (0,703 - 0,820) |
| | **DURR** | 0,814 (0,763 - 0,863) | 0,813 (0,762 - 0,863) | 0,810 (0,758 - 0,862) | 0,808 (0,754 - 0,858) | 0,792 (0,738 - 0,842) |
| | **IURR** | 0,814 (0,762 - 0,862) | 0,813 (0,763 - 0,863) | 0,810 (0,759 - 0,859) | 0,808 (0,754 - 0,857) | 0,793 (0,738 - 0,841) |
| **Kappa** | **Referans** | 0,551 (0,461 - 0,643) | 0,551 (0,461 - 0,643) | 0,551 (0,461 - 0,643) | 0,551 (0,461 - 0,643) | 0,551 (0,461 - 0,643) |
| | **Ortalama** | 0,548 (0,448 - 0,626) | 0,515 (0,428 - 0,601) | 0,482 (0,381 - 0,572) | 0,461 (0,363 - 0,548) | 0,394 (0,294 - 0,498) |
| | **Medyan** | 0,545 (0,449 - 0,626) | 0,515 (0,431 - 0,601) | 0,481 (0,384 - 0,570) | 0,461 (0,364 - 0,545) | 0,396 (0,294 - 0,497) |
| | **Rastgele** | 0,537 (0,454 - 0,615) | 0,510 (0,424 - 0,599) | 0,471 (0,374 - 0,555) | 0,436 (0,339 - 0,517) | 0,380 (0,281 - 0,469) |
| | **KNN** | 0,548 (0,458 - 0,633) | 0,530 (0,432 - 0,614) | 0,507 (0,412 - 0,597) | 0,473 (0,387 - 0,564) | 0,426 (0,330 - 0,537) |
| | **I-RF** | 0,552 (0,461 - 0,636) | 0,551 (0,460 - 0,636) | 0,545 (0,437 - 0,616) | 0,516 (0,426 - 0,601) | 0,474 (0,377 - 0,570) |
| | **MICE-CART** | 0,553 (0,464 - 0,629) | 0,549 (0,464 - 0,630) | 0,540 (0,459 - 0,630) | 0,536 (0,445 - 0,612) | 0,480 (0,385 - 0,561) |
| | **DURR** | 0,551 (0,461 - 0,635) | 0,549 (0,459 - 0,630) | 0,543 (0,453 - 0,628) | 0,539 (0,447 - 0,627) | 0,512 (0,430 - 0,601) |
| | **IURR** | 0,550 (0,463 - 0,630) | 0,550 (0,463 - 0,629) | 0,546 (0,461 - 0,629) | 0,542 (0,452 - 0,624) | 0,516 (0,432 - 0,602) |



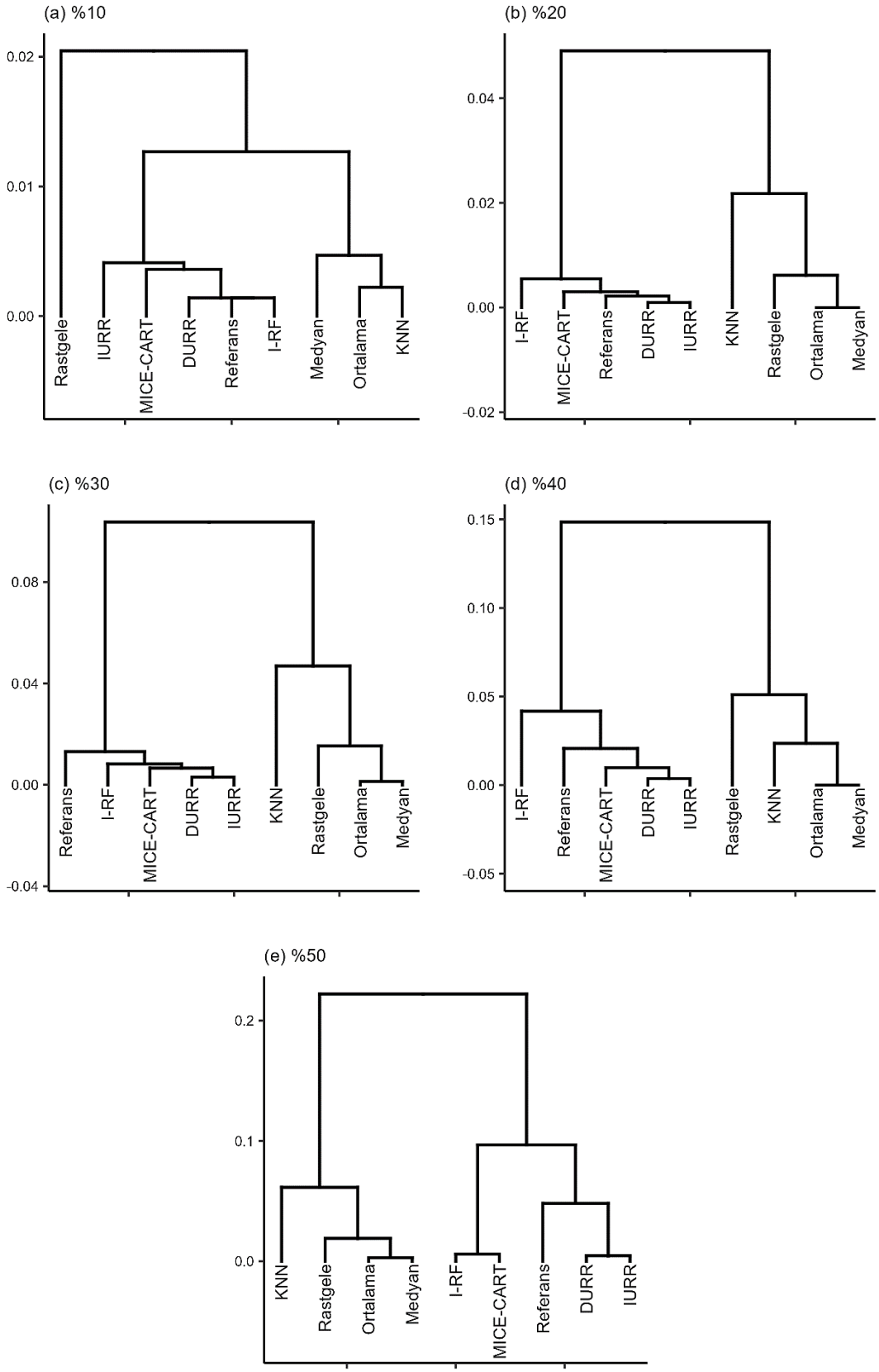

**Şekil 11.** -0,5 ≤ r ≤ 0,5 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranlarının orman grafiği




**Şekil 12.** -0,5 ≤ r ≤ 0,5 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin AUC değerlerinin orman grafiği




**Şekil 13.** -0,5 ≤ r ≤ 0,5 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin kappa değerlerinin orman grafiği




**Şekil 14.** -0,5 ≤ r ≤ 0,5 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için farklı eksik oranlarında yöntemlerin dendrogram grafikleri

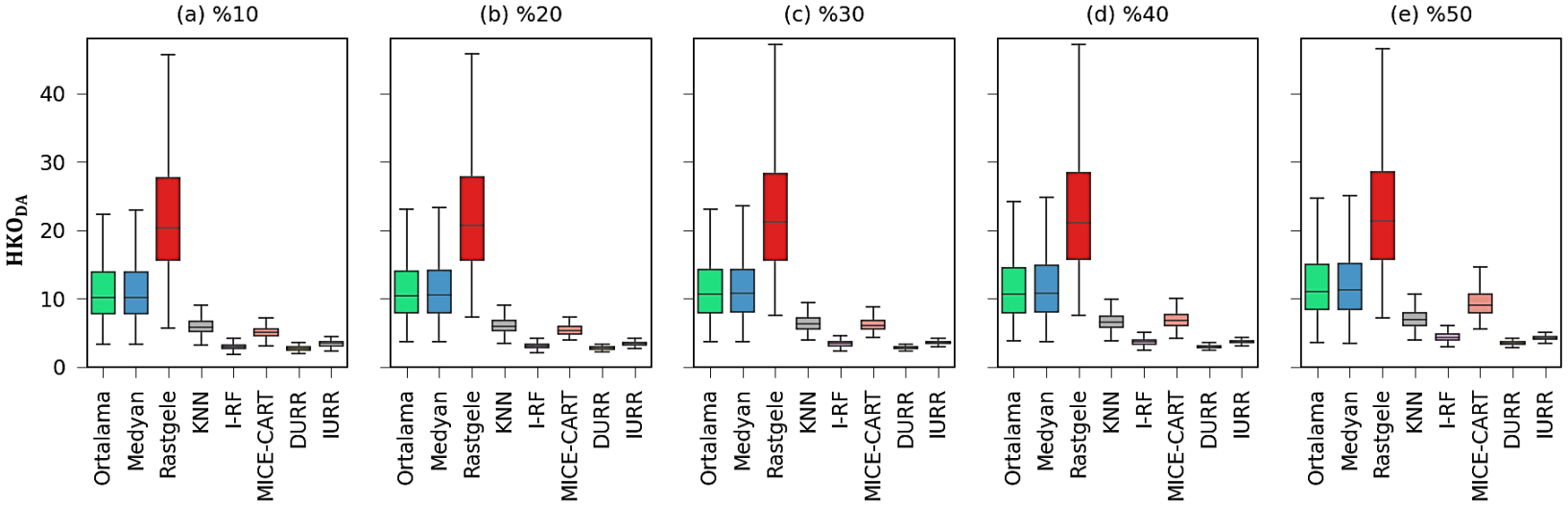
### 4.1.3. -0,8 ≤ r ≤ 0,8 Aralığına göre Rastgele Bir Değişken Setinin Doğrusal Kombinasyonundan Türetilen Eksik Verili Değişkenler için Bulgular

$HKO_{DA}$ değerlerinin medyan değişim aralığı eksik oranı %10 için 2,704-20,396; %20 için 2,773-20,744; %30 için 2,870-21,153; %40 için 2,994-21,065; %50 için 3,459-21,310'dur. Ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR yöntemlerinin $HKO_{DA}$ değerlerinin ilişkin ilişkin medyan değişim aralığı ise sırasıyla 10,149-11,072; 10,211-11,276; 20,396-21,310; 5,838-6,914; 2,914-4,320; 5,058-9,006; 2,704-3,459; 3,384-4,185'tir (Tablo 6).

Tüm değer atama yöntemlerinin $HKO_{DA}$ değerleri farklı eksik oranlarına göre incelendiğinde genel olarak eksik oranı arttıkça yöntemlerin $HKO_{DA}$ değerlerinin artış eğiliminde olduğu görülmüştür. Tüm eksik oranları için en düşük $HKO_{DA}$ değerine sahip yöntemin DURR; en yüksek $HKO_{DA}$ değerine sahip yöntemin ise rastgele değer atama yöntemi olduğu belirlenmiştir (Tablo 6 ve Şekil 15).

**Tablo 6.** -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerleri

| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| $HKO_{DA}$ | **Ortalama** | 10,149 (07,825 - 13,854) | 10,424 (07,873 - 13,987) | 10,659 (07,866 - 14,201) | 10,673 (07,921 - 14,557) | 11,072 (08,371 - 14,950) |
| | **Medyan** | 10,211 (07,825 - 13,887) | 10,467 (07,904 - 14,134) | 10,730 (07,975 - 14,281) | 10,731 (08,041 - 14,821) | 11,276 (08,404 - 15,096) |
| | **Rastgele** | 20,396 (15,663 - 27,656) | 20,744 (15,631 - 27,831) | 21,153 (15,566 - 28,240) | 21,065 (15,758 - 28,357) | 21,310 (15,770 - 28,494) |
| | **KNN** | 5,838 (5,137 - 6,699) | 5,986 (5,343 - 6,830) | 6,306 (5,598 - 7,148) | 6,489 (5,750 - 7,423) | 6,914 (6,073 - 7,963) |
| | **I-RF** | 2,914 (2,661 - 3,252) | 3,066 (2,777 - 3,349) | 3,380 (3,101 - 3,704) | 3,633 (3,313 - 3,997) | 4,320 (3,943 - 4,786) |
| | **MICE-CART** | 5,058 (4,549 - 5,582) | 5,280 (4,842 - 5,864) | 6,033 (5,521 - 6,847) | 6,773 (6,110 - 7,694) | 9,006 (7,877 - 10,617) |
| | **DURR** | 2,704 (2,506 - 2,924) | 2,773 (2,606 - 2,932) | 2,870 (2,741 - 2,998) | 2,994 (2,865 - 3,129) | 3,459 (3,287 - 3,661) |
| | **IURR** | 3,384 (3,121 - 3,646) | 3,432 (3,246 - 3,613) | 3,551 (3,411 - 3,717) | 3,688 (3,540 - 3,864) | 4,185 (4,004 - 4,410) |




**Şekil 15**. -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği

Referans veri setinden elde edilen dengeli doğruluk oranı değerlerinin medyanı 0,824 olarak elde edilmiştir. Atanmış veri setinden elde edilen sonuçlara göre eksik oranı %10, %20, %30, %40 ve %50 için dengeli doğruluk oranı değerlerinin medyan değişim aralığı sırasıyla 0,815-0,821; 0,810-0,824; 0,791-0,820; 0,775-0,817; 0,753-0,809'dur. Değer atama yöntemlerinin dengeli doğruluk oranı değerlerine ilişkin medyan değişim aralığı ise ortalama ve medyan için 0,768-0,815; rastgele için 0,753-0,816; KNN için 0,785-0,820; I-RF için 0,800-0,822; MICE-CART için 0,797-0,824; DURR için 0,807-0,821; IURR için 0,809-0,822'dir (Tablo 7 ve Şekil 16).

Referans veri setinden elde edilen AUC değerlerinin medyanı 0,877 olarak elde edilmiştir. Atanmış veri setinden elde edilen sonuçlara göre AUC değerlerinin medyan değişim aralığı eksik oranı %10 için 0,866-0,872; %20 için 0,860-0,876; %30 için 0,837-0,872; %40 için 0,820-0,871; %50 için 0,787-0,861'dir. Ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR yöntemlerinin AUC değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,808-0,870; 0,807-0,870; 0,787-0,866; 0,832-0,871; 0,848-0,874; 0,842-0,876; 0,861-0,874; 0,857-0,874'tür (Tablo 7 ve Şekil 17).

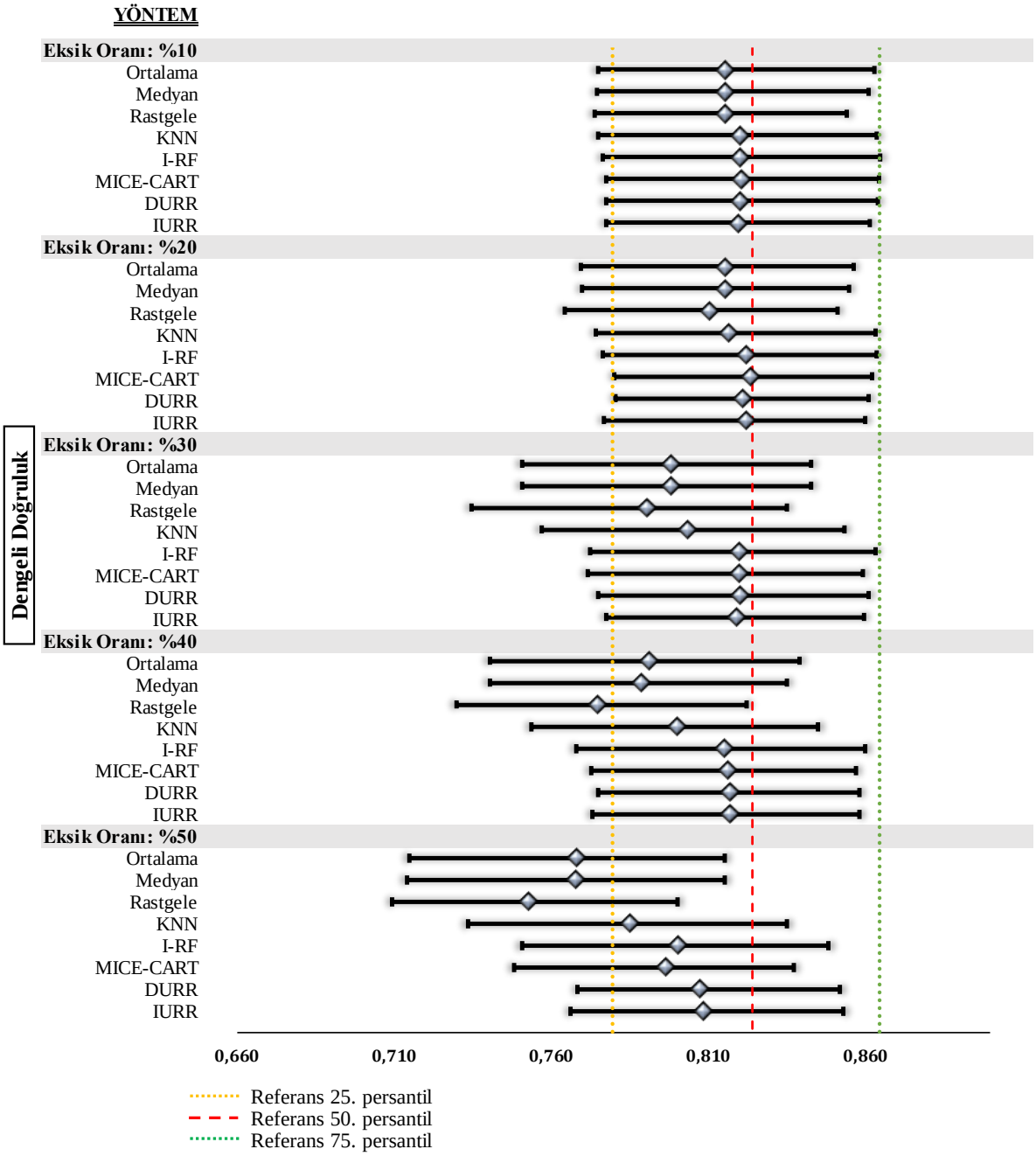
Referans veri setinden elde edilen kappa değerlerinin medyanı 0,645 olarak elde edilmiştir. Atanmış veri setinden elde edilen sonuçlara göre kappa değerlerinin medyan değişim aralığı eksik oranı %10 için 0,636-0,643; %20 için 0,629-0,650; %30 için 0,589-0,642; %40 için 0,554-0,640; %50 için 0,510-0,618'dir. Değer atama yöntemlerinin kappa değerlerine ilişkin medyan değişim aralığı ise ortalama için 0,545-0,637; medyan için 0,543-0,639; rastgele için 0,510-0,636; KNN için 0,571-0,643; I-RF için 0,601-0,644; MICE-CART için 0,594-0,650; DURR için 0,617-0,643; IURR için 0,618-0,644'tür (Tablo 7 ve Şekil 18).

Tüm değer atama yöntemlerinin dengeli doğruluk oranı, AUC ve kappa değeri performansları değişen eksik oranına göre incelendiğinde; %10 eksik oranında en kötü performansı rastgele değer atama yöntemi gösterirken, KNN, I-RF, MICE-CART, DURR ve IURR yöntemlerinin performanslarının diğer yöntemlere göre daha iyi olduğu; %20 eksik oranından itibaren ortalama, medyan, rastgele ve KNN yöntemlerinin performanslarının azalmaya başladığı gözlenmiştir. Eksik oranı %30'un üzerine çıktığında ortalama, medyan, rastgele ve KNN yöntemlerinin büyük bir performans kaybı yaşadığı; %40 eksik oranında I-RF, MICE-CART, DURR ve IURR yöntemlerinin; %50 eksik oranında ise DURR ve IURR yöntemlerini takiben I-RF ve MICE-CART yöntemlerinin referansı tahmin uyumlarının diğer yöntemlere göre daha iyi olduğu belirlenmiştir.

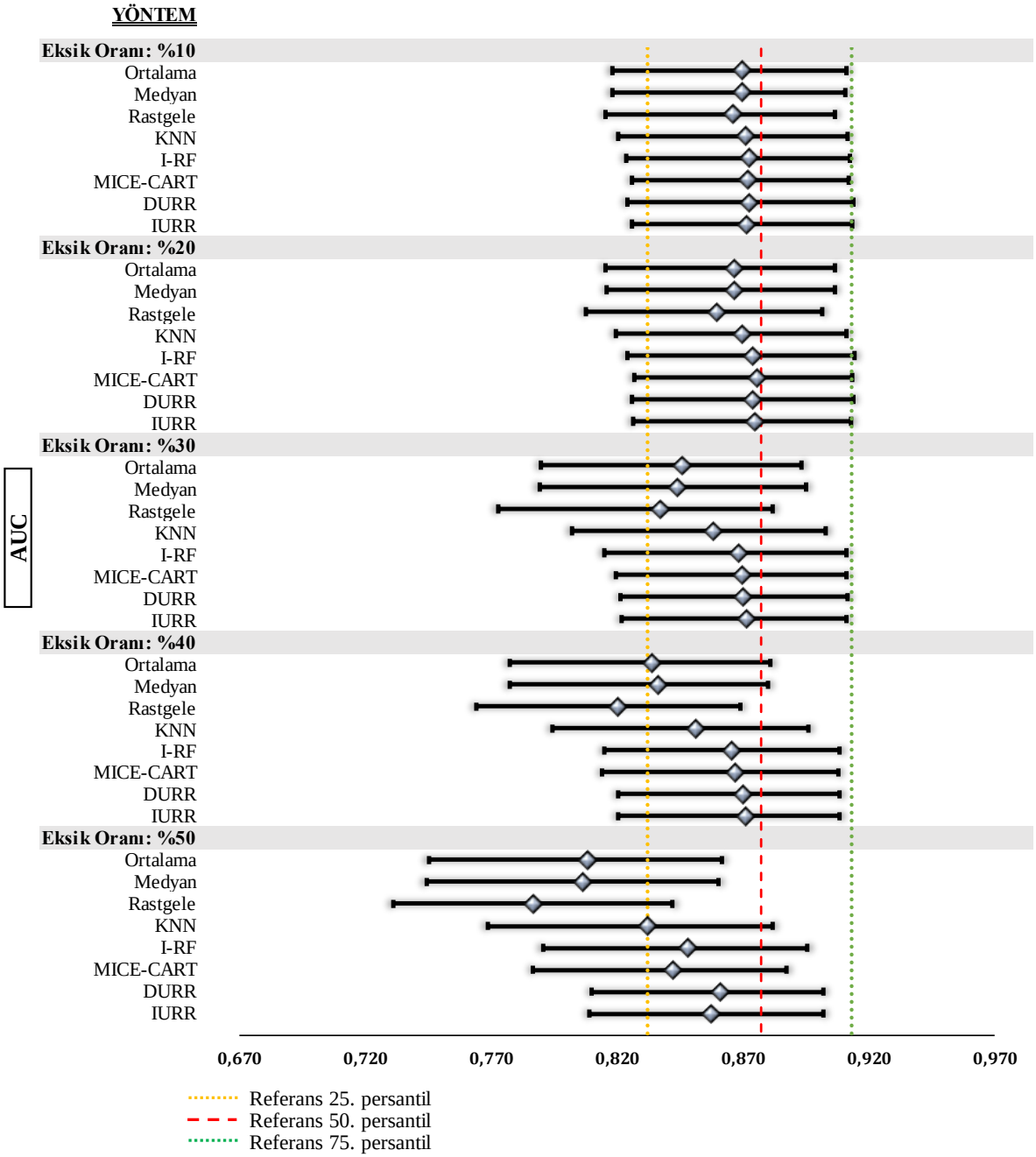
Dengeli doğruluk oranı, AUC ve kappa sonuçlarına göre uygulanan aşamalı kümeleme analizi ile elde edilen dendrogram grafikleri Şekil 19'da verilmiştir. Buna göre eksik oranı %10 için DURR, IURR, MICE-CART, I-RF ve KNN yöntemlerinin; eksik oranı %20, %30, %40 ve %50 için DURR, IURR, MICE-CART ve I-RF yöntemlerinin referans ile aynı kümede yer aldıkları görülmüştür. Ayrıca eksik oranı %10 için rastgele, ortalama ve medyan yöntemlerinin; eksik oranı %20, %30, %40 ve %50 için rastgele, ortalama, medyan ve KNN yöntemlerinin benzer performans gösterip ayrı bir küme oluşturdukları belirlenmiştir (Şekil 19).

**Tablo 7**. -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri

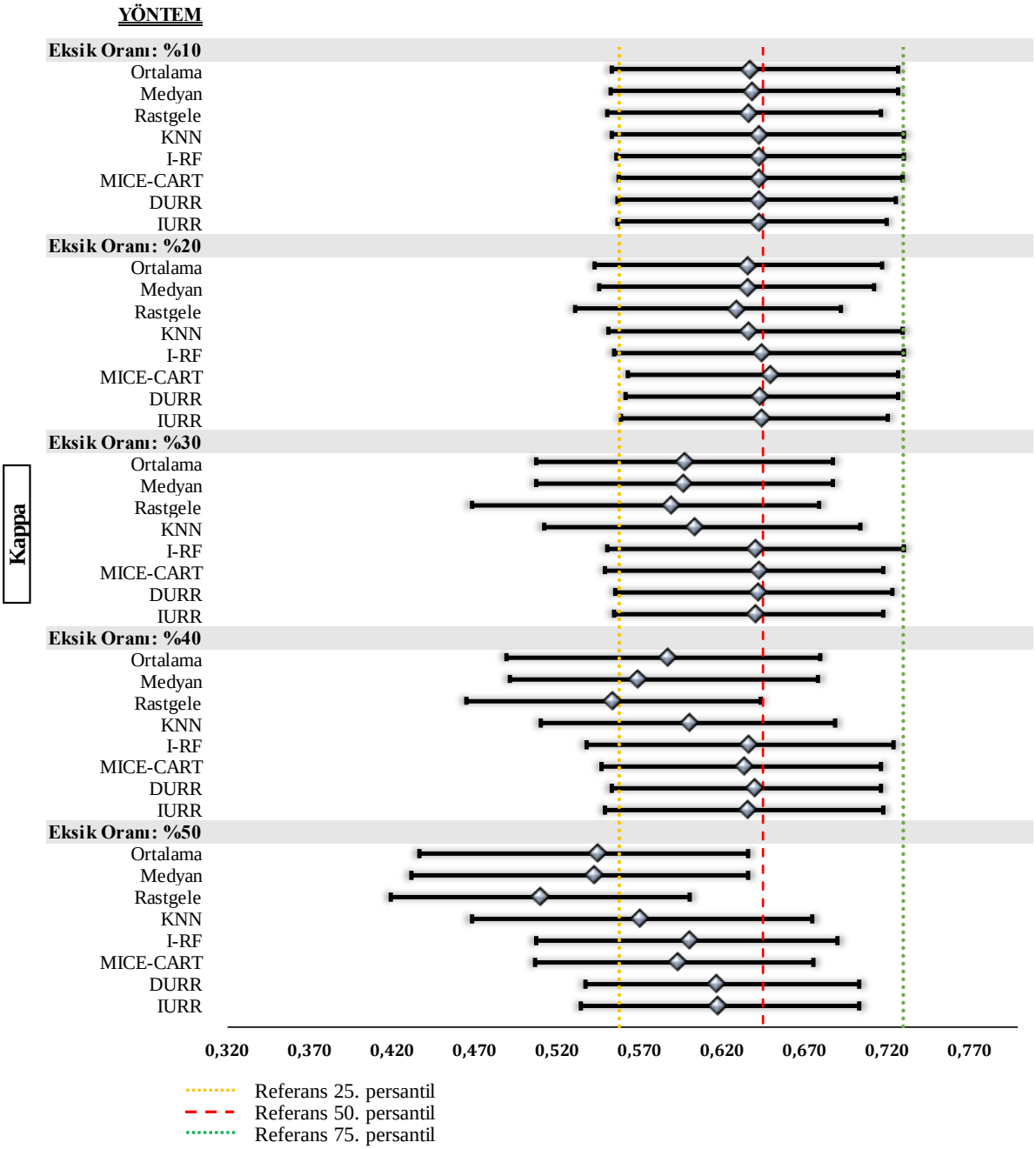
| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| **Dengeli Doğruluk** | **Referans** | 0,824 (0,780 - 0,865) | 0,824 (0,780 - 0,865) | 0,824 (0,780 - 0,865) | 0,824 (0,780 - 0,865) | 0,824 (0,780 - 0,865) |
| | **Ortalama** | 0,815 (0,775 - 0,863) | 0,815 (0,770 - 0,856) | 0,798 (0,751 - 0,843) | 0,791 (0,741 - 0,839) | 0,768 (0,715 - 0,815) |
| | **Medyan** | 0,815 (0,775 - 0,861) | 0,815 (0,770 - 0,855) | 0,798 (0,751 - 0,843) | 0,789 (0,741 - 0,835) | 0,768 (0,714 - 0,815) |
| | **Rastgele** | 0,816 (0,774 - 0,854) | 0,810 (0,765 - 0,851) | 0,791 (0,735 - 0,835) | 0,775 (0,730 - 0,822) | 0,753 (0,709 - 0,800) |
| | **KNN** | 0,820 (0,775 - 0,864) | 0,817 (0,774 - 0,863) | 0,804 (0,757 - 0,853) | 0,800 (0,754 - 0,845) | 0,785 (0,734 - 0,835) |
| | **I-RF** | 0,820 (0,776 - 0,865) | 0,822 (0,777 - 0,864) | 0,820 (0,773 - 0,863) | 0,815 (0,768 - 0,860) | 0,800 (0,751 - 0,848) |
| | **MICE-CART** | 0,821 (0,778 - 0,865) | 0,824 (0,780 - 0,862) | 0,820 (0,772 - 0,859) | 0,816 (0,773 - 0,857) | 0,797 (0,748 - 0,837) |
| | **DURR** | 0,820 (0,778 - 0,864) | 0,821 (0,781 - 0,861) | 0,820 (0,775 - 0,861) | 0,817 (0,775 - 0,858) | 0,807 (0,768 - 0,852) |
| | **IURR** | 0,820 (0,778 - 0,862) | 0,822 (0,777 - 0,860) | 0,819 (0,778 - 0,860) | 0,817 (0,773 - 0,858) | 0,809 (0,766 - 0,853) |
| **AUC** | **Referans** | 0,877 (0,832 - 0,913) | 0,877 (0,832 - 0,913) | 0,877 (0,832 - 0,913) | 0,877 (0,832 - 0,913) | 0,877 (0,832 - 0,913) |
| | **Ortalama** | 0,870 (0,818 - 0,911) | 0,866 (0,815 - 0,907) | 0,846 (0,790 - 0,893) | 0,834 (0,777 - 0,881) | 0,808 (0,745 - 0,862) |
| | **Medyan** | 0,870 (0,818 - 0,911) | 0,866 (0,816 - 0,907) | 0,844 (0,789 - 0,895) | 0,836 (0,777 - 0,880) | 0,807 (0,744 - 0,860) |
| | **Rastgele** | 0,866 (0,815 - 0,907) | 0,860 (0,808 - 0,901) | 0,837 (0,773 - 0,882) | 0,820 (0,764 - 0,869) | 0,787 (0,731 - 0,842) |
| | **KNN** | 0,871 (0,820 - 0,912) | 0,870 (0,819 - 0,911) | 0,858 (0,802 - 0,903) | 0,851 (0,794 - 0,896) | 0,832 (0,769 - 0,882) |
| | **I-RF** | 0,872 (0,824 - 0,913) | 0,874 (0,824 - 0,914) | 0,868 (0,815 - 0,911) | 0,866 (0,815 - 0,908) | 0,848 (0,791 - 0,895) |
| | **MICE-CART** | 0,872 (0,826 - 0,912) | 0,876 (0,827 - 0,913) | 0,870 (0,819 - 0,911) | 0,867 (0,814 - 0,908) | 0,842 (0,786 - 0,887) |
| | **DURR** | 0,872 (0,824 - 0,914) | 0,874 (0,826 - 0,914) | 0,870 (0,821 - 0,912) | 0,870 (0,820 - 0,909) | 0,861 (0,810 - 0,902) |
| | **IURR** | 0,871 (0,826 - 0,913) | 0,874 (0,826 - 0,913) | 0,872 (0,822 - 0,911) | 0,871 (0,820 - 0,909) | 0,857 (0,809 - 0,902) |
| **Kappa** | **Referans** | 0,645 (0,558 - 0,731) | 0,645 (0,558 - 0,731) | 0,645 (0,558 - 0,731) | 0,645 (0,558 - 0,731) | 0,645 (0,558 - 0,731) |
| | **Ortalama** | 0,637 (0,554 - 0,727) | 0,636 (0,543 - 0,717) | 0,597 (0,507 - 0,687) | 0,587 (0,489 - 0,680) | 0,545 (0,437 - 0,636) |
| | **Medyan** | 0,639 (0,553 - 0,727) | 0,636 (0,545 - 0,712) | 0,597 (0,507 - 0,687) | 0,569 (0,492 - 0,679) | 0,543 (0,432 - 0,636) |
| | **Rastgele** | 0,636 (0,551 - 0,717) | 0,629 (0,531 - 0,692) | 0,589 (0,469 - 0,679) | 0,554 (0,465 - 0,644) | 0,510 (0,419 - 0,601) |
| | **KNN** | 0,643 (0,554 - 0,731) | 0,636 (0,552 - 0,730) | 0,604 (0,512 - 0,704) | 0,600 (0,510 - 0,689) | 0,571 (0,469 - 0,675) |
| | **I-RF** | 0,643 (0,556 - 0,731) | 0,644 (0,554 - 0,731) | 0,640 (0,551 - 0,731) | 0,636 (0,538 - 0,724) | 0,601 (0,507 - 0,690) |
| | **MICE-CART** | 0,643 (0,557 - 0,730) | 0,650 (0,563 - 0,727) | 0,642 (0,549 - 0,718) | 0,633 (0,547 - 0,717) | 0,594 (0,507 - 0,676) |
| | **DURR** | 0,643 (0,557 - 0,726) | 0,643 (0,562 - 0,727) | 0,642 (0,556 - 0,724) | 0,640 (0,553 - 0,717) | 0,617 (0,537 - 0,703) |
| | **IURR** | 0,643 (0,557 - 0,720) | 0,644 (0,559 - 0,721) | 0,641 (0,555 - 0,718) | 0,636 (0,549 - 0,718) | 0,618 (0,535 - 0,704) |



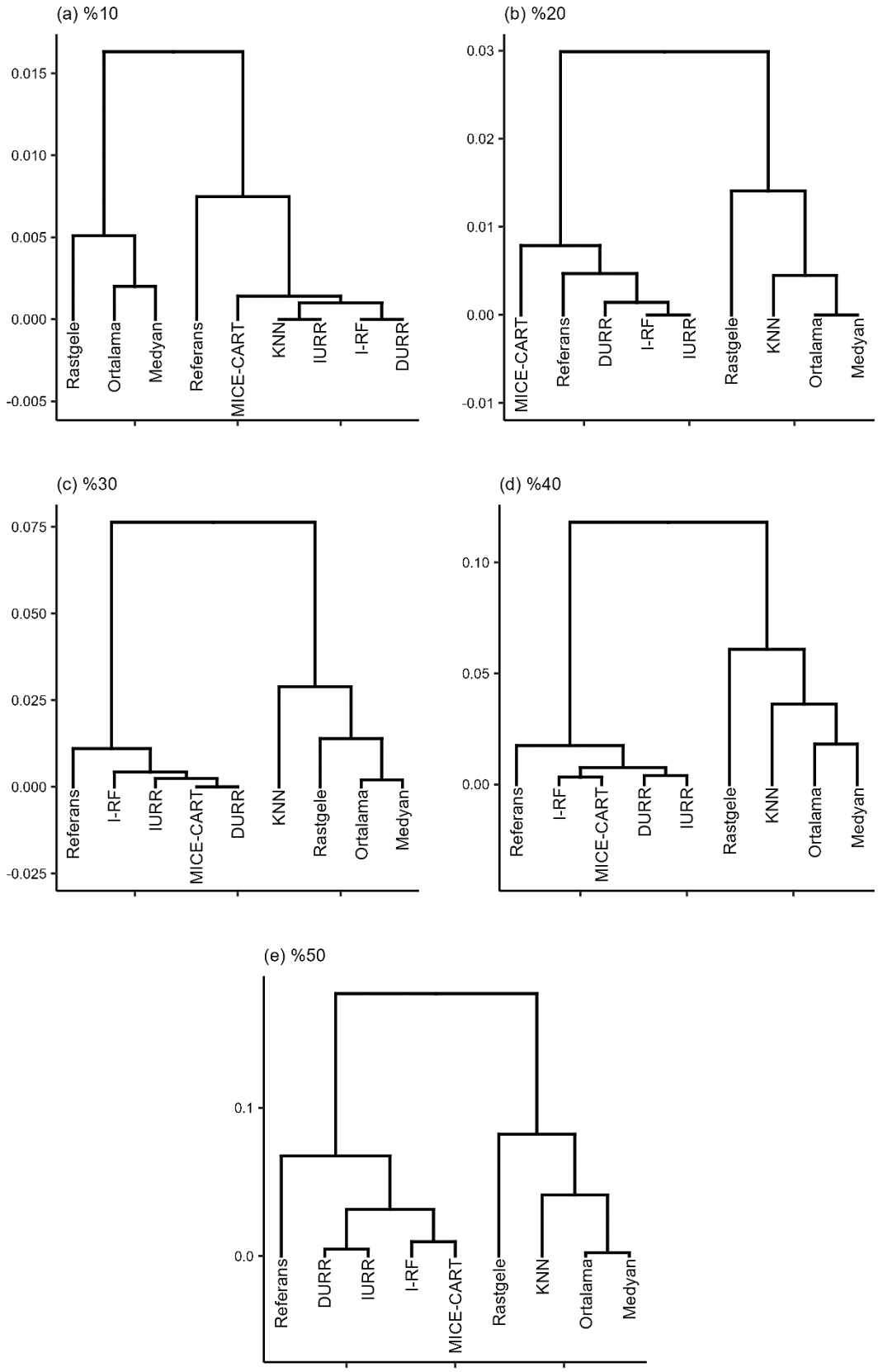

**Şekil 16.** -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin dengeli doğruluk oranlarının orman grafiği




**Şekil 17**. -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin AUC değerlerinin orman grafiği

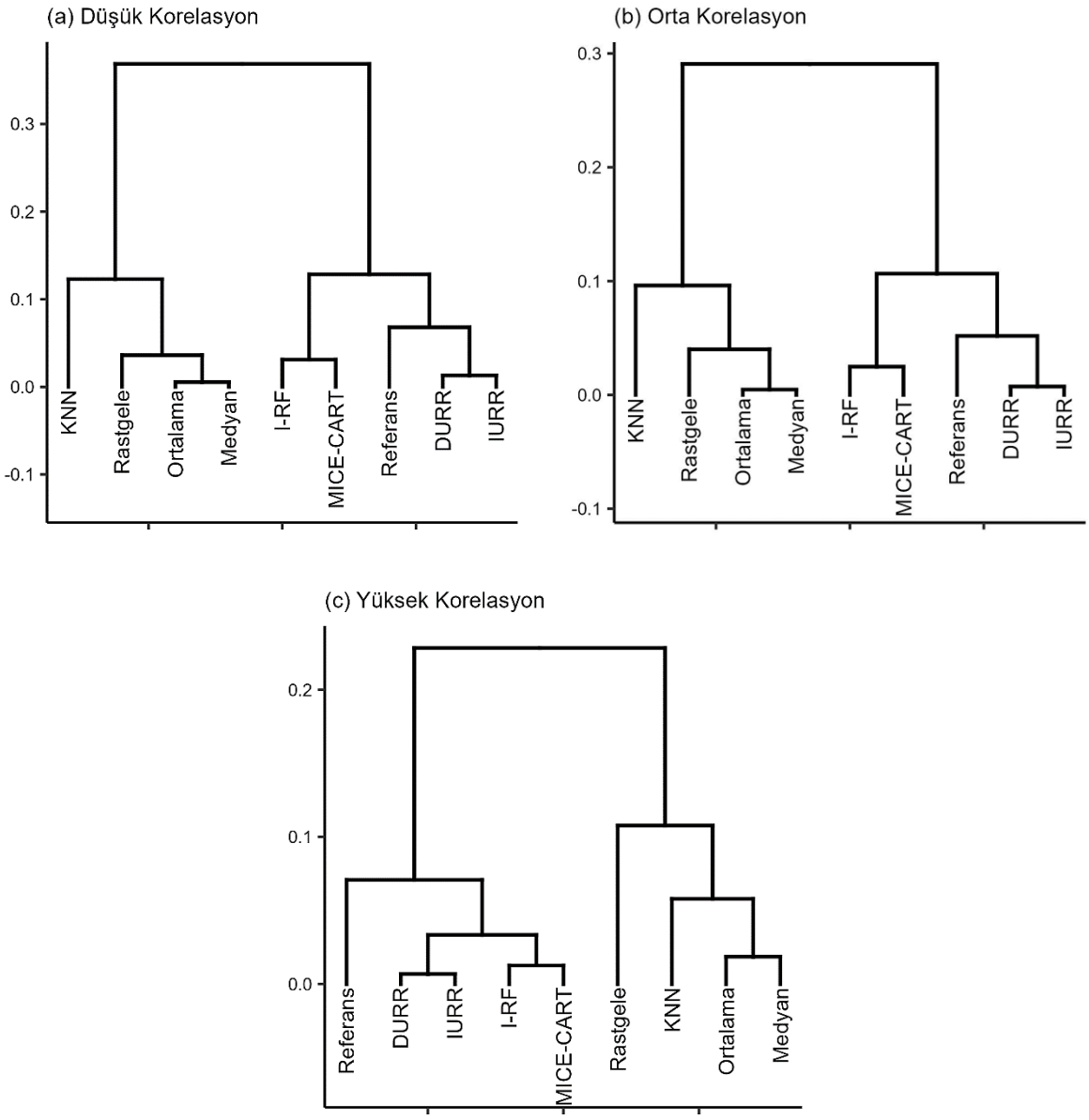



**Şekil 18**. -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için yöntemlerin kappa değerlerinin orman grafiği




**Şekil 19.** -0,8 ≤ r ≤ 0,8 aralığına göre rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için farklı eksik oranlarında yöntemlerin dendrogram grafikleri

Korelasyon katsayısı $-0,1 \leq r \leq 0,1$; $-0,5 \leq r \leq 0,5$ ve $-0,8 \leq r \leq 0,8$ için yöntemlerin tüm eksik oranlarındaki dengeli doğruluk oranı, AUC ve kappa sonuçları kullanılarak aşamalı kümeleme analizi yapılmış ve bu analiz sonucu elde edilen dendrogramlar Şekil 20'de verilmiştir. Buna göre tüm korelasyon düzeyleri için DURR, IURR, MICE-CART ve I-RF yöntemlerinin referans ile aynı küme içerisinde yer aldıkları görülmüştür. Ayrıca ortalama, medyan, rastgele ve KNN yöntemleri de benzer performans göstererek ayrı bir küme oluşturduğu belirlenmiştir. Buna ek olarak DURR ve IURR yöntemleri düşük ve orta korelasyon düzeylerinde referansa diğer yöntemlerden daha yakınken yüksek korelasyon düzeyinde MICE-CART ve I-RF yöntemlerinin performansı da bu iki yönteme ve referansa yaklaşmıştır (Şekil 20).




**Şekil 20**. Rastgele bir değişken setinin doğrusal kombinasyonundan türetilen eksik verili değişkenler için farklı korelasyon düzeylerinde yöntemlerin dendrogram grafikleri

## 4.2. Tamamen Rastgele Türetilen Veriler için Bulgular

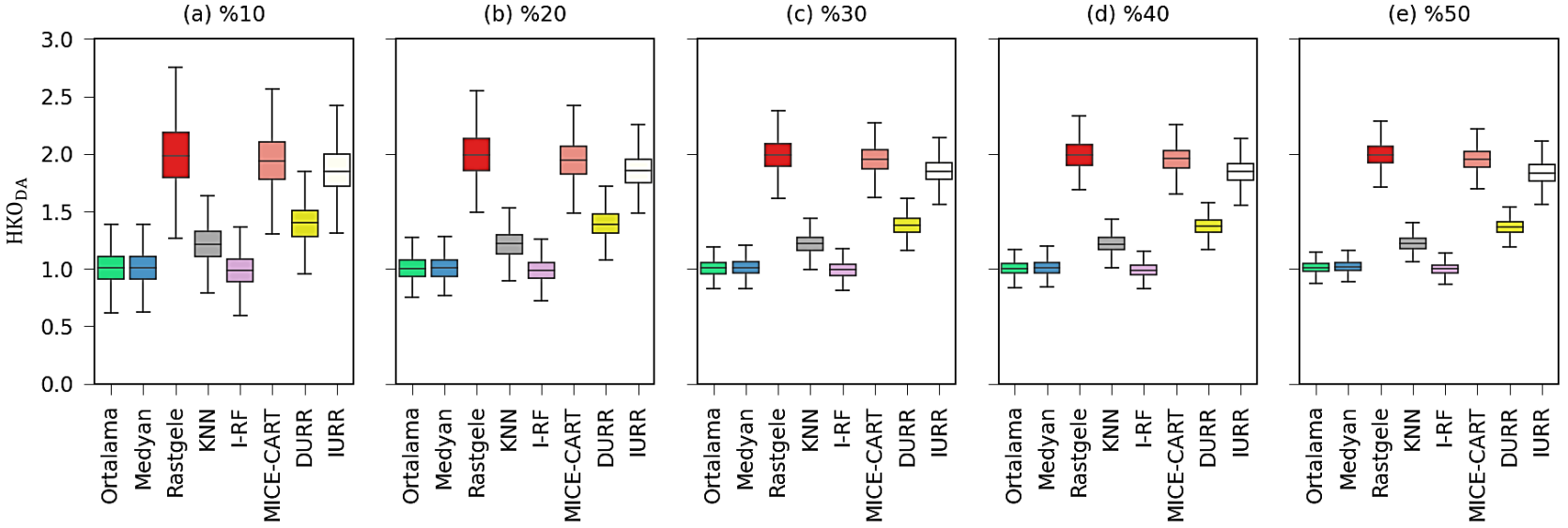
### 4.2.1. Tamamen Rastgele Türetilen Verilerde -0,1 ≤ r ≤ 0,1 Aralığı için Bulgular

Eksik oranı %10, %20, %30, %40 ve %50 için $HKO_{DA}$ değerlerinin medyan değişim aralığı sırasıyla 0,986-1,982; 0,987-1,990; 0,991-1,992; 0,990-1,991; 1,001-1,991'dir. Ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR yöntemlerinin $HKO_{DA}$ değerlerine ilişkin medyan değişim aralığı ise sırasıyla 1,004-1,012; 1,010-1,020; 1,982-1,992; 1,214-1,224; 0,986-1,001; 1,937-1,958; 1,363-1,401; 1,831-1,857'dir (Tablo 8).

Tüm değer atama yöntemlerinin $HKO_{DA}$ değerleri farklı eksik oranlarına göre incelendiğinde tüm eksik oranlarında rastgele, MICE-CART ve IURR yöntemlerinin diğer yöntemlere göre daha yüksek $HKO_{DA}$ değerine sahip olduğu belirlenmiştir (Tablo 8 ve Şekil 21).

**Tablo 8.** Tamamen rastgele türetilen verilerde -0,1 ≤ r ≤ 0,1 aralığı için yöntemlerin $HKO_{DA}$ değerleri

| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| **$HKO_{DA}$** | **Ortalama** | 1,008 (0,908 - 1,106) | 1,004 (0,932 - 1,074) | 1,006 (0,957 - 1,055) | 1,006 (0,962 - 1,049) | 1,012 (0,977 - 1,047) |
| | **Medyan** | 1,010 (0,913 - 1,109) | 1,010 (0,935 - 1,075) | 1,013 (0,964 - 1,061) | 1,012 (0,965 - 1,058) | 1,020 (0,985 - 1,055) |
| | **Rastgele** | 1,982 (1,798 - 2,187) | 1,990 (1,854 - 2,136) | 1,992 (1,892 - 2,088) | 1,991 (1,903 - 2,078) | 1,991 (1,921 - 2,069) |
| | **KNN** | 1,214 (1,108 - 1,327) | 1,223 (1,134 - 1,294) | 1,222 (1,163 - 1,275) | 1,216 (1,167 - 1,274) | 1,224 (1,178 - 1,269) |
| | **I-RF** | 0,986 (0,890 - 1,086) | 0,987 (0,920 - 1,057) | 0,991 (0,945 - 1,040) | 0,990 (0,949 - 1,032) | 1,001 (0,964 - 1,035) |
| | **MICE-CART** | 1,937 (1,780 - 2,105) | 1,946 (1,825 - 2,066) | 1,951 (1,871 - 2,039) | 1,958 (1,877 - 2,031) | 1,953 (1,887 - 2,023) |
| | **DURR** | 1,401 (1,281 - 1,510) | 1,388 (1,309 - 1,474) | 1,377 (1,319 - 1,437) | 1,372 (1,321 - 1,425) | 1,363 (1,320 - 1,410) |
| | **IURR** | 1,848 (1,721 - 2,002) | 1,857 (1,752 - 1,956) | 1,851 (1,781 - 1,927) | 1,848 (1,771 - 1,917) | 1,831 (1,765 - 1,909) |




**Şekil 21.** Tamamen rastgele türetilen verilerde -0,1 ≤ r ≤ 0,1 aralığı için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği

Referans veri setinden elde edilen dengeli doğruluk oranlarının medyanı 0,574 olarak elde edilmiştir. Atanmış veri setlerinden elde edilen sonuçlara göre dengeli doğruluk oranı değerlerinin medyan değişim aralığı eksik oranı %10 ve %40 için 0,580-0,583, %20 ve %50 için 0,580-0,582; %30 için 0,579-0,583'dür. Değer atama yöntemlerinin dengeli doğruluk oranlarına ilişkin medyan değişim aralığı ise ortalama, medyan ve rastgele için 0,580-0,582; KNN, I-RF ve IURR için 0,580-0,583; MICE-CART için 0,579-0,583; DURR için 0,582-0,583'dür (Tablo 9 ve Şekil 22).

Referans veri setinden elde edilen AUC değerlerinin medyanı 0,562 olarak elde edilmiştir. Atanmış veri setlerinden elde edilen sonuçlara göre AUC değerlerinin medyan değişim aralığı eksik oranı %10, %20, %30, %40 ve %50 için sırasıyla 0,566-0,567; 0,565-0,567; 0,564-0,568; 0,562-0,566; 0,563-0,567'dir. Değer atama yöntemlerinin AUC oranlarına ilişkin medyan değişim aralığı ise ortalama ve medyan için 0,565-0,567; rastgele için 0,565-0,566; KNN için 0,566-0,568; I-RF için 0,565-0,568; MICE-CART ve DURR için 0,563-0,566; IURR için 0,562-0,566'dır (Tablo 9 ve Şekil 23).

Referans veri setinden elde edilen kappa değerlerinin medyanı 0,156 olarak elde edilmiştir. Atanmış veri setlerinden elde edilen sonuçlara göre kappa değerlerinin medyan değişim aralığı eksik oranı %10, %20, %30, %40 ve %50 için sırasıyla 0,166-0,173; 0,165-0,167; 0,164-0,172; 0,166-0,172; 0,166-0,171'dir. Değer atama yöntemlerinin kappa oranlarına ilişkin medyan değişim aralığı ise ortalama için 0,167-0,169; medyan ve I-RF için 0,166-0,168; rastgele için 0,166-0,167; KNN için 0,165-0,172; MICE-CART için 0,164-0,171; DURR için 0,167-0,173; IURR için 0,166-0,172'dir (Tablo 9 ve Şekil 24).

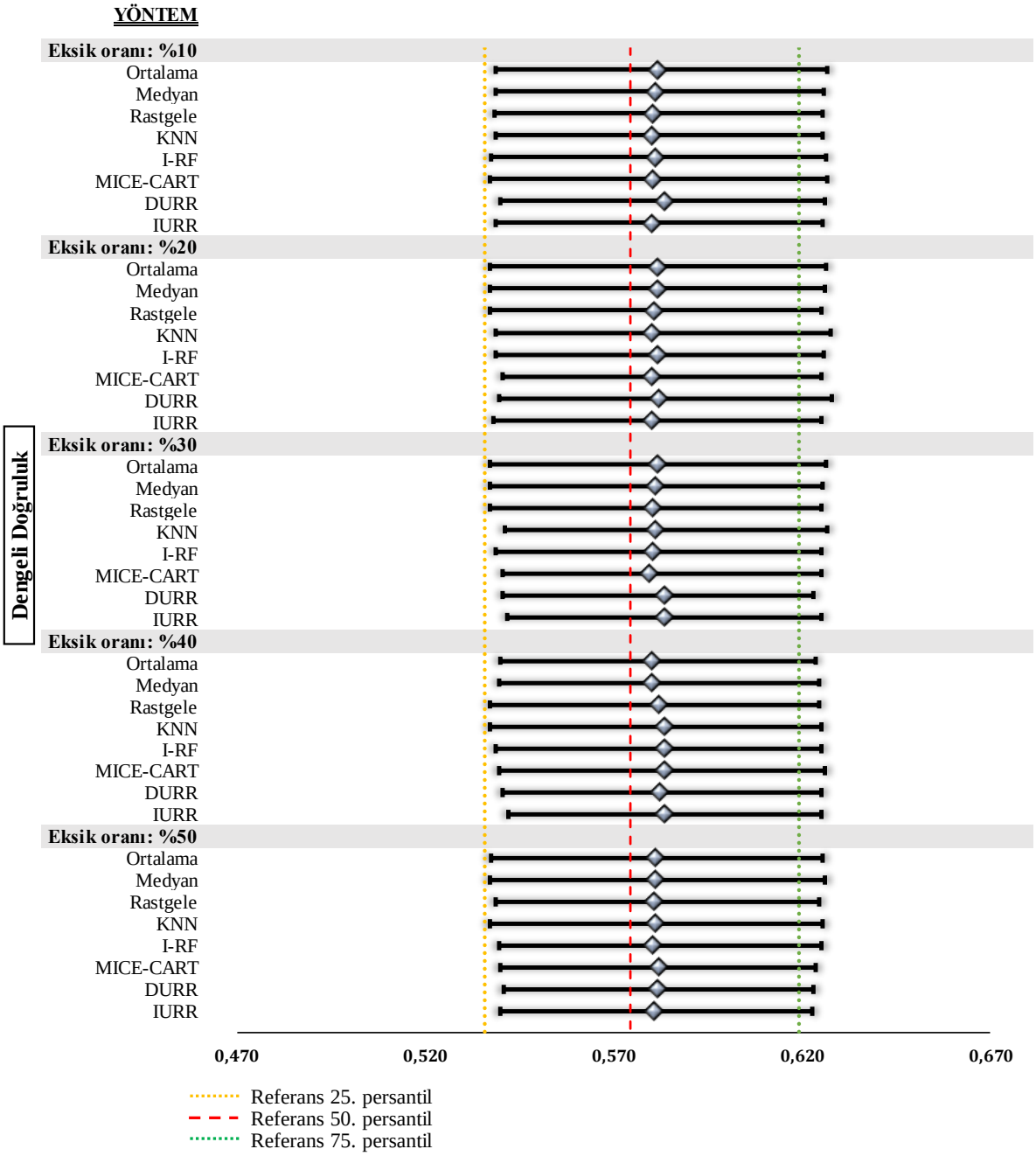
Tüm değer atama yöntemlerinin performansları değişen eksik oranına göre incelendiğinde, yöntemlerin eksik oranındaki artıştan genel olarak etkilenmedikleri; tüm yöntemlerin dengeli doğruluk oranı ve kappa değerleri bakımından tüm eksik oranlarında referans ile orta düzeyde yakın; AUC değerleri bakımından ise düşük eksik oranlarında referans ile orta düzeyde yakın, yüksek eksik oranlarında MICE-CART, DURR ve IURR yöntemlerinin yüksek düzeyde, diğer yöntemlerin orta düzeyde yakın tahminlerde bulunduğu görülmektedir.

Dengeli doğruluk oranı, AUC ve kappa sonuçlarına göre uygulanan aşamalı kümeleme analizi ile elde edilen dendrogram grafikleri Şekil 25'te verilmiştir. Buna göre referans ile tüm yöntemlerin birbirinden ayrıldığı gözlenmiştir. Eksik oranı %10 için MICE-CART, ortalama, medyan, rastgele, IURR, KNN ve I-RF yöntemlerinin birbirine yakın performans göstererek bir küme oluşturdukları görülmektedir (Şekil 25a). Eksik oranı %20 için tüm yöntemler tek bir küme içerisinde yer alırken birbirine en yakın iki yöntemin ortalama ve DURR olduğu

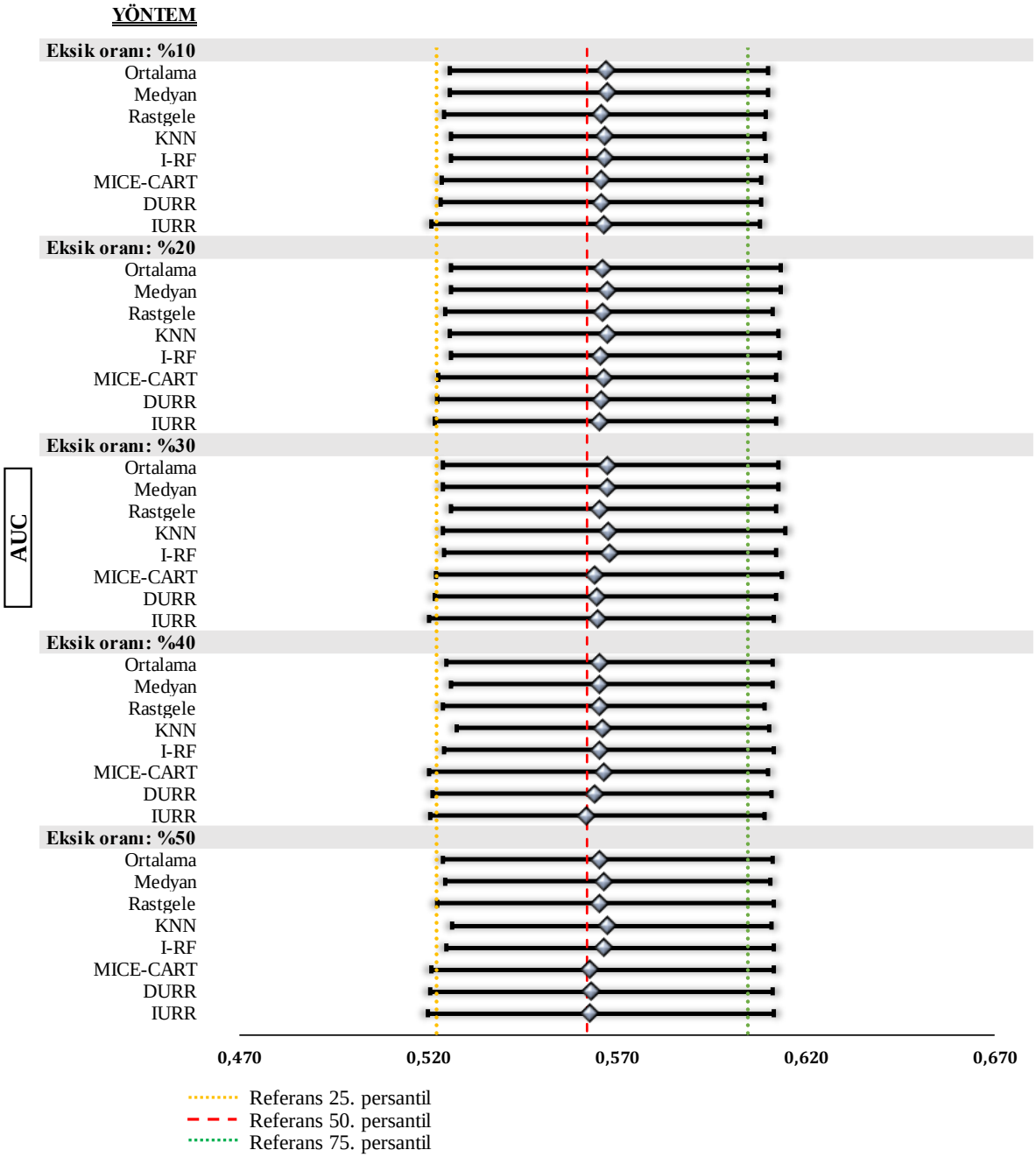
belirlenmiştir (Şekil 25b). Eksik oranı %30 için DURR ve IURR; rastgele, MICE-CART, I-RF, KNN, ortalama ve medyan yöntemlerinin iki ayrı küme oluşturduğu gözlenmiştir (Şekil 25c). Eksik oranı %40 için ortalama, medyan, rastgele ve I-RF; KNN, MICE-CART, DURR ve IURR yöntemleri iki ayrı küme oluştururken daha sonra bu iki küme birleşerek tek bir küme oluşturmuştur (Şekil 25d). Eksik oranı %50 için DURR ve IURR; rastgele, ortalama, medyan, KNN ve I-RF yöntemlerinin birbirine benzer performans göstererek iki ayrı küme içerisinde yer aldıkları bulgusuna ulaşılmıştır (Şekil 25e).

**Tablo 9.** Tamamen rastgele türetilen verilerde -0,1 ≤ r ≤ 0,1 aralığı için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri

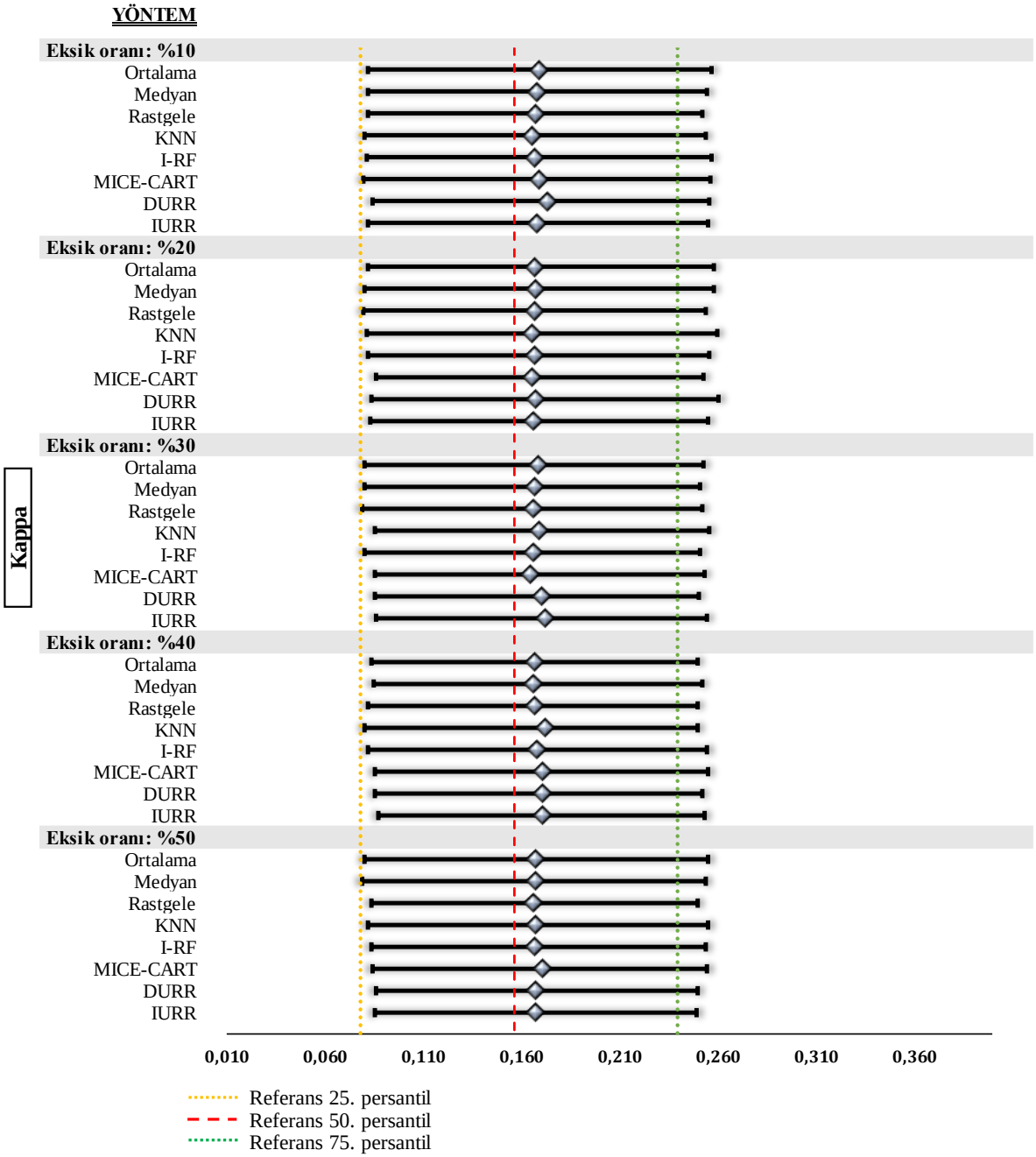
| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| **Dengeli Doğruluk** | **Referans** | 0,574 (0,536 - 0,619) | 0,574 (0,536 - 0,619) | 0,574 (0,536 - 0,619) | 0,574 (0,536 - 0,619) | 0,574 (0,536 - 0,619) |
| | **Ortalama** | 0,581 (0,539 - 0,627) | 0,582 (0,537 - 0,626) | 0,581 (0,537 - 0,626) | 0,580 (0,540 - 0,624) | 0,581 (0,537 - 0,625) |
| | **Medyan** | 0,581 (0,539 - 0,626) | 0,582 (0,537 - 0,626) | 0,581 (0,537 - 0,625) | 0,580 (0,540 - 0,625) | 0,581 (0,537 - 0,626) |
| | **Rastgele** | 0,580 (0,538 - 0,625) | 0,581 (0,537 - 0,625) | 0,580 (0,537 - 0,625) | 0,582 (0,537 - 0,625) | 0,581 (0,539 - 0,625) |
| | **KNN** | 0,580 (0,539 - 0,625) | 0,580 (0,539 - 0,628) | 0,581 (0,541 - 0,627) | 0,583 (0,537 - 0,625) | 0,581 (0,537 - 0,625) |
| | **I-RF** | 0,581 (0,537 - 0,626) | 0,581 (0,539 - 0,626) | 0,580 (0,539 - 0,625) | 0,583 (0,539 - 0,625) | 0,580 (0,540 - 0,625) |
| | **MICE-CART** | 0,580 (0,537 - 0,627) | 0,580 (0,541 - 0,625) | 0,579 (0,541 - 0,625) | 0,583 (0,540 - 0,626) | 0,582 (0,540 - 0,624) |
| | **DURR** | 0,583 (0,540 - 0,626) | 0,582 (0,540 - 0,628) | 0,583 (0,541 - 0,623) | 0,582 (0,541 - 0,625) | 0,582 (0,541 - 0,623) |
| | **IURR** | 0,580 (0,539 - 0,626) | 0,580 (0,538 - 0,625) | 0,583 (0,542 - 0,625) | 0,583 (0,542 - 0,625) | 0,581 (0,540 - 0,623) |
| **AUC** | **Referans** | 0,562 (0,522 - 0,605) | 0,562 (0,522 - 0,605) | 0,562 (0,522 - 0,605) | 0,562 (0,522 - 0,605) | 0,562 (0,522 - 0,605) |
| | **Ortalama** | 0,567 (0,526 - 0,610) | 0,566 (0,526 - 0,613) | 0,567 (0,524 - 0,613) | 0,565 (0,525 - 0,611) | 0,565 (0,524 - 0,611) |
| | **Medyan** | 0,567 (0,526 - 0,610) | 0,567 (0,526 - 0,613) | 0,567 (0,524 - 0,613) | 0,565 (0,526 - 0,611) | 0,566 (0,524 - 0,611) |
| | **Rastgele** | 0,566 (0,524 - 0,609) | 0,566 (0,524 - 0,611) | 0,565 (0,526 - 0,612) | 0,565 (0,524 - 0,609) | 0,565 (0,522 - 0,612) |
| | **KNN** | 0,567 (0,526 - 0,609) | 0,567 (0,526 - 0,613) | 0,568 (0,524 - 0,615) | 0,566 (0,527 - 0,610) | 0,567 (0,526 - 0,611) |
| | **I-RF** | 0,567 (0,526 - 0,609) | 0,565 (0,526 - 0,613) | 0,568 (0,524 - 0,612) | 0,565 (0,524 - 0,612) | 0,566 (0,525 - 0,611) |
| | **MICE-CART** | 0,566 (0,524 - 0,608) | 0,566 (0,522 - 0,612) | 0,564 (0,522 - 0,614) | 0,566 (0,520 - 0,610) | 0,563 (0,521 - 0,611) |
| | **DURR** | 0,566 (0,523 - 0,608) | 0,566 (0,522 - 0,611) | 0,565 (0,522 - 0,612) | 0,564 (0,521 - 0,611) | 0,563 (0,520 - 0,611) |
| | **IURR** | 0,566 (0,521 - 0,608) | 0,565 (0,522 - 0,612) | 0,565 (0,520 - 0,611) | 0,562 (0,521 - 0,609) | 0,563 (0,520 - 0,612) |
| **Kappa** | **Referans** | 0,156 (0,078 - 0,239) | 0,156 (0,078 - 0,239) | 0,156 (0,078 - 0,239) | 0,156 (0,078 - 0,239) | 0,156 (0,078 - 0,239) |
| | **Ortalama** | 0,169 (0,082 - 0,257) | 0,167 (0,082 - 0,258) | 0,168 (0,080 - 0,253) | 0,167 (0,083 - 0,250) | 0,167 (0,080 - 0,255) |
| | **Medyan** | 0,168 (0,082 - 0,255) | 0,167 (0,080 - 0,258) | 0,167 (0,080 - 0,251) | 0,166 (0,085 - 0,252) | 0,167 (0,079 - 0,254) |
| | **Rastgele** | 0,167 (0,082 - 0,252) | 0,167 (0,079 - 0,254) | 0,166 (0,079 - 0,252) | 0,167 (0,082 - 0,250) | 0,166 (0,083 - 0,250) |
| | **KNN** | 0,166 (0,080 - 0,254) | 0,165 (0,081 - 0,260) | 0,169 (0,085 - 0,256) | 0,172 (0,080 - 0,250) | 0,167 (0,082 - 0,255) |
| | **I-RF** | 0,167 (0,081 - 0,257) | 0,167 (0,082 - 0,256) | 0,166 (0,080 - 0,251) | 0,168 (0,082 - 0,255) | 0,167 (0,083 - 0,254) |
| | **MICE-CART** | 0,169 (0,079 - 0,256) | 0,165 (0,086 - 0,253) | 0,164 (0,085 - 0,254) | 0,171 (0,085 - 0,255) | 0,171 (0,084 - 0,254) |
| | **DURR** | 0,173 (0,084 - 0,255) | 0,167 (0,083 - 0,261) | 0,170 (0,085 - 0,250) | 0,171 (0,085 - 0,252) | 0,167 (0,086 - 0,249) |
| | **IURR** | 0,168 (0,082 - 0,255) | 0,166 (0,083 - 0,255) | 0,172 (0,086 - 0,254) | 0,171 (0,087 - 0,253) | 0,167 (0,085 - 0,249) |



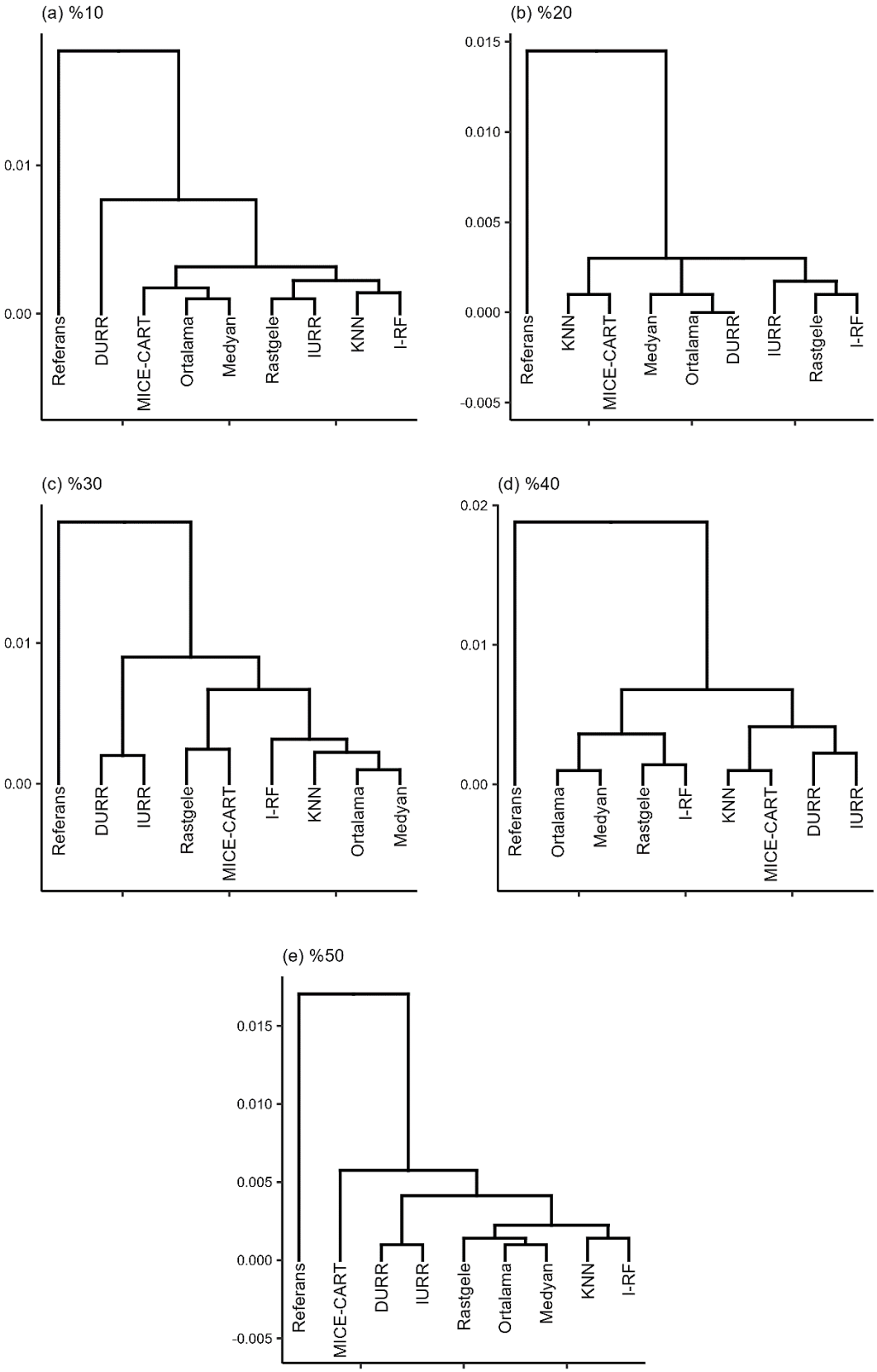

**Şekil 22.** Tamamen rastgele türetilen verilerde -0,1 ≤ r ≤ 0,1 aralığı için yöntemlerin dengeli doğruluk oranlarının orman grafiği




**Şekil 23.** Tamamen rastgele türetilen verilerde $-0{,}1 \leq r \leq 0{,}1$ aralığı için yöntemlerin AUC değerlerinin orman grafiği




**Şekil 24.** Tamamen rastgele türetilen verilerde $-0{,}1 \leq r \leq 0{,}1$ aralığı için yöntemlerin kappa değerlerinin orman grafiği




**Şekil 25.** Tamamen rastgele türetilen veri setlerinde $-0{,}1 \leq r \leq 0{,}1$ aralığı ve farklı eksik oranları için yöntemlerin dendrogram grafikleri

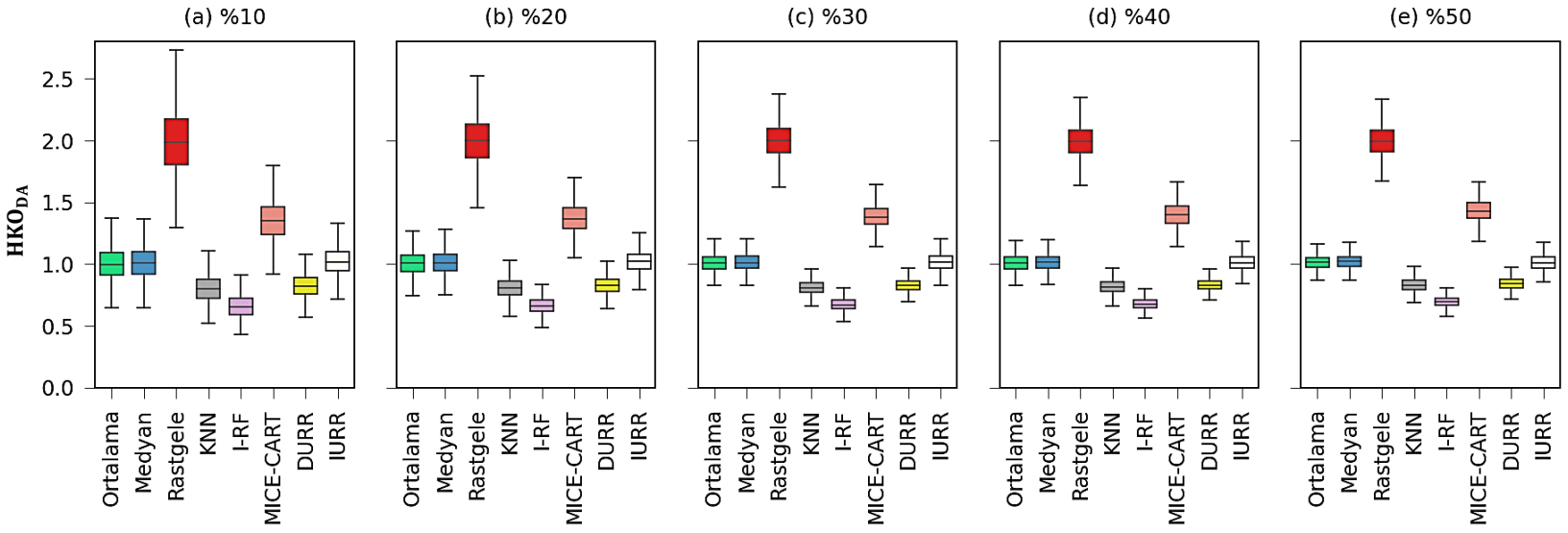
### 4.2.2. Tamamen Rastgele Türetilen Verilerde -0,5 ≤ r ≤ 0,5 Aralığı için Bulgular

Eksik oranı %10, %20, %30, %40 ve %50 için $HKO_{DA}$ değerlerinin medyan değişim aralığı sırasıyla 0,653-1,985; 0,661-1,998; 0,668-1,999; 0,673-1,992; 0,692-1,996'dır. Ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR yöntemlerinin $HKO_{DA}$ değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,999-1,014; 1,007-1,022; 1,985-1,999; 0,798-0,830; 0,653-0,692; 1,353-1,431; 0,823-0,844; 1,011-1,023'tür (Tablo 10).

Tüm değer atama yöntemlerinin $HKO_{DA}$ değerleri farklı eksik oranlarına göre incelendiğinde genel olarak rastgele değer atama hariç diğer yöntemlerin $HKO_{DA}$ değerleri arasında büyük bir fark olmadığı; tüm eksik oranları için KNN, I-RF ve DURR yöntemlerinin diğer yöntemlere göre daha düşük $HKO_{DA}$ değerlerine sahip olduğu; en yüksek $HKO_{DA}$ değerine sahip olan yöntemin ise rastgele değer atama yöntemi olduğu belirlenmiştir (Tablo 10 ve Şekil 26).

**Tablo 10**. Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin $HKO_{DA}$ değerleri

| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| $HKO_{DA}$ | **Ortalama** | 0,999 (0,911 - 1,096) | 1,006 (0,943 - 1,075) | 1,008 (0,964 - 1,060) | 1,006 (0,964 - 1,055) | 1,014 (0,977 - 1,052) |
| | **Medyan** | 1,007 (0,917 - 1,104) | 1,010 (0,944 - 1,080) | 1,013 (0,968 - 1,065) | 1,014 (0,970 - 1,060) | 1,022 (0,984 - 1,062) |
| | **Rastgele** | 1,985 (1,802 - 2,176) | 1,998 (1,857 - 2,130) | 1,999 (1,905 - 2,099) | 1,992 (1,900 - 2,083) | 1,996 (1,910 - 2,084) |
| | **KNN** | 0,798 (0,727 - 0,878) | 0,806 (0,748 - 0,862) | 0,810 (0,774 - 0,851) | 0,815 (0,778 - 0,855) | 0,830 (0,794 - 0,870) |
| | **I-RF** | 0,653 (0,593 - 0,721) | 0,661 (0,617 - 0,707) | 0,668 (0,639 - 0,707) | 0,673 (0,644 - 0,707) | 0,692 (0,665 - 0,725) |
| | **MICE-CART** | 1,353 (1,241 - 1,464) | 1,365 (1,287 - 1,456) | 1,382 (1,324 - 1,451) | 1,400 (1,333 - 1,468) | 1,431 (1,369 - 1,495) |
| | **DURR** | 0,823 (0,761 - 0,893) | 0,826 (0,777 - 0,877) | 0,826 (0,794 - 0,863) | 0,827 (0,798 - 0,864) | 0,844 (0,810 - 0,877) |
| | **IURR** | 1,017 (0,943 - 1,098) | 1,023 (0,963 - 1,082) | 1,018 (0,967 - 1,066) | 1,011 (0,970 - 1,056) | 1,013 (0,970 - 1,056) |




**Şekil 26.** Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği

Referans veri setinden elde edilen dengeli doğruluk oranlarının medyanı 0,645 olarak elde edilmiştir. Atanmış veri setlerinden elde edilen sonuçlara göre dengeli doğruluk oranı değerlerinin medyan değişim aralığı eksik oranı %10 için 0,648-0,650; %20 için 0,646-0,647; %30 ve %40 için 0,646-0,648; %50 için 0,644-0,648'dir. Değer atama yöntemlerinin dengeli doğruluk oranlarına ilişkin medyan değişim aralığı ise ortalama, medyan ve KNN için 0,646-0,648; rastgele, I-RF ve DURR için 0,645-0,648; MICE-CART için 0,646-0,650; IURR için 0,644-0,648'dir (Tablo 11 ve Şekil 27).

Referans veri setinden elde edilen AUC değerlerinin medyanı 0,633 olarak elde edilmiştir. Atanmış veri setlerinden elde edilen sonuçlara göre AUC değerlerinin medyan değişim aralığı eksik oranı %10 için 0,638-0,641; %20 için 0,636-0,639; %30 ve %40 için 0,636-0,640; %50 için 0,633-0,637'dir. Değer atama yöntemlerinin AUC değerlerine ilişkin medyan değişim aralığı ise ortalama için 0,634-0,640; medyan için 0,636-0,641; rastgele için 0,633-0,638; KNN ve MICE-CART için 0,636-0,640; I-RF için 0,633-0,639; DURR için 0,637-0,639; IURR için 0,635-0,639'dur (Tablo 11 ve Şekil 28).

Referans veri setinden elde edilen kappa değerlerinin medyanı 0,292 olarak elde edilmiştir. Atanmış veri setlerinden elde edilen sonuçlara göre eksik oranı %10, %20, %30, %40 ve %50 için kappa değerlerinin medyan değişim aralığı sırasıyla 0,302-0,305; 0,294-0,301; 0,296-0,301; 0,294-0,299; 0,292-0,301'dir. Değer atama yöntemlerinin kappa değerlerine ilişkin medyan değişim aralığı ise ortalama için 0,292-0,303; medyan için 0,293-0,303; rastgele için 0,292-0,302; KNN için 0,294-0,302; I-RF ve IURR için 0,293-0,302; MICE-CART için 0,296-0,305; DURR için 0,297-0,303'dür (Tablo 11 ve Şekil 29).

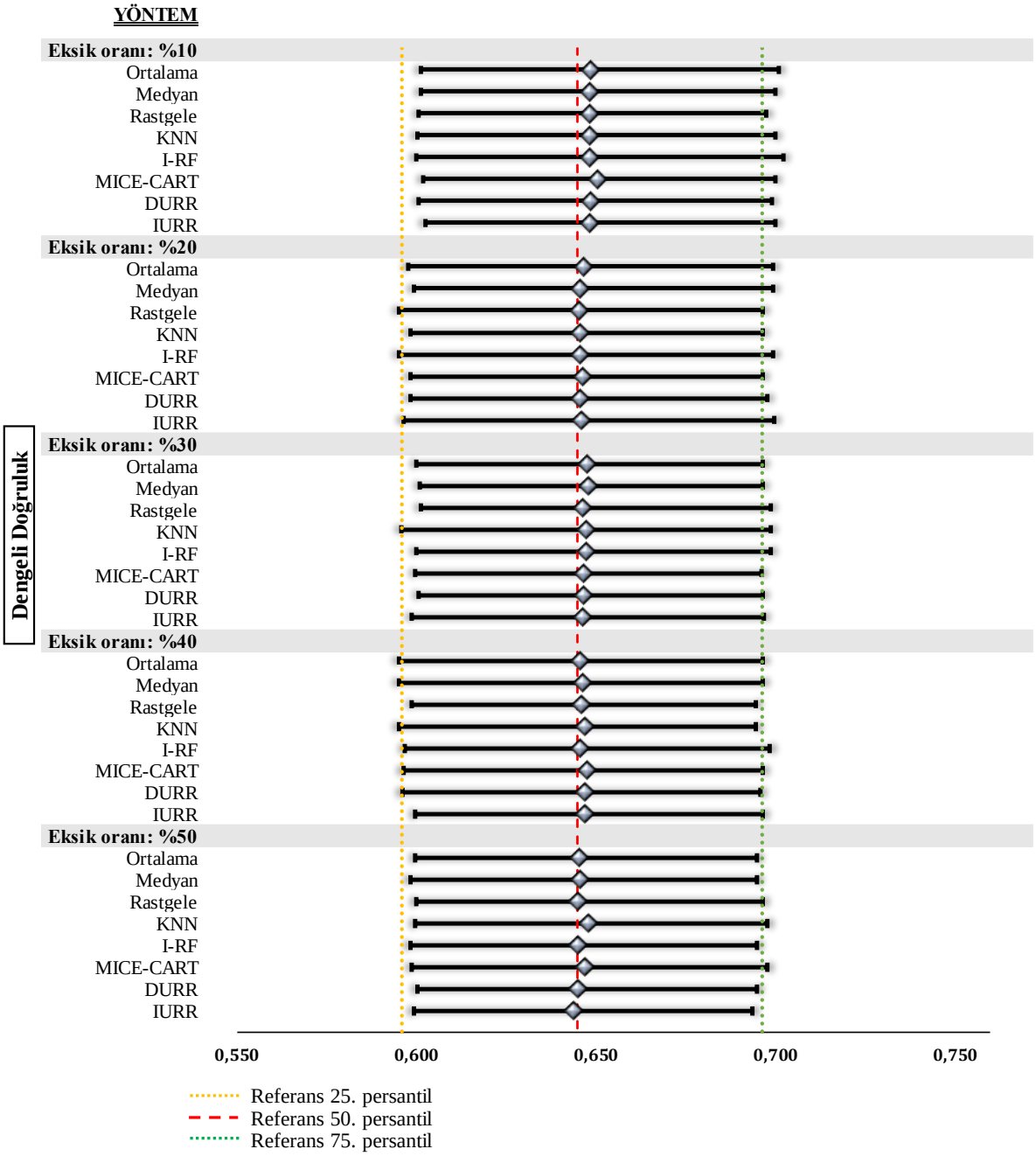
Tüm değer atama yöntemlerinin dengeli doğruluk oranı, AUC ve kappa değeri performansları değişen eksik oranına göre incelendiğinde, yöntemlerin eksik oranındaki artıştan etkilenmedikleri ve tüm yöntemlerin referansa ve birbirine orta-yüksek düzeyde yakın tahminlerde bulunduğu görülmektedir.

Dengeli doğruluk oranı, AUC ve kappa sonuçlarına göre uygulanan aşamalı kümeleme analizi ile elde edilen dendrogram grafikleri Şekil 30'da verilmiştir. Buna göre eksik oranı %10, %20, %30 ve %40 için referans ile diğer yöntemlerin birbirinden ayrıldığı gözlenmiştir. Eksik oranı %10 için KNN, ortalama, medyan, DURR, IURR, rastgele ve I-RF yöntemlerini takiben MICE-CART yönteminin birbirine yakın performans gösterip aynı kümede yer aldıkları belirlenmiştir (Şekil 30a). Eksik oranı %20 için ortalama ve medyan; MICE-CART, KNN ve I-RF; rastgele, DURR ve IURR yöntemleri üç ayrı küme içinde yer alırken, daha sonra bu üç kümenin birleşerek tek bir küme oluşturduğu gözlenmiştir (Şekil 30b). Eksik oranı %30 için iki

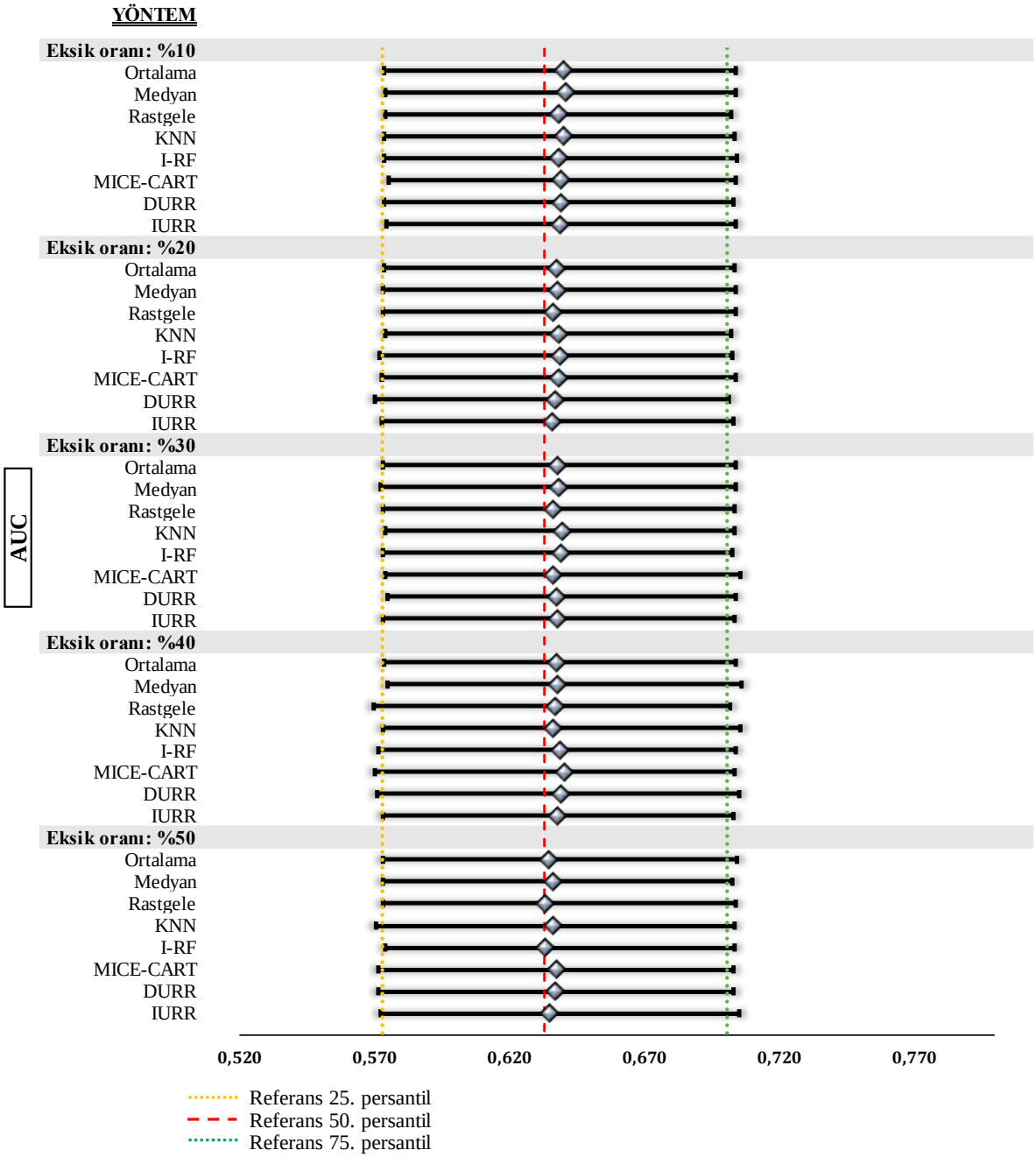
ayrı kümede yer alan yöntemler DURR, rastgele ve MICE-CART; ortalama, medyan, I-RF ve IURR yöntemleridir (Şekil 30c). Eksik oranı %40 için MICE-CART, DURR ve IURR; ortalama, I-RF, KNN, medyan ve rastgele yöntemlerinin iki ayrı kümede toplandığı belirlenmiştir (Şekil 30d). Eksik oranı %50 için referans ile aynı kümede yer alan yöntemlerin rastgele, I-RF ve ortalama yöntemleri olduğu, daha sonra medyan ve IURR yöntemlerinin de bu küme ile birleştiği; KNN, MICE-CART ve DURR yöntemlerinin ise benzer performans göstererek ayrı bir küme içinde toplandığı bulgusuna ulaşılmıştır (Şekil 30e).

**Tablo 11.** Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri

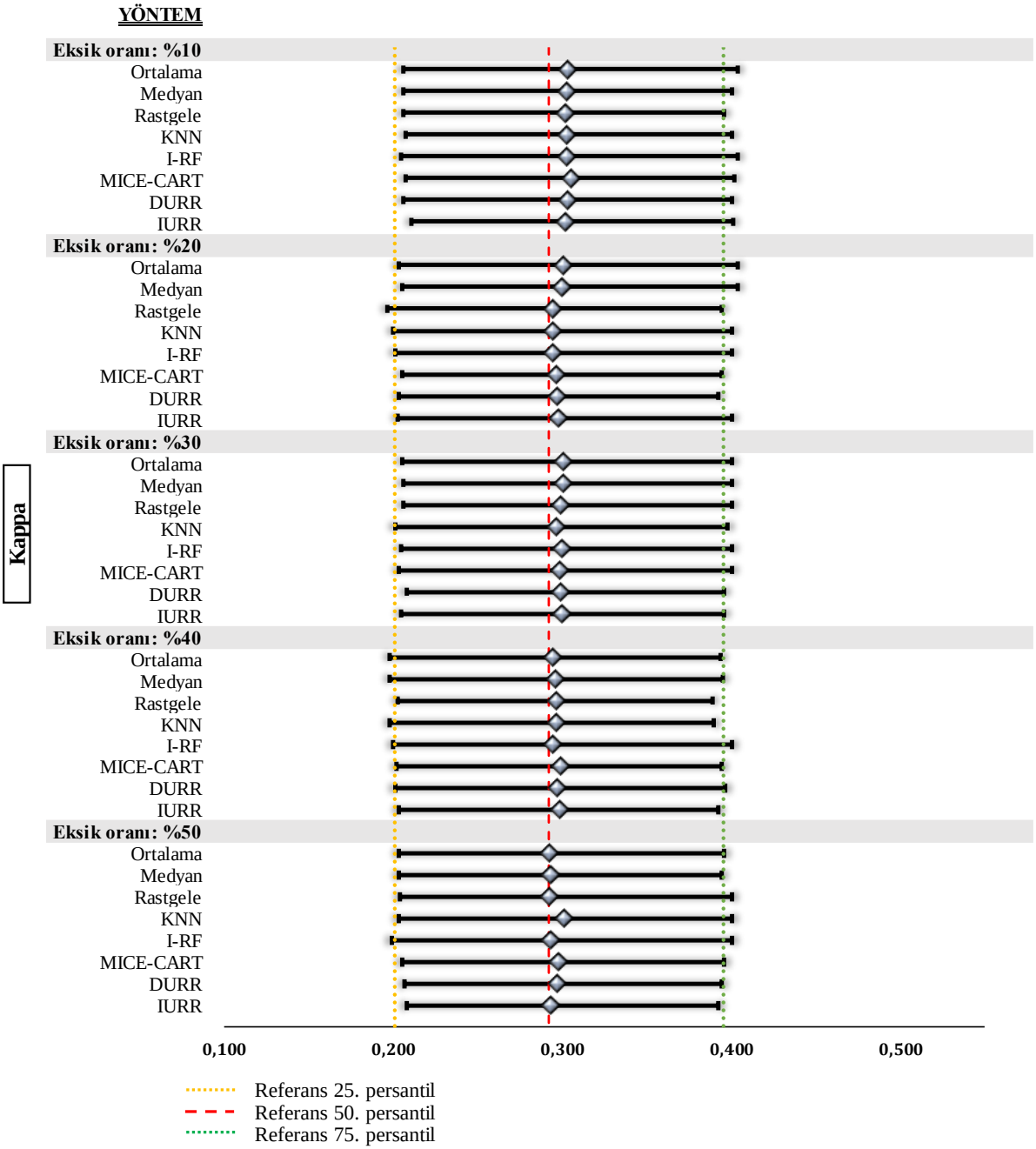
| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| **Dengeli Doğruluk** | **Referans** | 0,645 (0,596 - 0,696) | 0,645 (0,596 - 0,696) | 0,645 (0,596 - 0,696) | 0,645 (0,596 - 0,696) | 0,645 (0,596 - 0,696) |
| | **Ortalama** | 0,648 (0,601 - 0,701) | 0,647 (0,598 - 0,699) | 0,648 (0,600 - 0,696) | 0,646 (0,595 - 0,696) | 0,646 (0,600 - 0,695) |
| | **Medyan** | 0,648 (0,601 - 0,700) | 0,646 (0,599 - 0,699) | 0,648 (0,601 - 0,696) | 0,646 (0,595 - 0,696) | 0,646 (0,598 - 0,695) |
| | **Rastgele** | 0,648 (0,601 - 0,698) | 0,646 (0,595 - 0,697) | 0,646 (0,601 - 0,699) | 0,646 (0,599 - 0,695) | 0,645 (0,600 - 0,697) |
| | **KNN** | 0,648 (0,600 - 0,700) | 0,646 (0,598 - 0,696) | 0,647 (0,596 - 0,699) | 0,647 (0,595 - 0,695) | 0,648 (0,600 - 0,698) |
| | **I-RF** | 0,648 (0,600 - 0,702) | 0,646 (0,595 - 0,699) | 0,647 (0,600 - 0,699) | 0,646 (0,597 - 0,698) | 0,645 (0,598 - 0,695) |
| | **MICE-CART** | 0,650 (0,602 - 0,700) | 0,646 (0,598 - 0,697) | 0,647 (0,600 - 0,696) | 0,648 (0,596 - 0,696) | 0,647 (0,599 - 0,698) |
| | **DURR** | 0,648 (0,601 - 0,699) | 0,646 (0,598 - 0,698) | 0,647 (0,601 - 0,697) | 0,647 (0,596 - 0,696) | 0,645 (0,600 - 0,695) |
| | **IURR** | 0,648 (0,603 - 0,700) | 0,646 (0,597 - 0,700) | 0,646 (0,599 - 0,697) | 0,647 (0,600 - 0,697) | 0,644 (0,599 - 0,694) |
| **AUC** | **Referans** | 0,633 (0,573 - 0,701) | 0,633 (0,573 - 0,701) | 0,633 (0,573 - 0,701) | 0,633 (0,573 - 0,701) | 0,633 (0,573 - 0,701) |
| | **Ortalama** | 0,640 (0,574 - 0,704) | 0,637 (0,573 - 0,704) | 0,638 (0,573 - 0,704) | 0,637 (0,573 - 0,704) | 0,634 (0,573 - 0,704) |
| | **Medyan** | 0,641 (0,574 - 0,704) | 0,638 (0,573 - 0,704) | 0,638 (0,572 - 0,704) | 0,638 (0,575 - 0,706) | 0,636 (0,573 - 0,703) |
| | **Rastgele** | 0,638 (0,574 - 0,702) | 0,636 (0,573 - 0,704) | 0,636 (0,573 - 0,704) | 0,637 (0,569 - 0,702) | 0,633 (0,573 - 0,704) |
| | **KNN** | 0,640 (0,573 - 0,704) | 0,638 (0,574 - 0,702) | 0,640 (0,574 - 0,704) | 0,636 (0,573 - 0,706) | 0,636 (0,571 - 0,704) |
| | **I-RF** | 0,638 (0,573 - 0,704) | 0,639 (0,572 - 0,703) | 0,639 (0,573 - 0,703) | 0,639 (0,571 - 0,704) | 0,633 (0,574 - 0,704) |
| | **MICE-CART** | 0,639 (0,575 - 0,704) | 0,639 (0,573 - 0,704) | 0,636 (0,574 - 0,706) | 0,640 (0,570 - 0,704) | 0,637 (0,571 - 0,703) |
| | **DURR** | 0,639 (0,573 - 0,703) | 0,637 (0,570 - 0,702) | 0,637 (0,575 - 0,704) | 0,639 (0,571 - 0,705) | 0,637 (0,571 - 0,703) |
| | **IURR** | 0,639 (0,574 - 0,704) | 0,636 (0,573 - 0,703) | 0,638 (0,573 - 0,704) | 0,638 (0,573 - 0,703) | 0,635 (0,572 - 0,705) |
| **Kappa** | **Referans** | 0,292 (0,201 - 0,395) | 0,292 (0,201 - 0,395) | 0,292 (0,201 - 0,395) | 0,292 (0,201 - 0,395) | 0,292 (0,201 - 0,395) |
| | **Ortalama** | 0,303 (0,206 - 0,404) | 0,301 (0,203 - 0,404) | 0,301 (0,205 - 0,400) | 0,294 (0,198 - 0,394) | 0,292 (0,203 - 0,395) |
| | **Medyan** | 0,303 (0,206 - 0,400) | 0,299 (0,205 - 0,404) | 0,301 (0,206 - 0,400) | 0,296 (0,198 - 0,395) | 0,293 (0,203 - 0,394) |
| | **Rastgele** | 0,302 (0,206 - 0,395) | 0,294 (0,196 - 0,394) | 0,299 (0,206 - 0,400) | 0,296 (0,202 - 0,389) | 0,292 (0,204 - 0,400) |
| | **KNN** | 0,302 (0,207 - 0,400) | 0,294 (0,200 - 0,400) | 0,296 (0,201 - 0,397) | 0,296 (0,198 - 0,390) | 0,301 (0,203 - 0,400) |
| | **I-RF** | 0,302 (0,204 - 0,404) | 0,294 (0,201 - 0,400) | 0,300 (0,204 - 0,400) | 0,294 (0,200 - 0,400) | 0,293 (0,199 - 0,400) |
| | **MICE-CART** | 0,305 (0,207 - 0,402) | 0,296 (0,205 - 0,394) | 0,298 (0,203 - 0,400) | 0,299 (0,202 - 0,394) | 0,298 (0,205 - 0,396) |
| | **DURR** | 0,303 (0,206 - 0,400) | 0,297 (0,203 - 0,392) | 0,299 (0,208 - 0,396) | 0,297 (0,201 - 0,396) | 0,297 (0,207 - 0,394) |
| | **IURR** | 0,302 (0,211 - 0,401) | 0,297 (0,202 - 0,400) | 0,300 (0,205 - 0,396) | 0,298 (0,203 - 0,392) | 0,293 (0,208 - 0,392) |



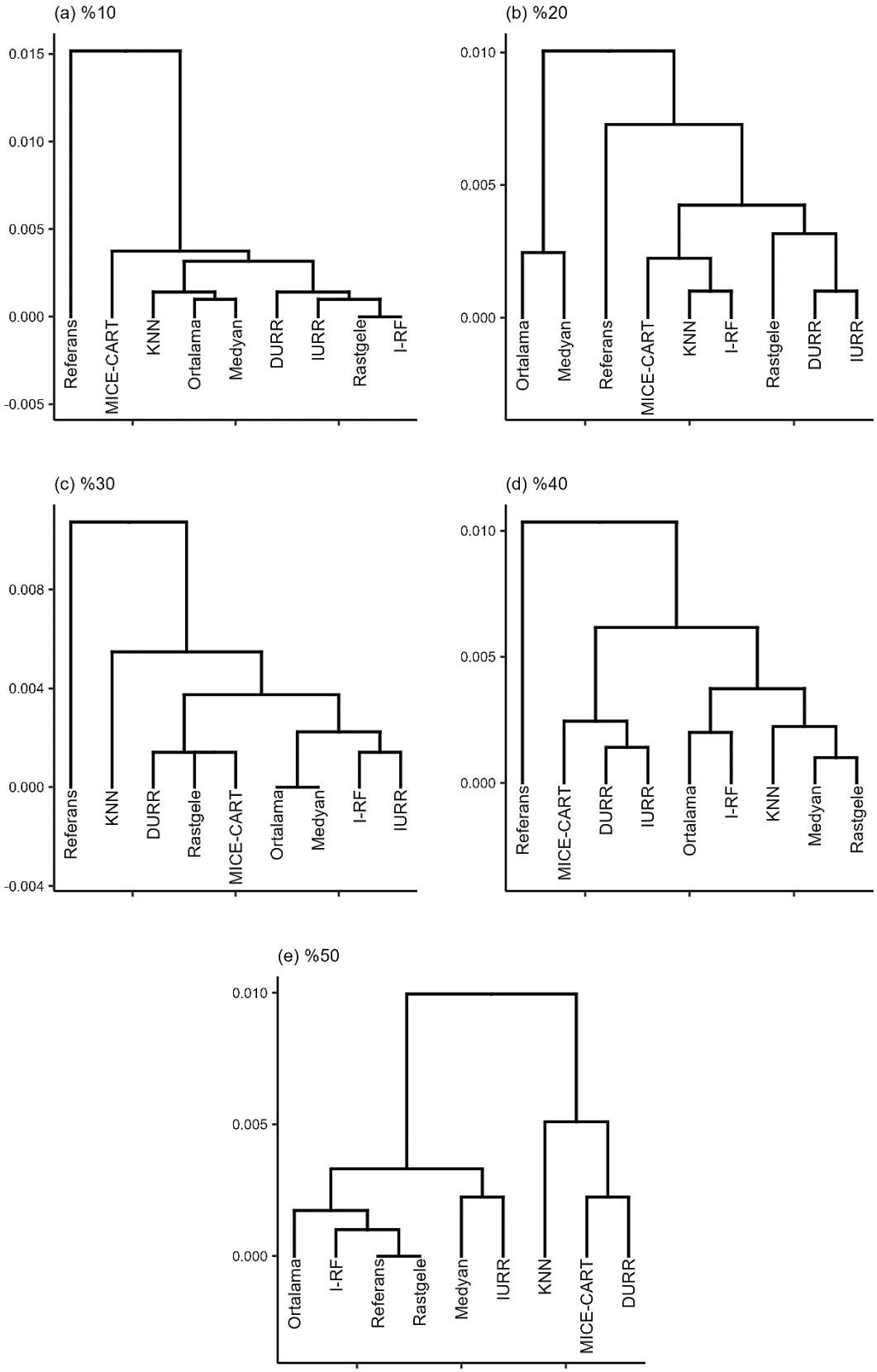

**Şekil 27.** Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin dengeli doğruluk oranlarının orman grafiği




**Şekil 28.** Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin AUC değerlerinin orman grafiği




**Şekil 29.** Tamamen rastgele türetilen verilerde -0,5 ≤ r ≤ 0,5 aralığı için yöntemlerin kappa değerlerinin orman grafiği




**Şekil 30.** Tamamen rastgele türetilen veri setlerinde $-0{,}5 \leq r \leq 0{,}5$ aralığı ve farklı eksik oranları için yöntemlerin dendrogram grafikleri

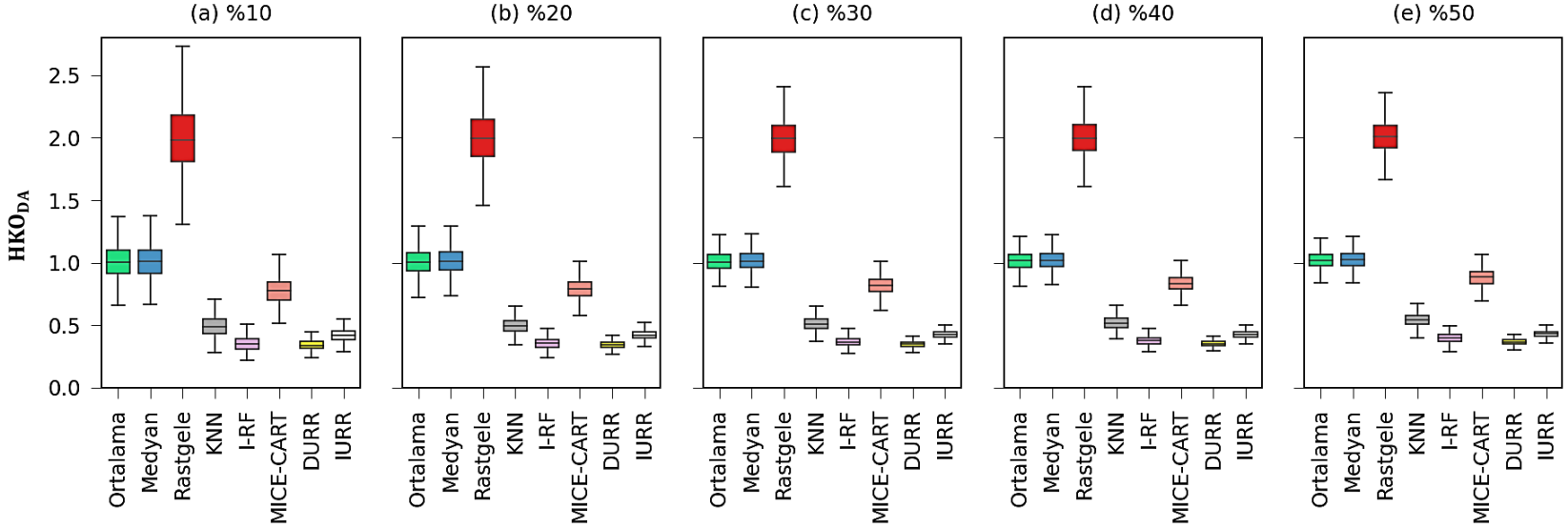
### 4.2.3. Tamamen Rastgele Türetilen Verilerde -0,8 ≤ r ≤ 0,8 Aralığı için Bulgular

$HKO_{DA}$ değerlerinin medyan değişim aralığı eksik oranı %10 için 0,338-1,984; %20 için 0,341-1,994; %30 için 0,346-1,993; %40 için 0,350-1,999; %50 için 0,365-2,009'dur. Değer atama yöntemlerinin $HKO_{DA}$ değerlerine ilişkin medyan değişim aralığı ise ortalama için 1,004-1,018; medyan için 1,011-1,026; rastgele için 1,984-2,009; KNN için 0,488-0,541; I-RF için 0,346-0,395; MICE-CART için 0,775-0,886; DURR için 0,338-0,365; IURR için 0,417-0,430'dur (Tablo 12).

Tüm değer atama yöntemlerinin $HKO_{DA}$ değerleri farklı eksik oranlarına göre incelendiğinde genel olarak rastgele değer atama hariç diğer yöntemlerin $HKO_{DA}$ değerleri arasında büyük bir fark olmadığı; tüm eksik oranları için KNN, I-RF, DURR ve IURR yöntemlerinin diğer yöntemlere göre daha düşük $HKO_{DA}$ değerlerine sahip olduğu; en yüksek $HKO_{DA}$ değerine sahip olan yöntemin ise rastgele değer atama yöntemi olduğu belirlenmiştir (Tablo 12 ve Şekil 31).

**Tablo 12.** Tamamen rastgele türetilen verilerde -0,8 ≤ r ≤ 0,8 aralığı için yöntemlerin $HKO_{DA}$ değerleri

| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| **$HKO_{DA}$** | **Ortalama** | 1,004 (0,913 - 1,099) | 1,007 (0,935 - 1,079) | 1,007 (0,957 - 1,065) | 1,014 (0,965 - 1,065) | 1,018 (0,973 - 1,064) |
| | **Medyan** | 1,011 (0,918 - 1,102) | 1,011 (0,940 - 1,084) | 1,013 (0,964 - 1,070) | 1,020 (0,969 - 1,072) | 1,026 (0,978 - 1,071) |
| | **Rastgele** | 1,984 (1,809 - 2,179) | 1,994 (1,855 - 2,145) | 1,993 (1,888 - 2,097) | 1,999 (1,901 - 2,103) | 2,009 (1,918 - 2,097) |
| | **KNN** | 0,488 (0,433 - 0,546) | 0,492 (0,451 - 0,533) | 0,509 (0,474 - 0,546) | 0,518 (0,483 - 0,554) | 0,541 (0,505 - 0,579) |
| | **I-RF** | 0,346 (0,310 - 0,389) | 0,354 (0,324 - 0,384) | 0,366 (0,340 - 0,393) | 0,375 (0,351 - 0,401) | 0,395 (0,369 - 0,422) |
| | **MICE-CART** | 0,775 (0,704 - 0,849) | 0,788 (0,734 - 0,844) | 0,816 (0,766 - 0,864) | 0,832 (0,788 - 0,881) | 0,886 (0,835 - 0,930) |
| | **DURR** | 0,338 (0,312 - 0,368) | 0,341 (0,322 - 0,362) | 0,346 (0,330 - 0,363) | 0,350 (0,335 - 0,367) | 0,365 (0,350 - 0,382) |
| | **IURR** | 0,417 (0,385 - 0,451) | 0,421 (0,396 - 0,445) | 0,423 (0,404 - 0,444) | 0,425 (0,405 - 0,444) | 0,430 (0,410 - 0,448) |




**Şekil 31**. Tamamen rastgele türetilen verilerde 0,8 ≤ r ≤ 0,8 aralığı için yöntemlerin $HKO_{DA}$ değerlerinin kutu grafiği

Referans veri setinden elde edilen dengeli doğruluk oranlarının medyanı 0,738 olarak elde edilmiştir. Atanmış veri setlerinden elde edilen sonuçlara dengeli doğruluk oranı değerlerinin medyan değişim aralığı eksik oranı %10 için 0,735-0,738; %20 için 0,739-0,742; %30 için 0,738-0,741; %40 için 0,737-0,741; %50 için 0,739-0,741'dir. Değer atama yöntemlerinin dengeli doğruluk oranlarına ilişkin medyan değişim aralığı ise ortalama ve medyan için 0,736-0,742; rastgele için 0,738-0,741; KNN için 0,735-0,741; I-RF ve IURR için 0,736-0,741; MICE-CART için 0,738-0,742; DURR için 0,737-0,741'dir (Tablo 13 ve Şekil 32).

Referans veri setinden elde edilen AUC değerlerinin medyanı 0,765 olarak elde edilmiştir. Atanmış veri setlerinden elde edilen sonuçlara göre eksik oranı %10, %20, %30, %40 ve %50 için AUC değerlerinin medyan değişim aralığı sırasıyla 0,765-0,767; 0,766-0,770; 0,766-0,771; 0,766-0,768 ve 0,767-0,770'dir. Değer atama yöntemlerinin AUC değerlerine ilişkin medyan değişim aralığı ise ortalama ve MICE-CART için 0,766-0,768; medyan için 0,765-0,769; rastgele için 0,767-0,770; KNN ve DURR için 0,766-0,770; I-RF için 0,766-0,769; IURR için 0,766-0,771'dir (Tablo 13 ve Şekil 33).

Referans veri setinden elde edilen kappa değerlerinin medyanı 0,474 olarak elde edilmiştir. Atanmış veri setlerinden elde edilen sonuçlara göre kappa değerlerinin medyan değişim aralığı eksik oranı %10 için 0,471-0,481; %20 için 0,480-0,495; %30 için 0,476-0,483; %40 için 0,474-0,486; %50 için 0,481-0,496'dır. Ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR yöntemlerinin kappa oranlarına ilişkin medyan değişim aralığı ise sırasıyla 0,475-0,492; 0,471-0,496; 0,476-0,487; 0,471-0,487; 0,473-0,487; 0,481-0,483; 0,474-0,486; 0,472-0,488'dir (Tablo 13 ve Şekil 34).

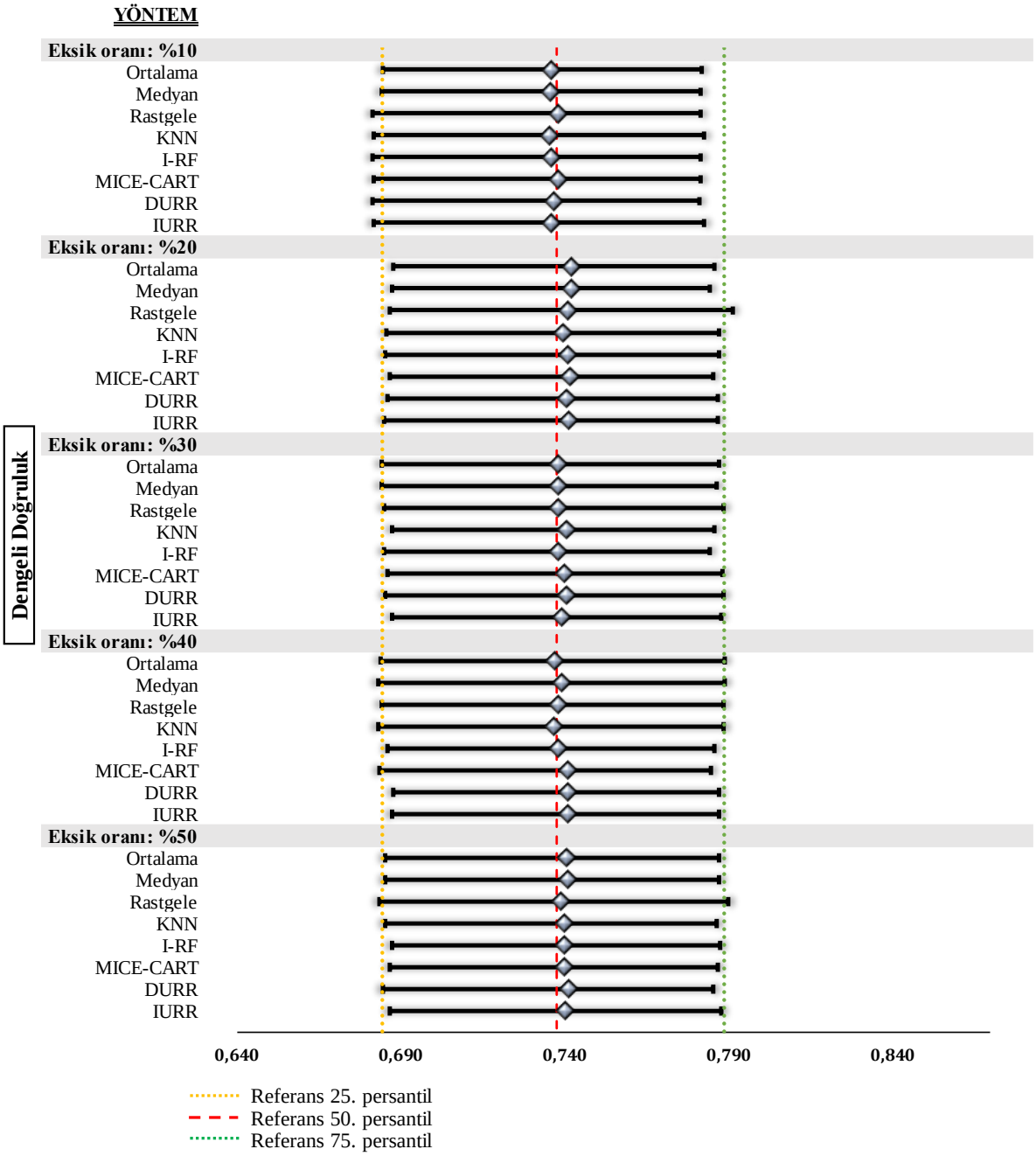
Tüm değer atama yöntemlerinin dengeli doğruluk oranı, AUC ve kappa değeri performansları değişen eksik oranına göre incelendiğinde, yöntemlerin eksik oranındaki artıştan etkilenmedikleri ve referans ile orta-yüksek düzeyde yakın tahminlerde bulunduğu ve birbirine yakın performans gösterdiği görülmektedir.

Dengeli doğruluk oranı, AUC ve kappa sonuçlarına göre uygulanan aşamalı kümeleme analizi ile elde edilen dendrogram grafikleri Şekil 35'de verilmiştir. Buna göre referans ile aynı küme içerisinde yer alan yöntemler eksik oranı %10 için DURR, ortalama ve rastgele yöntemleri; eksik oranı %20 için KNN, MICE-CART ve DURR yöntemleri; eksik oranı %30 için IURR, KNN, rastgele ve I-RF yöntemleri; eksik oranı %40 için KNN, ortalama ve rastgele yöntemleri; eksik oranı %50 için MICE-CART, IURR, DURR, rastgele, KNN ve I-RF yöntemleri olarak belirlenmiştir Ayrıca eksik oranı %10 için medyan, KNN, I-RF ve IURR yöntemlerinin; eksik oranı %20 için ortalama, medyan, rastgele, I-RF ve IURR yöntemlerinin;

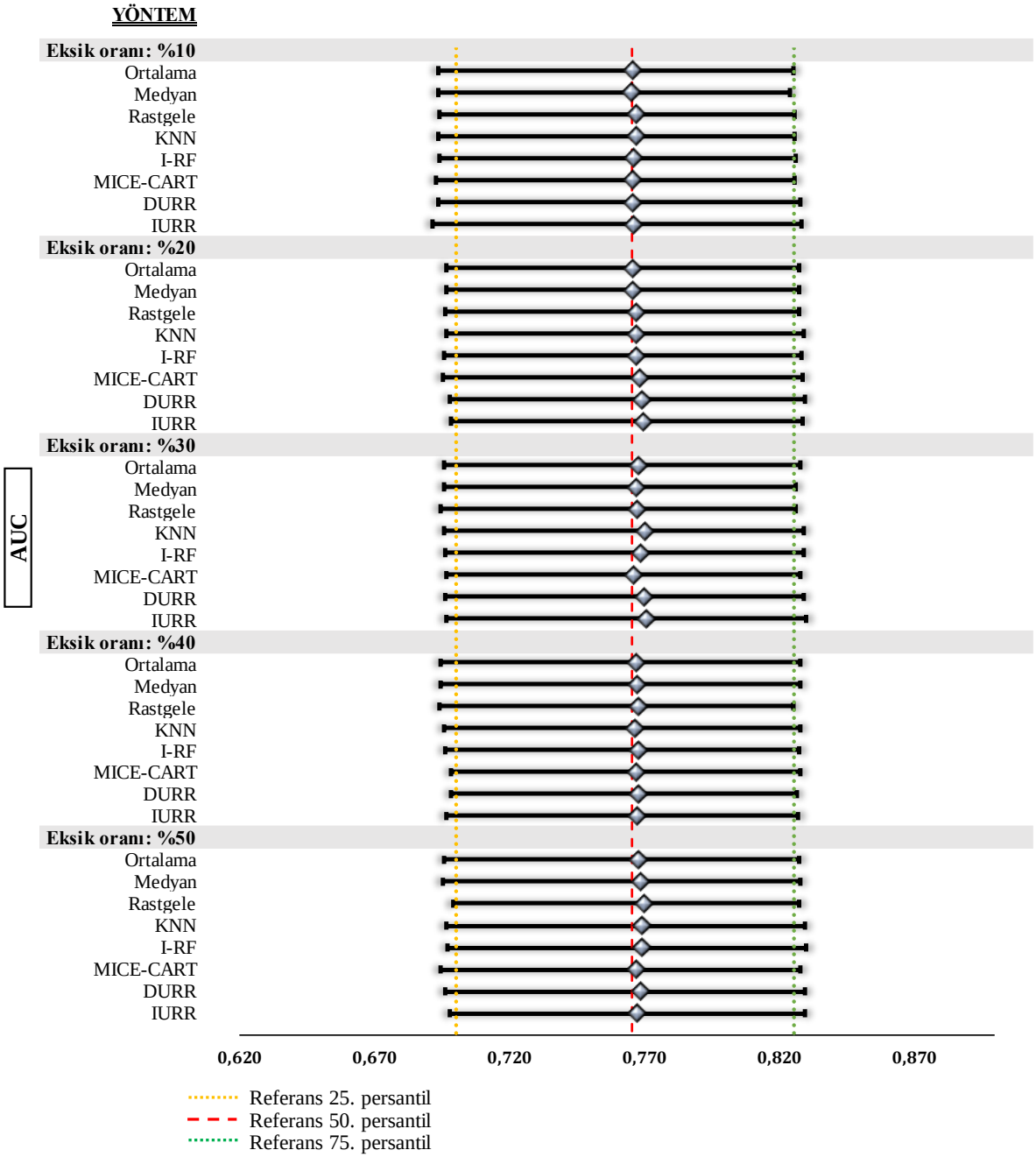
eksik oranı %30 için ortalama, medyan, MICE-CART ve DURR yöntemlerinin; eksik oranı %40 medyan, MICE-CART, DURR, IURR ve I-RF yöntemlerinin; eksik oranı %50 için ortalama ve medyan yöntemlerinin benzer performans gösterip ayrı bir küme oluşturdukları belirlenmiştir (Şekil 35).

**Tablo 13.** Tamamen rastgele türetilen verilerde -0,8 ≤ r ≤ 0,8 aralığı için yöntemlerin dengeli doğruluk oranları, AUC ve kappa değerleri

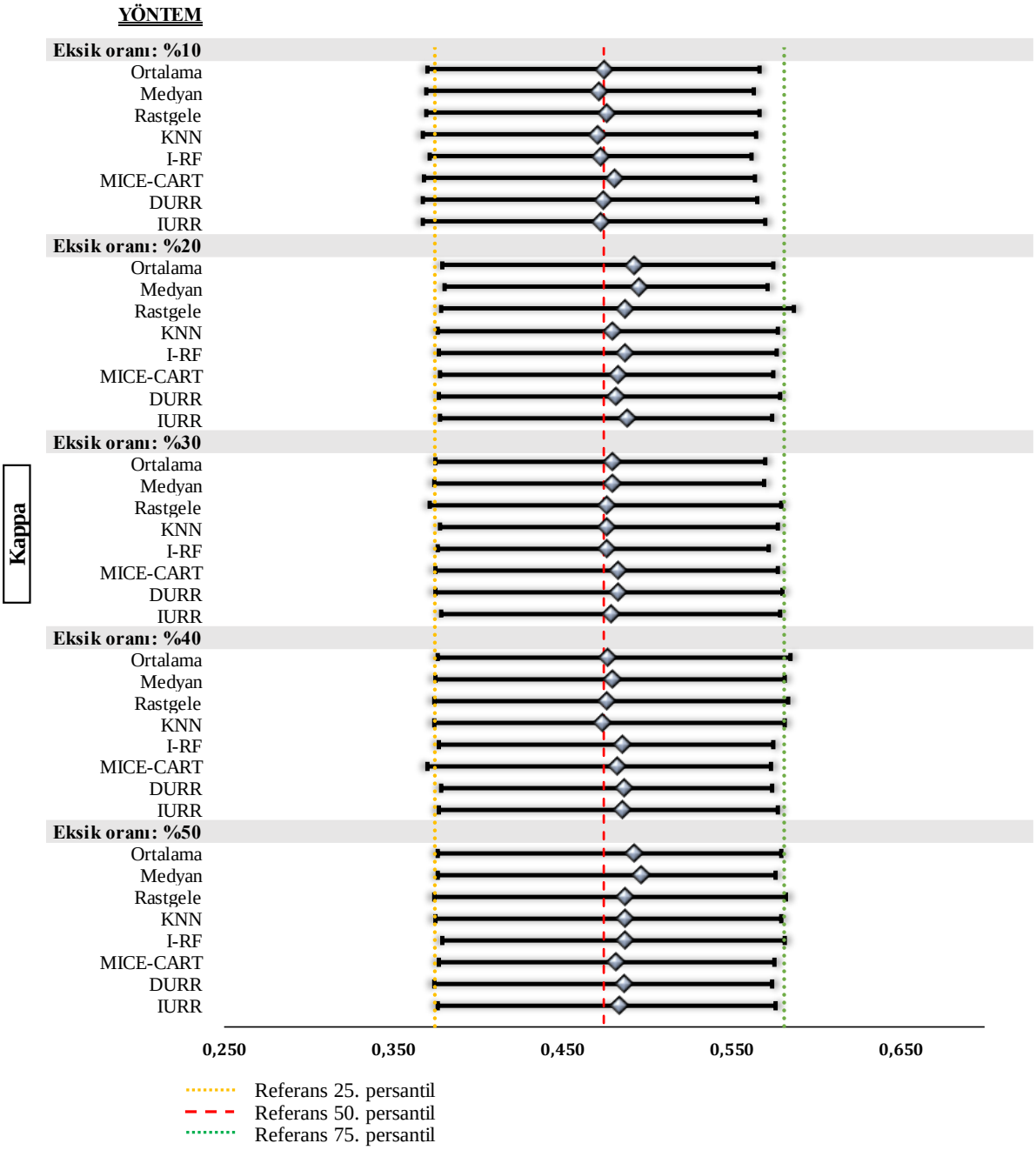
| | | EKSİK ORANI | | | | |
|---|---|---|---|---|---|---|
| | **YÖNTEM** | **%10** | **%20** | **%30** | **%40** | **%50** |
| **Dengeli Doğruluk** | **Referans** | 0,738 (0,685 - 0,789) | 0,738 (0,685 - 0,789) | 0,738 (0,685 - 0,789) | 0,738 (0,685 - 0,789) | 0,738 (0,685 - 0,789) |
| | **Ortalama** | 0,736 (0,684 - 0,782) | 0,742 (0,688 - 0,786) | 0,738 (0,684 - 0,787) | 0,737 (0,684 - 0,789) | 0,741 (0,685 - 0,787) |
| | **Medyan** | 0,736 (0,684 - 0,781) | 0,742 (0,687 - 0,784) | 0,738 (0,684 - 0,787) | 0,739 (0,683 - 0,789) | 0,741 (0,685 - 0,787) |
| | **Rastgele** | 0,738 (0,682 - 0,781) | 0,741 (0,687 - 0,791) | 0,738 (0,685 - 0,789) | 0,738 (0,684 - 0,789) | 0,739 (0,683 - 0,790) |
| | **KNN** | 0,735 (0,682 - 0,783) | 0,739 (0,686 - 0,787) | 0,741 (0,687 - 0,786) | 0,737 (0,683 - 0,789) | 0,740 (0,685 - 0,787) |
| | **I-RF** | 0,736 (0,682 - 0,781) | 0,741 (0,685 - 0,787) | 0,738 (0,685 - 0,784) | 0,738 (0,686 - 0,786) | 0,740 (0,688 - 0,787) |
| | **MICE-CART** | 0,738 (0,682 - 0,781) | 0,742 (0,687 - 0,785) | 0,740 (0,686 - 0,788) | 0,741 (0,683 - 0,785) | 0,740 (0,687 - 0,787) |
| | **DURR** | 0,737 (0,681 - 0,781) | 0,741 (0,686 - 0,787) | 0,741 (0,685 - 0,789) | 0,741 (0,688 - 0,787) | 0,741 (0,685 - 0,785) |
| | **IURR** | 0,736 (0,682 - 0,783) | 0,741 (0,685 - 0,787) | 0,739 (0,687 - 0,788) | 0,741 (0,688 - 0,787) | 0,740 (0,687 - 0,788) |
| **AUC** | **Referans** | 0,765 (0,700 - 0,826) | 0,765 (0,700 - 0,826) | 0,765 (0,700 - 0,826) | 0,765 (0,700 - 0,826) | 0,765 (0,700 - 0,826) |
| | **Ortalama** | 0,766 (0,694 - 0,825) | 0,766 (0,696 - 0,828) | 0,768 (0,696 - 0,828) | 0,767 (0,694 - 0,828) | 0,768 (0,696 - 0,827) |
| | **Medyan** | 0,765 (0,694 - 0,824) | 0,766 (0,696 - 0,828) | 0,767 (0,696 - 0,826) | 0,767 (0,694 - 0,828) | 0,769 (0,695 - 0,828) |
| | **Rastgele** | 0,767 (0,694 - 0,826) | 0,767 (0,696 - 0,828) | 0,768 (0,694 - 0,826) | 0,768 (0,694 - 0,826) | 0,770 (0,699 - 0,827) |
| | **KNN** | 0,767 (0,694 - 0,826) | 0,767 (0,696 - 0,829) | 0,770 (0,696 - 0,829) | 0,766 (0,696 - 0,828) | 0,769 (0,696 - 0,829) |
| | **I-RF** | 0,766 (0,694 - 0,826) | 0,767 (0,696 - 0,828) | 0,769 (0,696 - 0,829) | 0,768 (0,696 - 0,827) | 0,769 (0,697 - 0,830) |
| | **MICE-CART** | 0,766 (0,693 - 0,826) | 0,768 (0,695 - 0,829) | 0,766 (0,697 - 0,828) | 0,767 (0,698 - 0,828) | 0,767 (0,695 - 0,828) |
| | **DURR** | 0,766 (0,693 - 0,828) | 0,769 (0,698 - 0,829) | 0,770 (0,696 - 0,829) | 0,768 (0,698 - 0,827) | 0,769 (0,696 - 0,829) |
| | **IURR** | 0,766 (0,691 - 0,828) | 0,770 (0,698 - 0,829) | 0,771 (0,696 - 0,830) | 0,767 (0,697 - 0,827) | 0,767 (0,698 - 0,829) |
| **Kappa** | **Referans** | 0,474 (0,374 - 0,581) | 0,474 (0,374 - 0,581) | 0,474 (0,374 - 0,581) | 0,474 (0,374 - 0,581) | 0,474 (0,374 - 0,581) |
| | **Ortalama** | 0,475 (0,370 - 0,567) | 0,492 (0,379 - 0,575) | 0,480 (0,375 - 0,570) | 0,477 (0,376 - 0,585) | 0,492 (0,376 - 0,579) |
| | **Medyan** | 0,471 (0,369 - 0,563) | 0,495 (0,380 - 0,571) | 0,480 (0,374 - 0,569) | 0,480 (0,375 - 0,581) | 0,496 (0,376 - 0,576) |
| | **Rastgele** | 0,476 (0,369 - 0,566) | 0,487 (0,378 - 0,587) | 0,476 (0,371 - 0,580) | 0,476 (0,374 - 0,584) | 0,487 (0,374 - 0,582) |
| | **KNN** | 0,471 (0,368 - 0,565) | 0,480 (0,376 - 0,577) | 0,476 (0,377 - 0,577) | 0,474 (0,374 - 0,581) | 0,487 (0,375 - 0,579) |
| | **I-RF** | 0,473 (0,371 - 0,562) | 0,487 (0,377 - 0,577) | 0,476 (0,376 - 0,572) | 0,486 (0,377 - 0,575) | 0,487 (0,379 - 0,582) |
| | **MICE-CART** | 0,481 (0,368 - 0,564) | 0,483 (0,377 - 0,575) | 0,483 (0,374 - 0,578) | 0,482 (0,370 - 0,573) | 0,481 (0,377 - 0,576) |
| | **DURR** | 0,474 (0,368 - 0,565) | 0,481 (0,377 - 0,579) | 0,483 (0,375 - 0,580) | 0,486 (0,378 - 0,574) | 0,486 (0,374 - 0,574) |
| | **IURR** | 0,472 (0,368 - 0,570) | 0,488 (0,377 - 0,574) | 0,479 (0,378 - 0,579) | 0,485 (0,377 - 0,577) | 0,483 (0,376 - 0,576) |



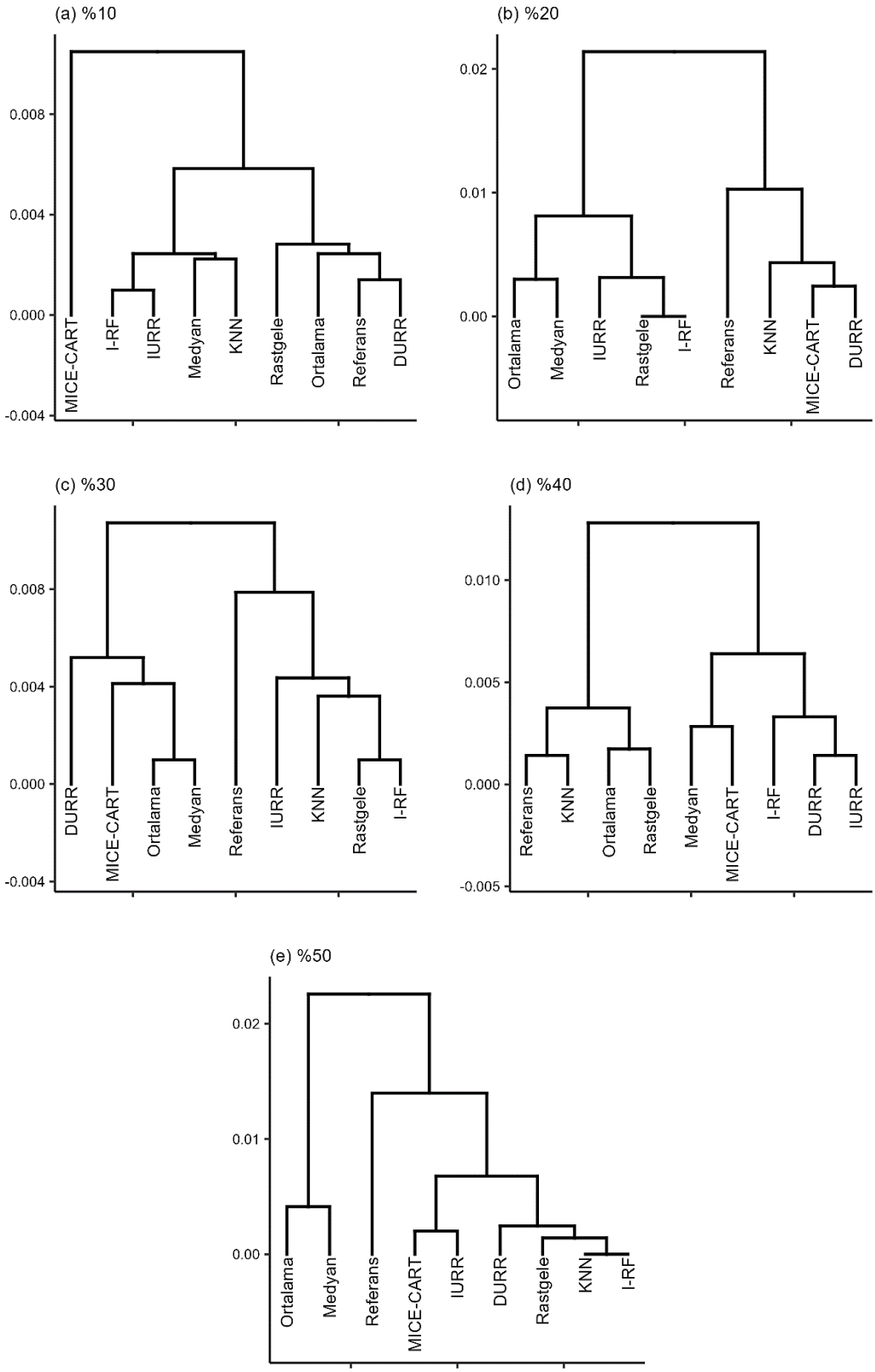

**Şekil 32.** Tamamen rastgele türetilen verilerde -0,8 ≤ r ≤ 0,8 aralığı için yöntemlerin dengeli doğruluk oranlarının orman grafiği




**Şekil 33.** Tamamen rastgele türetilen verilerde $-0{,}8 \leq r \leq 0{,}8$ aralığı için yöntemlerin AUC değerlerinin orman grafiği

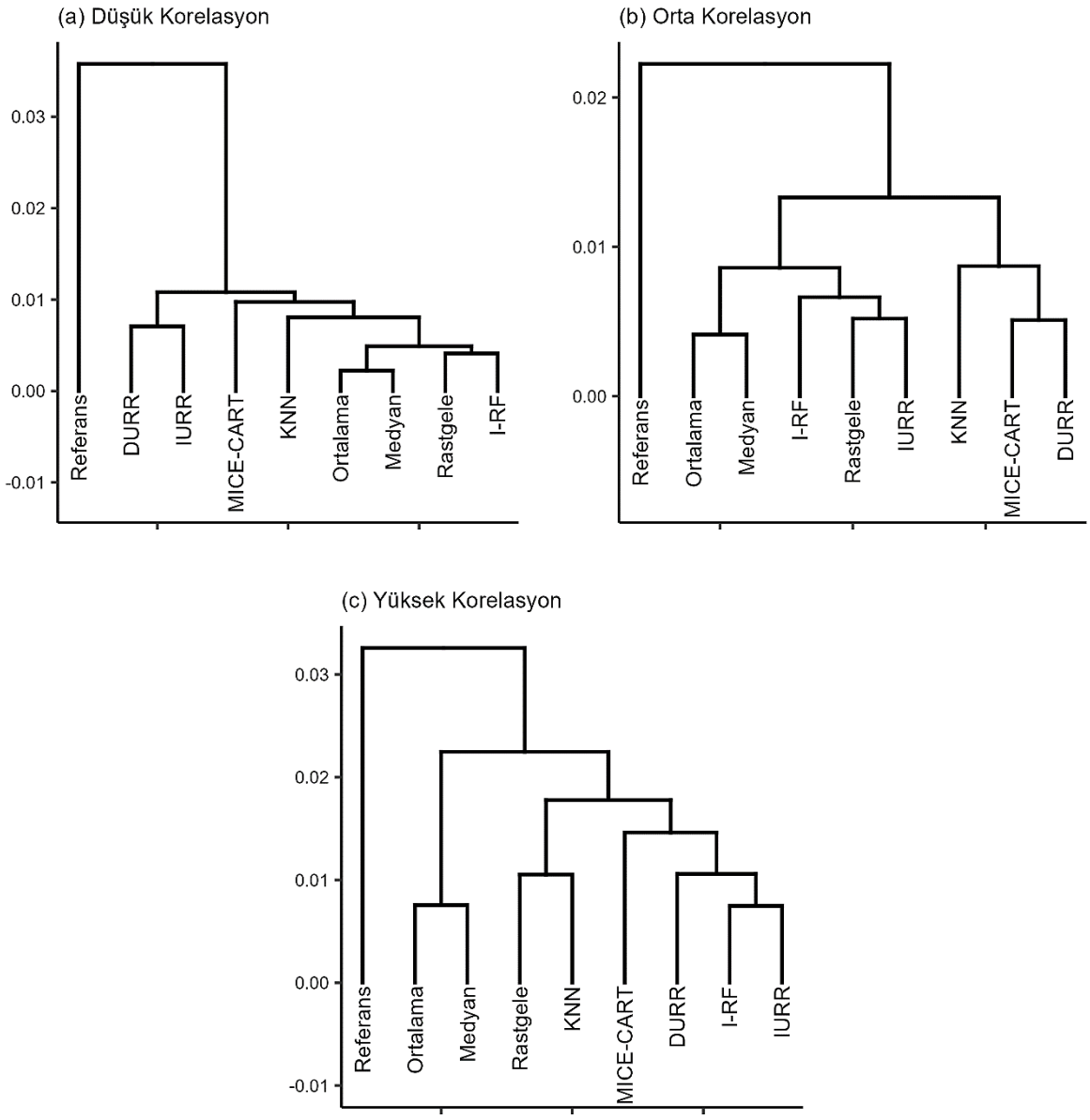



**Şekil 34.** Tamamen rastgele türetilen verilerde $-0,8 \leq r \leq 0,8$ aralığı için yöntemlerin kappa değerlerinin orman grafiği




**Şekil 35.** Tamamen rastgele türetilen veri setlerinde $-0{,}8 \leq r \leq 0{,}8$ aralığı ve farklı eksik oranları için yöntemlerin dendrogram grafikleri

Korelasyon katsayısı $-0,1 \leq r \leq 0,1$; $-0,5 \leq r \leq 0,5$ ve $-0,8 \leq r \leq 0,8$ için yöntemlerin tüm eksik oranlarındaki dengeli doğruluk oranı, AUC ve kappa sonuçları kullanılarak aşamalı kümeleme analizi yapılmış ve bu analiz ile elde edilen dendrogramlar Şekil 36'da verilmiştir. Buna göre tüm korelasyon düzeyleri için referans ile diğer yöntemlerin birbirinden ayrıldığı gözlenmiştir. Düşük korelasyon düzeyi için tüm değer atama yöntemleri birbirine yakın performans göstererek bir kümede toplanmıştır (Şekil 36a). Orta korelasyon düzeyi için ortalama, medyan, I-RF, rastgele ve IURR; KNN, MICE-CART ve DURR yöntemleri benzer performans göstererek ayrı iki kümede yer alırken daha sonra bu iki küme birleşerek tek bir küme oluşturmuştur (Şekil 36b). Yüksek korelasyon düzeyi için 2 farklı kümede yer alan yöntemler ortalama ve medyan; rastgele, KNN, MICE-CART, DURR, I-RF, ve IURR yöntemleri olurken daha sonra bu iki küme birleşerek tek bir küme içinde toplanmıştır (Şekil 36c).




**Şekil 36.** Tamamen rastgele türetilen veri setlerinde farklı korelasyon düzeyleri için yöntemlerin dendrogram grafikleri

# 5. TARTIŞMA

İstatistiksel analiz süreçlerini olumsuz etkilediği bilinen eksik değer içeren veri setleri son yıllarda hem teorik hem de pratik açıdan dikkate değer bir araştırma konusu haline gelmiştir. Sağlık alanındaki çalışmalarda özellikle hastalar hakkında çok fazla bilgi barındıran ve sıklıkla eksik değer içeren yüksek boyutlu verilerdeki bilgi kaybını önlemek ve verilerin doğru bir şekilde elde edilmesiyle oluşturulan modellerin performansını arttırmak için eksik verileri en az hata ile tahmin etmek büyük önem arz etmektedir. Schafer ve Graham (2002), gerçek veriler ile yaptıkları çalışmada eksik gözlemlerin veri setinden silinmesi halinde, özellikle eksik oranı arttıkça istatistiksel gücün azaldığını ve hatalı çıkarımlar elde edildiğini göstermişlerdir. Sun ve diğerleri (2010), gen ifade verileri ile yaptıkları çalışmada değer atama yöntemleri ile tamamlanan veri setleri ile yapılan istatistiksel analizlerin performanslarının, eksik birimleri veri setinden çıkararak yapılan analizlere göre önemli düzeyde yüksek olduğunu belirtmişlerdir. Bunun yanı sıra, eksik gözlemlerin veri setinden çıkarılması ya da uygun olmayan bir yöntem ile eksik verilerin tamamlanması; örneklem hacminde küçülme, istatistiksel gücün azalması, standart hataların artması, yanlış ve genellenemeyen çıkarımlar elde etme gibi sorunlara yol açtığı için eksik verilerin en az hata ile tahminde bulunan yöntemler ile tamamlandıktan sonra istatistiksel analizlerin uygulanması büyük önem arz etmektedir (Enders, 2022; Little ve Rubin, 2019; Schafer ve Olsen, 1998; Van Buuren, 2007).

Literatürde, eksik veri problemini ele alan birçok eksik veri değer atama yöntemi vardır. Bu yöntemlerin performansları veri yapısından etkilenmektedir. Van Buuren (2018), eksik veri problemini çözmek için en iyi yöntemin farklı veri yapılarına göre değişeceğini ve uygun yöntem seçiminin denemeler sonucu belirlenmesi gerektiğini belirtmiştir. Liao ve diğerleri (2014), farklı veri yapıları ve eksik oranlarında, 3 farklı senaryo için türetilmiş karma ilişkili verilerde MAR mekanizmalı eksik veri yapıları oluşturup, eksik verileri ortalama, MICE, I-RF ve KNN tabanlı yöntemler ile tamamlamışlardır. Veri yapısındaki farklılığın ve eksik veri oranındaki artışın yöntemlerin performansını etkilediğini göstermişlerdir. Çalışmamızda her iki simülasyon algoritması için de yöntemlerin $HKO_{DA}$ değerlerinin, değişen eksik oranına ve korelasyon yapısına göre farklılaştığı belirlenmiştir.

Değer atama yöntemlerinin performansı, eksik veri oranındaki değişimden etkilenebilmektedir. Mohammed ve diğerleri (2021), ortalama, medyan, rastgele, KNN ve MICE değer atama yöntemlerinin etkisini %5 ile %45 arasında değişen farklı eksik oranlarına

göre gerçek veri setlerinde incelemişler ve eksik oranındaki artışın yöntemlerin performansını düşürdüğünü bildirmişlerdir. D. J. Stekhoven ve Buhlmann (2012), farklı geçek veri setlerinde I-RF, KNN ve MICE yöntemlerinin performansını, %10, %20 ve %30 eksik oranları için incelemişler ve I-RF yönteminin tüm veri setlerinde diğer iki yönteme üstünlük sağladığını, ayrıca eksik oranındaki artışın yöntemlerin performansını düşürdüğünü bildirmişlerdir. Costantini ve diğerleri (2022), rastgele türettikleri karma ilişkili ve yüksek boyutlu verilerden, belirli bir değişken seti ile ilişkili olarak oluşturdukları 6 değişken için MAR mekanizmalı, %10 ve %30 oranlarında eksik değerler oluşturmuşlardır. Daha sonra bu veri setlerini DURR, IURR ve düzenlileştirme fonksiyona dayalı ve MICE tabanlı farklı yöntemler ile tamamlamışlardır. Elde ettikleri sonuçlara göre IURR yönteminin diğer yöntemlere göre performansının daha iyi olduğunu ve artan eksik oranının yöntemlerin performansını düşürdüğünü bildirmişlerdir. Eksik oranındaki değişimin yanı sıra, yapılan bazı çalışmalarda değişkenler arasındaki ilişkinin yöntemlerin performansına etkisi de incelenmiştir. Zhao ve Long (2016), aralarında korelasyon olmayan, orta ve yüksek korelasyonlu (0; 0,5; 0,9), n=100 birimden oluşan yüksek boyutlu veriler türetmişler; bu verilerden rastgele olarak seçtikleri 50 değişken ile ilişkili olarak oluşturdukları bir değişkende MAR mekanizmalı eksik değerler oluşturmuşlardır. Eksik değerli veri setlerini lasso ve elastik net düzenlileştirme fonksiyonlarına dayalı DURR ve IURR eksik değer atama yöntemleri ve düzenlileştirilmiş regresyona dayalı farklı yöntemler ile tamamladıktan sonra bu yöntemlerin performanslarını karşılaştırmışlardır. Elde ettikleri sonuçlara göre lasso düzenlileştirme fonksiyonuna dayalı DURR ve IURR yöntemlerinin diğer yöntemlere göre daha iyi performans gösterdiğini bildirmişlerdir. Ayrıca korelasyon düzeyindeki artışın yöntemlerin genel performansını arttırdığını, ancak performans sıralamasını değiştirmediğini bildirmişlerdir. Y. Deng ve diğerleri (2016), benzer bir çalışmada düşük, orta ve yüksek korelasyonlu (0,1; 0,5; 0,9), n=100 birimden oluşan yüksek boyutlu veriler türetmişlerdir. Bu veriler arasından rastgele 20 değişken belirleyerek bu değişken seti ile doğrusal olarak ilişkilendirilen 3 değişken için MAR mekanizmalı eksik değerler oluşturmuşlardır. Eksik değerli veri setlerini KNN algoritmasından geliştirilen 2 farklı yöntem, RF tabanlı MICE (MICE-RF), DURR ve IURR yöntemlerini kullanarak tamamladıktan sonra bu yöntemlerin performanslarını karşılaştırmışlardır. Elde ettikleri sonuçlara göre IURR yönteminin en iyi tahmini verdiğini ve bu yöntemi DURR yönteminin takip ettiğini bildirmişlerdir. Ayrıca korelasyon düzeylerinin yöntemlerin tahmin performanslarını etkilemesine rağmen, sıralamasını değiştirmediğini bildirmişlerdir. Zahid ve diğerleri (2021), $p < n$ ve $p \geq n$ için orta ve yüksek korelasyonlu (0,5; 0,8), n=100 birimden oluşan veriler türettikleri çalışmalarında, rastgele seçtikleri 10 değişkende farklı eksik veri oranlarında, MAR

mekanizmalı eksik değerler oluşturmuşlardır. Eksik değerli veri setlerini, istatistiksel paket programlarda yaygın olarak kullanılan çoklu değer atama yöntemleri ve düzenlileştirilmiş regresyona dayalı yöntemler ile tamamladıktan sonra bu yöntemlerin performansını karşılaştırmışlardır. Elde ettikleri sonuçlara göre yüksek boyutlu olan ve olmayan verilerde düzenlileştirilmiş regresyona dayalı yöntemlerin, diğer yöntemlere göre daha iyi performans gösterdiğini ve korelasyon düzeyindeki değişimin yöntemlerin performansını etkilemediğini, ancak eksik oranındaki artışın yöntemlerin performansını düşürdüğünü bildirmişlerdir. Çalışmamızda eksik değerlerin ortaya çıkmasında, değişkenler arasındaki ilişkilerin farklı olması amaçlandığı için sabit korelasyon yapıları kullanılmamış; düşük, orta ve yüksek korelasyonlu, n=150 birim, p=500 değişkenden oluşan yüksek boyutlu veriler türetilmiştir. Birinci simülasyon algoritmasında, rastgele olarak 50 değişken belirlenerek; bu 50 değişken ile ilişkili olarak oluşturulan 10 değişkende MAR mekanizmalı eksik değerler oluşturulmuştur. İkinci simülasyon algoritmasında ise, belirli bir değişken seti oluşturulmadan tüm değişkenler rastgele türetilmiştir. Birinci simülasyon algoritması için, eksik veri oranındaki artış tüm korelasyon düzeylerinde yöntemlerin tahmin hatalarını artırırken, eksik verileri en az hata ile tahmin eden yöntemlerin DURR, IURR ve I-RF olduğu görülmüştür. İkinci simülasyon algoritmasında ise düşük korelasyon düzeyinde, eksik veri oranındaki artış ile birlikte DURR ve IURR dışındaki yöntemlerin tahmin hataları artmış; eksik verileri en az hata ile tahmin eden yöntem I-RF olmuştur. Orta korelasyon düzeyinde, eksik veri oranındaki artış ile birlikte IURR dışındaki yöntemlerin tahmin hataları artmış, I-RF yönteminin eksik verileri diğer yöntemlere göre daha az hata ile tahmin ettiği görülmüştür. Yüksek korelasyon düzeyinde ise eksik veri oranındaki artış ile birlikte tüm yöntemlerin tahmin hataları artmış, eksik verileri en az hata ile tahmin eden yöntemler DURR, IURR ve I-RF olmuştur. Ayrıca yapılan denemeler sonucu daha kararlı sonuçlar vermesi ve uygulama kolaylığı nedeniyle, DURR ve IURR yöntemleri için lasso düzenlileştirme fonksiyonu kullanılmıştır.

Eksik değer atama yöntemlerinin sınıflandırma performansına etkisini inceleyen çalışmalarda ise yöntemler farklı performans ölçütlerine göre değerlendirilmiştir. Shrive ve diğerleri (2006), eksik değerli bir HIV veri setinin RF ile sınıflandırma performansına etkisini, I-RF ve yapay sinir ağları tabanlı 2 yöntem ile karşılaştırmışlardır. Elde ettikleri sonuçlara göre I-RF yöntemini diğer 2 yönteme göre oldukça etkili bulmuşlardır. Acuna ve Rodriguez (2004), eksik gözlemleri silme yöntemi ile ortalama, medyan ve KNN değer atama yöntemlerinin KNN ve doğrusal diskriminant analizi ile sınıflandırma performansına etkisini eksik oranı %1 ve %21 arasında değişen 11 gerçek veri seti üzerinde değerlendirmişlerdir. Elde ettikleri sonuçlara göre

düşük eksik oranlarında yöntemlerin birbirine yakın performans gösterdiğini; eksik oranı arttıkça KNN değer atama yönteminin diğer yöntemlere göre daha iyi performans gösterdiğini bildirmişlerdir. Çalışmamızda ise yüksek öğrenme hızı ve yüksek boyutlu verilerin boyut indirgemeden sınıflandırılmasına olanak sağladığı için sınıflandırma yöntemi olarak ELM tercih edilmiştir. Ayrıca literatürdeki çalışmalardan farklı olarak, değer atama yöntemlerinin sınıflandırma performansına etkileri yüksek boyutlu verilerde incelenmiştir. İlk simülasyon sonuçlarına göre düşük eksik oranlarında KNN, MICE-CART, I-RF, DURR ve IURR yöntemleri; yüksek eksik oranlarında DURR, IURR ve bunları takiben I-RF yöntemleri ön plana çıkmaktadır. İkinci simülasyon sonuçlarına göre ise yöntemlerin tahmin performanslarının benzer olduğu görülmüştür.

# 6. SONUÇ VE ÖNERİLER

Türetilmiş veriler ile 1000 döngü olarak gerçekleştirilen çalışmamızda n=150 gözlemden oluşan iki kategorili bağımlı değişken ve korelasyon katsayısı birinci durumda $-0,1 \leq r \leq 0,1$; ikinci durumda $-0,5 \leq r \leq 0,5$; üçüncü durumda $-0,8 \leq r \leq 0,8$ arasında değişen karma ilişkili, p=500 değişkenden oluşan rastgele veriler türetildi. Bağımsız değişkenlerden %10, %20, %30, %40 ve %50 oranlarında MAR mekanizmalı eksik değerler oluşturuldu. Daha sonra eksik veri değer atama yöntemlerinden; ortalama, medyan, rastgele, KNN, I-RF, MICE-CART, DURR ve IURR yöntemleri ile eksik değerler tahmin edildi. Yöntemlerin, orijinal gözlem değerlerini tahmin hataları $HKO_{DA}$ değerleri ile değerlendirilirken; sınıflandırma performansları ELM'den elde edilen dengeli doğruluk oranı, AUC ve kappa performans ölçütlerinin referans veri setinden elde edilen değerlere yakınlığına göre incelendi. Ayrıca referansa ve birbirine yakın performans gösteren yöntemler aşamalı kümeleme analizi ile belirlendi.

Yöntemlerin performanslarının veri yapısından etkilendiği belirlendi. Eksik değerli değişkenlerin veri setindeki belirli bir grup değişkenin doğrusal kombinasyonundan türetildiği birinci simülasyon algoritmasında, yöntemlerin eksik değer tahmin başarısı ve sınıflandırma performansına etkisinin eksik veri oranına ve korelasyon düzeyine göre değişiklik gösterdiği belirlendi. Buna göre genel olarak korelasyon düzeyi arttıkça yöntemlerin performanslarının yükseldiği, referansa ve birbirine daha yakın tahminlerde bulunduğu ancak yöntemlerin performans sıralamasının değişmediği görüldü. Ayrıca eksik veri oranı arttıkça, her üç korelasyon düzeyi için de yöntemlerin performanslarının düştüğü gözlendi. Tüm bağımsız değişkenlerin çok değişkenli standart normal dağılımdan türetildiği ikinci simülasyon algoritmasında ise yöntemlerin performanslarının korelasyon düzeyi ve eksik veri oranındaki artıştan genel olarak etkilenmediği gözlendi.

Birinci simülasyon algoritmasında; dengeli doğruluk oranı, AUC ve kappa sonuçlarına göre uygulanan aşamalı kümeleme analizinde tüm korelasyon düzeyleri için I-RF, MICE-CART, DURR ve IURR yöntemlerinin benzer performans gösterdiği ve referans ile aynı kümede bulundukları; ortalama, medyan ve rastgele yöntemlerinin de birbirlerine benzer performans gösterirken referansa daha uzak oldukları gözlenmiştir. KNN yönteminin ise $-0,1 \leq r \leq 0,1$ ve $-0,8 \leq r \leq 0,8$ aralıkları için düşük eksik oranlarında performansının arttığı görülmüştür. Ayrıca eksik oranı %50 olduğunda $-0,1 \leq r \leq 0,1$ ve $-0,5 \leq r \leq 0,5$ aralıkları için referansa en yakın yöntemler DURR ve IURR olurken; $-0,8 \leq r \leq +0,8$ aralığı için I-RF ve

MICE-CART yöntemlerinin de performanslarının DURR, IURR ve referansa yaklaştığı belirlenmiştir. Farklı korelasyon düzeyleri için değerlendirildiğinde düşük ve orta korelasyon düzeylerinde referansın DURR ve IURR yöntemleri ile bir küme oluşturduğu; yüksek korelasyon düzeyinde referansın DURR ve IURR yöntemlerine ek olarak I-RF ve MICE-CART yöntemleri ile aynı kümede yer aldığı tespit edilmiştir. Verilerin tamamen rastgele türetildiği ikinci algoritmada ise $-0,1 \leq r \leq 0,1$ aralığı için tüm yöntemlerin referanstan ayrı bir şekilde kümelendiği; $-0,5 \leq r \leq 0,5$ aralığı için, %50 eksik oranına kadar yöntemlerin referanstan ayrı bir şekilde kümelendiği, %50 eksik oranında referansın ortalama, rastgele ve I-RF yöntemleri ile bir küme oluşturduğu görülmüştür. $-0,8 \leq r \leq 0,8$ aralığında ise referansın %10 eksik oranında ortalama, rastgele ve DURR yöntemleriyle; %20 eksik oranında KNN, MICE-CART ve DURR yöntemleriyle; %30 eksik oranında rastgele, KNN, I-RF ve IURR yöntemleriyle; %40 eksik oranında ortalama, rastgele ve KNN yöntemleriyle bir küme oluşturduğu; %50 eksik oranında ise referansın tüm yöntemlerden ayrıldığı görülmüştür. Farklı korelasyon düzeyleri için referansın değer atama yöntemlerinden ayrıldığı ve yöntemlerin genel olarak performanslarının birbirine yakın olduğu görülmüştür.

Sonuç olarak; birinci simülasyon algoritmasında ve her üç korelasyon düzeyinde, $HKO_{DA}$ değerleri açısından değerlendirildiğinde orijinal gözlem değerlerini tahmin başarısı en yüksek olan yöntemlerin DURR, IURR ve I-RF olduğu belirlendi. Sınıflandırma performansı açısından ise her üç korelasyon düzeyinde, dengeli doğruluk oranı, AUC ve kappa ölçütlerine göre; %10, %20 ve %30 eksik oranları için I-RF, MICE-CART, DURR ve IURR yöntemlerinin; %40 eksik oranı için DURR ve IURR yöntemlerini takiben I-RF ve MICE-CART yöntemlerinin ön plana çıkan yöntemler olduğu; %50 eksik oranında ise DURR ve IURR yöntemlerinin diğer yöntemlere üstünlük sağladığı görülmüştür. Buna ek olarak $-0,1 \leq r \leq 0,1$ aralığı için %10 ve %20 eksik oranlarında; $-0,8 \leq r \leq 0,8$ aralığı için ise %10 eksik oranında KNN yönteminin de iyi performans gösterdiği belirlenmiştir. İkinci simülasyon algoritmasında ise, orijinal gözlem değerlerini tahmin performansı açısından, tüm eksik veri oranlarında, $-0,1 \leq r \leq 0,1$ aralığı için diğer yöntemlere göre daha düşük $HKO_{DA}$ değerine sahip olmaları nedeniyle ortalama, medyan, KNN, I-RF ya da DURR yöntemleri; $-0,5 \leq r \leq 0,5$ aralığı için ortalama, medyan, KNN, I-RF ya da DURR yöntemleri; $-0,8 \leq r \leq 0,8$ aralığı için KNN, I-RF, DURR ve IURR yöntemleri tercih edilebilir. Sınıflandırma performansına etki açısından ise $-0,1 \leq r \leq 0,1$ ve $-0,5 \leq r \leq 0,5$ aralıkları için yöntemler arasında bir üstünlük görülmezken; $-0,8 \leq r \leq 0,8$ aralığı için düşük eksik oranlarında ortalama, rastgele, KNN, MICE-CART ve DURR yöntemlerinden biriyle; bu yöntemlere ek olarak eksik oranı arttıkça IURR yöntemiyle de eksik veriler tahmin edilebilir.

%50 eksik oranında ise performansları birbirine yakın olduğu için çalışmamızda kullanılan yöntemlerden herhangi biri tercih edilebilir.

Bu çalışma; farklı veri yapıları, korelasyon düzeyleri ve eksik veri oranları için yüksek boyutlu verilerde değer atama yöntemlerinin kullanımı ve performanslarının değerlendirilmesi açısından literatüre önemli bilimsel katkıda bulunmaktadır. Özellikle sağlık alanında çoğunlukla yüksek boyutlu verilerde karşılaşılması kaçınılmaz olan eksik veri problemi için geliştirilen DURR ve IURR gibi yöntemlerin sınıflandırma performansına etkisini, literatürde yaygın olarak kullanılan basit ve gelişmiş yöntemlerle kıyaslayan başka bir çalışma bulunmaması çalışmamızı özgün kılan önemli bir noktadır. Eksik veriler ile sıklıkla karşılaşılan sağlık alanında, istatistiksel analiz süreçlerindeki başarıyı arttırmak ve veri kaybını önlemek için eksik veri problemi ayrıntılı olarak ele alınmalıdır. Özellikle veriler tamamen rastgele türetildiğinde, çalışmamızda kullanılan yöntemlerin sınıflandırma performansına etkisi değişkenler arasındaki ilişkiden ve eksik oranından etkilenmemektedir. Bu senaryoda, yöntemler birbirine yakın performans göstermektedir. Ancak eksik verili değişkenlerin veri setindeki belirli bir değişken seti ile ilişkili olduğu senaryoda, özellikle DURR ve IURR yöntemleri ön plana çıkmaktadır ve bu yöntemler değişkenler arasındaki ilişki ve eksik veri oranındaki farklılıktan diğer yöntemlere göre daha az etkilenmektedir. Literatürde aynı amaçla kullanılan birçok yöntem olmasına rağmen teorik yapıları gereği bazı yöntemlerin yüksek boyutlu eksik veri problemini çözmede yetersiz kalmaları nedeniyle, yüksek boyutlu verilerde ortaya çıkan bilgi kaybının önlenmesi ve modellerin tahmin performansının arttırılması için veriye en uygun yöntemin seçilmesi önem kazanmaktadır.

# KAYNAKLAR


Acuna, E., & Rodriguez, C. (2004). *The treatment of missing values and its effect on classifier accuracy*. Paper presented at the Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004.

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Aleryani, A., Wang, W., & De La Iglesia, B. (2018). *Dealing with missing data and uncertainty in the context of data mining*. Paper presented at the Hybrid Artificial Intelligent Systems: 13th International Conference, HAIS 2018, Oviedo, Spain, June 20-22, 2018, Proceedings 13.

Alpar, R. (2010). *Uygulamalı istatistik ve geçerlik-güvenirlik: spor, sağlık ve eğitim bilimlerinden örneklerle*: Detay Yayıncılık.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational statistics & data analysis, 52*(4), 2249-2260.

Belli, E., & Vantini, S. (2022). Measure inducing classification and regression trees for functional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 15*(5), 553-569.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123-140.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32.

Breiman, L. (2017). *Classification and regression trees*: Routledge.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). *The balanced accuracy and its posterior distribution*. Paper presented at the 2010 20th international conference on pattern recognition.

Burgette, L. F., & Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology, 172*(9), 1070-1076. doi:10.1093/aje/kwq260

Cantaş Türkiş, F., Kurt Ömürlü, İ., & Türe, M. (2024). Survival Prediction with Extreme Learning Machine, Supervised Principal Components and Regularized Cox Models in

High-Dimensional Survival Data by Simulation. *Gazi University Journal of Science*, 1-1. doi:10.35378/gujs.1223015

Chen, H., Peng, J., Zhou, Y., Li, L., & Pan, Z. (2014). Extreme learning machine for ranking: generalization analysis and applications. *Neural Netw, 53*, 119-126. doi:10.1016/j.neunet.2014.01.015

Choudhury, A., & Kosorok, M. R. (2020). Missing data imputation for classification problems. *arXiv preprint arXiv:2002.10709*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

Costantini, E., Lang, K. M., Reeskens, T., & Sijtsma, K. (2022). High-dimensional imputation for the social sciences: a comparison of state-of-the-art methods. *arXiv preprint arXiv:2208.13656*.

Cui, L., Zhai, H., & Lin, H. (2019). A Novel Orthogonal Extreme Learning Machine for Regression and Classification Problems. *Symmetry, 11*(10), 1284.

De Brevern, A. G., Hazout, S., & Malpertuy, A. (2004). Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC bioinformatics, 5*, 1-12.

Deng, W., Zheng, Q., & Chen, L. (2009). *Regularized extreme learning machine.* Paper presented at the 2009 IEEE symposium on computational intelligence and data mining.

Deng, Y., Chang, C., Ido, M. S., & Long, Q. (2016). Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Sci Rep, 6*, 21689. doi:10.1038/srep21689

Dobbin, K. K., & Simon, R. M. (2007). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics, 8*(1), 101-117.

Doğaner, A. (2020). *Topluluk öğrenme yöntemleri ile renal hücreli karsinom'un tahmin edilmesi*. Doktora Tezi, İnönü Üniversitesi Sağlık Bilimleri Enstitüsü, Malatya.

Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational statistics & data analysis, 72*, 92-104.

Enders, C. K. (2022). *Applied missing data analysis*: Guilford Publications.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874.

Fox-Wasylyshyn, S. M., & El-Masri, M. M. (2005). Handling missing data in self-report measures. *Research in nursing & health, 28*(6), 488-495.

García-Laencina, P. J., Sancho-Gómez, J.-L., Figueiras-Vidal, A. R., & Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing, 72*(7-9), 1483-1493.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention science, 8*, 206-213.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence, 20*(8), 832-844.

Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician, 55*(3), 244-254.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2004). *Extreme learning machine: a new learning scheme of feedforward neural networks.* Paper presented at the 2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541).

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing, 70*(1-3), 489-501.

Huang, G., Huang, G.-B., Song, S., & You, K. (2015). Trends in extreme learning machines: A review. *Neural Networks, 61*, 32-48.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence, 33*(10), 913-933.

Kalton, G. (1986). The treatment of missing survey data. *Survey methodology, 12*, 1-16.

Kasun, L. L., Yang, Y., Huang, G. B., & Zhang, Z. (2016). Dimension Reduction With Extreme Learning Machine. *IEEE Trans Image Process, 25*(8), 3906-3918. doi:10.1109/TIP.2016.2570569

Khan, S., Rahmani, H., Shah, S. A. A., & Bennamoun, M. (2018). A guide to convolutional neural networks for computer vision. *Synthesis lectures on computer vision, 8*(1), 1-207.

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of statistical software, 74*, 1-16.

Kurt Omurlu, I., Ture, M., Unubol, M., Katranci, M., & Guney, E. (2014). Comparing performances of logistic regression, classification & regression trees and artificial neural networks for predicting albuminuria in type 2 diabetes mellitus. *Int J Sci Basic Appl Res, 16*(1), 173-187.

Liang, N.-Y., Saratchandran, P., Huang, G.-B., & Sundararajan, N. (2006). Classification of mental tasks from EEG signals using extreme learning machine. *International journal of neural systems, 16*(01), 29-38.

Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F. C., & Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC bioinformatics, 15*(1), 1-12.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793): John Wiley & Sons.

Lu, J., Huang, J., & Lu, F. (2019). Distributed kernel extreme learning machines for aircraft engine failure diagnostics. *Applied Sciences, 9*(8), 1707.

McCleary, L. (2002). Using multiple imputation for analysis of incomplete data in clinical research. *Nursing Research, 51*(5), 339-343.

Mohammed, M., Zulkafli, H., Adam, M., Ali, N., & Baba, I. (2021). *Comparison of five imputation methods in handling missing data in a continuous frequency table.* Paper presented at the AIP Conference Proceedings.

Myers, W. R. (2000). Handling missing data in clinical trials: an overview. *Drug information journal: DIJ/Drug Information Association, 34*, 525-533.

Ozen, H., & Bal, C. (2020). A study on missing data problem in random Forest. *Osmangazi Tıp Dergisi, 42*(1), 103-109.

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, 157-166.

Raymond, M. R. (1986). Missing data in evaluation research. *Evaluation & the health professions, 9*(4), 395-420.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel psychology, 47*(3), 537-560.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Rubin, D. B. (1988). *An overview of multiple imputation.* Paper presented at the Proceedings of the survey research methods section of the American statistical association.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association, 91*(434), 473-489.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*: CRC press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods, 7*(2), 147.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research, 33*(4), 545-571.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal, 20*(1), 3-29.

Service, R. W. (2009). Book Review: Corbin, J., & Strauss, A.(2008). Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory . Thousand Oaks, CA: Sage. *Organizational Research Methods, 12*(3), 614-617.

Shrive, F. M., Stuart, H., Quan, H., & Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC medical research methodology, 6*, 1-10.

Stekhoven, D. J. (2011). Using the missForest package. *R package*, 1-11.

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112-118. doi:10.1093/bioinformatics/btr597

Sun, Y., Braga-Neto, U., & Dougherty, E. R. (2010). Impact of missing value imputation on classification for DNA microarray gene expression data—a model-based study. *EURASIP Journal on Bioinformatics and Systems Biology, 2009*, 1-17.

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 10*(6), 363-377.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research, 16*(3), 219-242.

Van Buuren, S. (2018). *Flexible imputation of missing data*: CRC press.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software, 45*, 1-67.

Wang, H., & Li, G. (2019). Extreme learning machine Cox model for high-dimensional survival analysis. *Statistics in medicine, 38*(12), 2139-2156.

Warrens, M. J. (2015). Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy, 5*(4), 1.

Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research, 10*(2).

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine, 30*(4), 377-399.

Wilkinson, L. (2004). Classification and regression trees. *Systat, 11*, 35-56.

Xing, E. P., Jordan, M. I., & Karp, R. M. (2001). *Feature selection for high-dimensional genomic microarray data.* Paper presented at the Icml.

Yin, X., Levy, D., Willinger, C., Adourian, A., & Larson, M. G. (2016). Multiple imputation and analysis for high-dimensional incomplete proteomics data. *Statistics in medicine, 35*(8), 1315-1326.

Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (Version 9.0). *SAS Institute Inc, Rockville, MD, 49*(1-11), 12.

Zahid, F. M., Faisal, S., & Heumann, C. (2021). Multiple imputation with compatibility for high-dimensional data. *PLoS One, 16*(7), e0254112. doi:10.1371/journal.pone.0254112

Zhang, R., Ye, B., & Liu, P. (2019). Dimension reduction of high-dimensional dataset with missing values. *Journal of Algorithms & Computational Technology, 13*, 1748302619867440.

Zhang, S., Gong, L., Zeng, Q., Li, W., Xiao, F., & Lei, J. (2021). Imputation of gps coordinate time series using missforest. *Remote Sensing, 13*(12), 2312.

Zhang, Z. (2016). Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of translational medicine, 4*(2), 30.

Zhao, Y., & Long, Q. (2016). Multiple imputation in the presence of high-dimensional data. *Stat Methods Med Res, 25*(5), 2021-2035. doi:10.1177/0962280213511027

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*: CRC press.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology), 67*(2), 301-320.