**T.C.**
**YILDIZ TEKNİK ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# VERİ MADENCİLİĞİ İLE MÜHENDİSLİK FAKÜLTESİ ÖĞRENCİLERİNİN OKUL BAŞARILARININ ANALİZİ

**AHMET SAYGILI**

**YÜKSEK LİSANS TEZİ**
**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**
**BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI**

**DANIŞMAN**
**YRD. DOÇ. DR. SONGÜL ALBAYRAK**

**İSTANBUL, 2013**

**T.C.**
**YILDIZ TEKNİK ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# Veri Madenciliği ile Mühendislik Fakültesi Öğrencilerinin Okul Başarılarının Analizi

Ahmet SAYGILI tarafından hazırlanan tez çalışması 27/05/2013 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

**Tez Danışmanı**

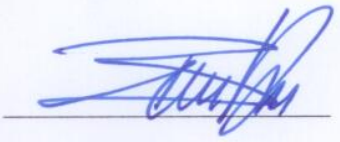
Yrd. Doç. Dr. Songül ALBAYRAK

Yıldız Teknik Üniversitesi

**Jüri Üyeleri**

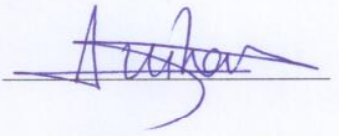
Doç. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi



Yrd. Doç. Dr. Nihan KAHRAMAN

Yıldız Teknik Üniversitesi



Yrd. Doç. Dr. Songül ALBAYRAK

Yıldız Teknik Üniversitesi

Bu çalışma, Yıldız Teknik Üniversitesi Bilimsel Araştırma Projeleri Koordinatörlüğü'nün 2012-04-01-KAP02 numaralı projesi ile desteklenmiştir.

# ÖNSÖZ

Bu çalışmada, bulanık kümeleme teknikleri ve katı kümeleme teknikleri kullanılarak mühendislik fakültesi lisans öğrencilerinin başarılarını etkileyen faktörlerin analizinin yapılması amaçlanmıştır. Bu amaç doğrultusunda ÖSYM kanalıyla Mühendislik Fakültesine gönderilen bilgiler ve öğrencinin okumakta olduğu bölümde ki mevcut başarısının yer aldığı verilerden oluşan bir veri seti oluşturulmuştur. Bunun yanında aile fertlerinin eğitim durumlarının ve mesleklerinin öğrencilerin başarısına olan etkisini tespit edebilmek için öğrencilerin belli bir bölümüne anket uygulanmıştır.

Çalışma boyunca değerli bilgileri ile beni yönlendiren ve aydınlatan tez danışmanım ve hocam Yrd. Doç. Dr. Songül ALBAYRAK'a teşekkürlerimi sunuyorum.

Bugünlere gelmem de en büyük emeğe sahip, yaşamım boyunca desteklerini ve dualarını esirgemeyen çok kıymetli anneme ve babama ayrı ayrı sonsuz teşekkürler ediyorum. Ayrıca tez süresi boyunca desteği ile beni cesaretlendiren eşime sevgilerimi ve teşekkürlerimi sunuyorum.

Mayıs, 2013

Ahmet SAYGILI

# İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ ..... viii

KISALTMA LİSTESİ ..... ix

ŞEKİL LİSTESİ ..... x

ÇİZELGE LİSTESİ ..... xii

ÖZET ..... xiv

ABSTRACT ..... xv

BÖLÜM 1

GİRİŞ ..... 1

1.1 Literatür Özeti ..... 1
1.2 Tezin Amacı ..... 10
1.3 Hipotez ..... 10

BÖLÜM 2

VERİ MADENCİLİĞİ ..... 13

2.1 Veri Madenciliği Süreci ..... 14
2.1.1 Problemin Tanımlanması (Problem Definition) ..... 15
2.1.2 Verinin Hazırlanması (Data Preparation) ..... 15
2.1.2.1 Veri Temizleme (Data Cleaning) ..... 15
2.1.2.2 Veri Bütünleştirme (Data Integration) ..... 16
2.1.2.3 Veri İndirgeme (Data Reduction) ..... 16
2.1.2.4 Veri Dönüştürme (Data Transformation) ..... 16
2.1.3 Modelin Kurulması ve Değerlendirilmesi (Evaluation of the Model) ..... 17
2.1.4 Modelin Kullanılması (Using the Model) ..... 18
2.2 Veri Madenciliğinin Farklı Disiplinlerle İlişkisi ..... 18
2.3 Veri Madenciliği Uygulama Alanları ..... 19
2.4 Veri Madenciliği Modelleri ..... 22
2.4.1 Sınıflama ve Regresyon (Classification and Regression) ..... 22
2.4.2 Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler (Association Rules and Sequential Patterns) ..... 24

BÖLÜM 3

UYGULAMADA KULLANILAN KÜMELEME VE SINIFLANDIRMA YÖNTEMLERİ ..... 26

3.1 Kümeleme (Clustering) ..... 26
3.1.1 Kümeleme Yöntemlerinde Uzaklık Ölçütleri ..... 27
3.1.1.1 Öklid Uzaklığı ..... 28
3.1.1.2 Manhattan Uzaklığı ..... 28
3.1.1.2 Minkowski Uzaklığı ..... 28
3.1.2 Katı Kümeleme Algoritmaları ..... 29
3.1.2.1 K-ortalamalar Algoritması ..... 29
3.1.2.1 K-medoids Algoritması ..... 30
3.1.3 Bulanık Kümeleme Algoritmaları ..... 32
3.1.3.1 Bulanık C-Ortalamalar Algoritması ..... 32
3.1.3.2 Gustafson-Kessel algoritması ..... 34
3.1.3.3 Gath-Geva algoritması ..... 36
3.1.4 Küme Geçerliliği (Cluster Validity) ..... 38
3.1.4.1 Bölünme Katsayısı (Partititon Coefficient) (PC) ..... 39
3.1.4.2 Sınıflandırma Entropisi (Classification Entropy) (CE) ..... 39
3.1.4.3 Bölümlendirme İndeksi (Partition Index) (SC) ..... 40
3.1.4.4 Ayırma İndeksi ( Seperation Index) (S) ..... 40
3.2 Karar Ağacı (Decision Trees) Sınıflandırma Yöntemi ..... 41

BÖLÜM 4

UYGULAMA ..... 43

4.1 Uygulamada Kullanılan Veri Madenciliği Programları ..... 43
4.1.1 SAS Enterprise Miner ..... 43
4.1.2 WEKA ..... 44
4.1.3 MATLAB ..... 44
4.2 Veri Setinin Tanımlanması ..... 45
4.2.1 Veri Ön İşleme ..... 47
4.3 Veri Setinin Analiz Edilmesi ..... 49
4.3.1 Yerleştirme Puanlarına ve Sıralarına Göre Öğrencilerin Dağılımı ..... 49
4.3.2 Yerleştikleri Kontenjan Tipine Göre Öğrencilerin Dağılımı ..... 50
4.3.3 Bölümlere Göre Öğrencilerin Dağılımı ..... 51
4.3.4 Cinsiyete Göre Öğrencilerin Dağılımı ..... 51
4.3.5 Genel Not Ortalamasına Göre Öğrencilerin Dağılımı ..... 52
4.3.6 Kredi ve Burs Durumlarına Göre Öğrencilerin Dağılımı ..... 52
4.3.7 Kayıt Yılına Göre Öğrencilerin Dağılımı ..... 53
4.3.8 Nüfusa Kayıtlı Oldukları Bölgelere Göre Öğrencilerin Dağılımı ..... 54
4.3.9 Öğretim Türüne Göre Öğrencilerin Dağılımı ..... 55
4.3.10 Okul Türlerine Göre Öğrencilerin Dağılımı ..... 55
4.3.11 Tercih Sıralarına Göre Öğrencilerin Dağılımı ..... 56
4.3.12 Yaşa Göre Öğrencilerin Dağılımı ..... 57
4.3.13 Bölüm - Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 57
4.3.14 Cinsiyet - Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 58
4.3.15 Kayıt Yılı - Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 58
4.3.16 Nüfusa Kayıtlı Bölge - Genel Not Ortalamasına Göre Öğrenci Dağılımı .. 59
4.3.17 Öğretim Türü - Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 59

4.3.18 Okul Türü - Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 60
4.3.19 Yerleştirme Sırası - Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 61
4.3.20 Bölüm-Nüfusa Kayıtlı Olduğu Bölgeye Göre Öğrenci Dağılımı ..... 61
4.3.21 Cinsiyet-Bölüme Göre Öğrenci Dağılımı ..... 62
4.3.22 Kayıt Yılı-Nüfusa Kayıtlı Olduğu Bölgeye Göre Öğrenci Dağılımı ..... 62
4.3.23 Burs Durumu - Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 63
4.3.24 Sınıfı - Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 63
4.3.25 Yerleştirme Sırası – Nüfusa Kayıtlı Olduğu Bölgeye Göre Öğrenci Dağılımı ..... 64
4.3.26 Yerleştirme Sırası - Bölümüne Göre Öğrenci Dağılımı ..... 65
4.3.27 Yerleştirme Sırası–Medeni Durum-Nüfusa Kayıtlı Olduğu Bölgeye Göre Öğrenci Dağılımı ..... 66
4.3.28 Genel Not Ortalaması-Yerleştiği Kontenjan Tipi-Yerleştirme Sırasına Göre Öğrenci Dağılımı ..... 67
4.3.29 Nüfusa Kayıtlı Olduğu Bölge-Cinsiyet-Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 67
4.3.30 Öğretim Türü-Bölüm- Genel Not Ortalamasına Göre Öğrenci Dağılımı .. 68
4.3.31 Kayıt Yılı-Nüfusa Kayıtlı Olduğu Bölge-Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 69
4.3.32 Yerleştirme Sırası –Burs Durumu- Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 69
4.4 Veri Setine Kümeleme Yöntemlerinin Uygulanması ..... 70
4.4.1 K-ortalamalar Yönteminin Verilere Uygulanması ..... 70
4.4.2 K-medoids Yönteminin Verilere Uygulanması ..... 76
4.4.3 Bulanık c-ortalamalar Yönteminin Verilere Uygulanması ..... 79
4.4.4 Gustafson Kessel Yönteminin Verilere Uygulanması ..... 83
4.4.5 Gath-Geva Yönteminin Verilere Uygulanması ..... 85
4.4.6 Geçerlilik Değerleri ile Yöntemlerin Karşılaştırılması ..... 87
4.5 Karar Ağacı Yönteminin Verilere Uygulanması ..... 88
4.6 Öğrenci Anketlerinin Değerlendirilmesi ..... 90
4.6.1 Anket Veri Setinin Analiz Edilmesi ..... 91
4.6.2 Anket Verilerinin K-Ortalamalar Yöntemi İle Kümelenmesi ..... 96
4.6.3 Anket Verilerinin K-Medoids Yöntemi ile Kümelenmesi ..... 98
4.6.4 Anket Verilerinin BCO Yöntemi İle Kümelenmesi ..... 100
4.6.5 Anket Verilerinin Gustafson-Kessel Yöntemi İle Kümelenmesi ..... 102
4.6.6 Anket Verilerinin Gath-Geva Yöntemi İle Kümelenmesi ..... 103

BÖLÜM 5

SONUÇ VE ÖNERİLER ..... 106

KAYNAKLAR ..... 110

ÖZGEÇMİŞ ..... 114

# SİMGE LİSTESİ

| | |
|---|---|
| $A$ | Simetrik ve pozitif tanımlı matris |
| $\alpha_i$ | i. kümenin ön seçilme olasılığı |
| $BYK_J$ | Her bir bölgenin yoğunluk katsayısı |
| $BN_J$ | j. bölgenin toplam nüfusu |
| $C$ | Küme sayısı |
| $c_{jih}$ | j. birey için i 'nin h'a katkısı |
| $d(x_i, x_j)$ | $x_i$ ve $x_j$ arasındaki uzaklık |
| $\varepsilon$ | İşlem bitirme kriteri |
| $F_i$ | i. kümenin kovaryans matrisi |
| $J$ | Amaç fonksiyonu |
| $KB_{ij}$ | i. kümedeki j. bölge öğrenci sayısı |
| $KBY_{ij}$ | i. kümedeki j. bölge yoğunluğu |
| $KBYK_{ij}$ | i. kümedeki j. bölgenin yoğunluk katsayısı |
| $KT_i$ | i. kümedeki toplam öğrenci sayısı |
| $m$ | Bulanıklaştırma katsayısı |
| $\mu_{i,k}$ | k. nesnenin i. kümeye üyelik derecesi |
| $N$ | Toplam nesne sayısı |
| $P_i$ | Önsel olasılık |
| $S$ | Benzerlik |
| $T_{ih}$ | Toplam yer değiştirme katsayısı |
| $TN$ | Tüm bölgelerin toplam nüfusu |
| $U$ | Fuzzy üyelik matrisi |
| $v_i$ | i. kümenin merkezi (prototipi) |
| $x_i$ | veri setindeki i. nesne |

# KISALTMA LİSTESİ

| | |
|---|---|
| BCO | Bulanık c-ortalamalar |
| CE | Sınıflandırma Entropisi |
| FCM | Fuzzy C-Means |
| GG | Gath-Geva |
| GNO | Genel Not Ortalaması |
| GK | Gustafson-Kessel |
| MATLAB | Matrix Laboratory |
| ÖSYM | Öğrenci Seçme Yerleştirme Merkezi |
| ÖSYS | Öğrenci Seçme Yerleştirme Sınavı |
| PC | Bölünme Katsayısı |
| S | Ayırma İndeksi |
| SAS | Statistical Analysis System |
| SC | Bölümlendirme İndeksi |
| SQL | Structured Query Language |
| SSE | Hata karelerinin toplamı |
| WEKA | Waikato Environment for Knowledge Analysis |

# ŞEKİL LİSTESİ

Sayfa


Şekil 1.1 Kullanılan yöntemlerin başarı oranları [2] ..... 3
Şekil 1.2 Kategorize edilmiş üniversiteye giriş notlarının öğrenciler üzerindeki dağılım grafiği [4] ..... 4
Şekil 1.3 Öğrenci ve öğrenci notları [10] ..... 10
Şekil 2.1 Veri Madenciliği Süreci [17] ..... 14
Şekil 2.2 Veri madenciliğinin ilişkili olduğu disiplinler [27] ..... 19
Şekil 2.3 Veri Madenciliği Modelleri ..... 22
Şekil 3.1 İki nokta arasındaki d uzaklığı ..... 28
Şekil 3.2 k-ortalamalar algoritmasının (a) k=2 için; (b)k=3 için ötelenişi [47] ..... 29
Şekil 3.3 BCO Algoritması ve GK Algoritması [56] ..... 34
Şekil 3.4 Gath-Geva algoritmasının sentetik veri seti için kümeleme sonuçları [58] 37
Şekil 3.5 Örnek bir karar ağacı [53] ..... 42
Şekil 4.1 SAS Enterprise Miner kullanıcı arayüzü ..... 43
Şekil 4.2 Weka kullanıcı arayüzü ..... 44
Şekil 4.3 MATLAB kullanıcı arayüzü ..... 45
Şekil 4.4 Öğrencilerin bölgelere göre dağılımı ..... 47
Şekil 4.5 Notların dönüştürülmeden önceki dağılımı ..... 48
Şekil 4.6 Notların dönüştürüldükten sonraki dağılımı ..... 49
Şekil 4.7 Yerleştirme puanlarına göre öğrencilerin dağılımı ..... 49
Şekil 4.8 Yerleştirme sıralarına göre öğrenci dağılımları ..... 50
Şekil 4.9 Öğrencilerin yerleştikleri kontenjan türleri ..... 50
Şekil 4.10 Bölümlere göre öğrenci dağılımları ..... 51
Şekil 4.11 Cinsiyete göre öğrenci dağılımları ..... 51
Şekil 4.12 Genel not ortalamasına göre öğrenci dağılımları ..... 52
Şekil 4.13 Katkı kredisi durumuna göre öğrenci dağılımları ..... 52
Şekil 4.14 Öğrenim kredisi durumuna göre öğrenci dağılımları ..... 53
Şekil 4.15 Burs durumuna göre öğrenci dağılımları ..... 53
Şekil 4.16 Kayıt yılına göre öğrenci dağılımları ..... 54
Şekil 4.17 Nüfusa kayıtlı olduğu bölgeye göre öğrenci dağılımları ..... 55
Şekil 4.18 Öğretim türüne göre öğrenci dağılımları ..... 55
Şekil 4.19 Okul türüne göre öğrenci dağılımları ..... 56
Şekil 4.20 Tercih sırasına göre öğrenci dağılımları ..... 56
Şekil 4.21 Yaşa göre öğrenci dağılımları ..... 57
Şekil 4.22 Bölüm- Genel not ortalamasına göre öğrenci dağılımları ..... 58
Şekil 4.23 Cinsiyet-Genel not ortalamasına göre öğrenci dağılımları ..... 58
Şekil 4.24 Kayıt yılı-Genel not ortalamasına göre öğrenci dağılımları ..... 59
Şekil 4.25 Nüfusa kayıtlı olduğu bölge-Genel not ortalamasına göre öğrenci dağılımları ..... 59

Şekil 4.26 Öğretim türü-Genel not ortalamasına göre öğrenci dağılımları ..... 60
Şekil 4.27 Okul Türü-Genel not ortalamasına göre öğrenci dağılımları ..... 60
Şekil 4.28 Yerleştirme sırası-Genel not ortalamasına göre öğrenci dağılımları ..... 61
Şekil 4.29 Bölüm-Nüfusa kayıtlı olduğu bölgeye göre öğrenci dağılımları ..... 62
Şekil 4.30 Cinsiyet-Bölüme göre öğrenci dağılımları ..... 62
Şekil 4.31 Kayıt yılı-Nüfusa kayıtlı olduğu bölgeye öğrenci dağılımları ..... 63
Şekil 4.32 Burs Durumu-Genel not ortalamasına göre öğrenci dağılımları ..... 63
Şekil 4.33 Sınıfı-Genel not ortalamasına göre öğrenci dağılımları ..... 64
Şekil 4.34 Yerleştirme sırası-Nüfusa kayıtlı olduğu bölgeye göre öğrenci dağılımları 65
Şekil 4.35 Yerleştirme sırası-Bölüme göre öğrenci dağılımları ..... 65
Şekil 4.36 Yerleştirme Sırası – Medeni Durum- Nüfusa Kayıtlı Olduğu Bölgeye Göre Öğrenci Dağılımı ..... 66
Şekil 4.37 Genel Not Ortalamasına-Yerleştiği Kontenjan Tipi-Yerleştirme Sırasına Göre Öğrenci Dağılımı ..... 67
Şekil 4.38 Nüfusa Kayıtlı Olduğu Bölge-Cinsiyet- Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 68
Şekil 4.39 Öğretim Türü-Bölüm- Genel Not Ortalamasına Göre Öğrenci Dağılımı..... 68
Şekil 4.40 Kayıt Yılı-Nüfusa Kayıtlı Olduğu Bölge-Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 69
Şekil 4.41 Yerleştirme Sırası –Burs Durumu- Genel Not Ortalamasına Göre Öğrenci Dağılımı ..... 70
Şekil 4.42 K-ortalamalar yöntemi için geçerlilik indeksleri ..... 71
Şekil 4.43 K-ortalamalar yöntemi GNO özelliği için Box-plot analizi ..... 76
Şekil 4.44 K-medoids yöntemi için geçerlilik indeksleri ..... 77
Şekil 4.45 K-medoids yöntemi GNO özelliği için Box plot analizi ..... 79
Şekil 4.46 Bulanık c-ortalamalar yöntemi için geçerlilik indeksleri ..... 80
Şekil 4.47 Bulanık c-ortalamalar yöntemi GNO özelliği için Box plot analizi ..... 82
Şekil 4.48 Gustafson-Kessel yöntemi için geçerlilik indeksleri ..... 83
Şekil 4.49 Gustafson-Kessel yöntemi GNO özelliği için Box plot analizi ..... 85
Şekil 4.50 Gath-Geva yöntemi için geçerlilik indeksleri ..... 85
Şekil 4.51 Gath-Geva yöntemi GNO özelliği için Box plot analizi ..... 87
Şekil 4.52 Karar Ağacı ..... 88
Şekil 4.53 Baba Eğitim Durumu-Öğrencinin Lise Mezuniyet Notu İlişkisi ..... 92
Şekil 4.54 Anne Eğitim Durumu- Öğrencinin Lise Mezuniyet Notu İlişkisi ..... 92
Şekil 4.55 Anne Eğitim Durumu- Öğrencinin Yerleştirme Sırası İlişkisi ..... 93
Şekil 4.56 Baba Eğitim Durumu- Öğrencinin Yerleştirme Sırası İlişkisi ..... 93
Şekil 4.57 Baba Eğitim Durumu- Öğrencinin Genel Not Ortalaması İlişkisi ..... 94
Şekil 4.58 Anne Eğitim Durumu- Öğrencinin Genel Not Ortalaması İlişkisi ..... 94
Şekil 4.59 Lise Mezuniyet Notu- Genel Not Ortalaması İlişkisi ..... 95
Şekil 4.60 Baba Eğitim Durumu-Kardeş-Okuyan Kardeş Sayısı İlişkisi ..... 95
Şekil 4.61 Anne Eğitim Durumu-Kardeş-Okuyan Kardeş Sayısı İlişkisi ..... 96
Şekil 4.62 GSM Operatörleri Dağılımı ..... 96

# ÇİZELGE LİSTESİ

Sayfa


Çizelge 1.1 Sınıflayıcıların doğruluk değerleri sonuçları [2] ..... 2
Çizelge 1.2 K-ortalamalar yöntemi sonuçları [4] ..... 4
Çizelge 1.3 Özelliklere göre kümelerin karakteristikleri [4] ..... 5
Çizelge 1.4 Yöntemlerin doğruluk oranları[5] ..... 6
Çizelge 1.5 Yöntemlerin doğruluk oranları[5] ..... 6
Çizelge 1.6 EM yöntemi ile öğrencilerin kümelenmesi [6] ..... 7
Çizelge 1.7 Yöntemlerin doğruluk değerleri [7] ..... 7
Çizelge 1.8 Karar ağacı sınıflandırma kuralları [8] ..... 8
Çizelge 1.9 Doğruluk sonuçları [8] ..... 9
Çizelge 1.10 Elde edilen kümeler [10] ..... 10
Çizelge 2.1 Veri madenciliğinin sektörler bazında kullanımı [30] ..... 21
Çizelge 4.1 Veri Seti Özellikleri ..... 46
Çizelge 4.2 Okul Türü Kodları ..... 56
Çizelge 4.3 Sınıflara göre öğrencilerin başarı yüzdeleri ..... 64
Çizelge 4.4 Farklı küme sayıları sonucunda k-ortalamalar yöntemi için doğruluk değerleri ..... 71
Çizelge 4.5 Bölgelere Göre Nüfus Dağılımları ..... 73
Çizelge 4.6 Bölgelere Göre Nüfus Katsayıları ..... 73
Çizelge 4.7 Küme-1 için Bölgelerden Gelen Öğrenci Sayıları ..... 73
Çizelge 4.8 Küme-1 için bölgelere göre ağırlık katsayıları ..... 74
Çizelge 4.9 Küme-1 için bölgelere göre yoğunluk katsayıları ..... 74
Çizelge 4.10 k-ortalamalar yöntemi sonucunda elde edilmiş küme merkezleri ..... 75
Çizelge 4.11 Farklı küme sayıları sonucunda k-medoids yöntemi için doğruluk değerleri ..... 76
Çizelge 4.12 k-medoids yöntemi sonucunda elde edilmiş küme merkezleri ..... 78
Çizelge 4.13 Bulanık c-ortalamalar yöntemi için doğruluk indeksi sonuçları ..... 79
Çizelge 4.14 Bulanık c-ortalamalar yöntemi sonucunda elde edilmiş küme merkezleri ..... 81
Çizelge 4.15 Bulanık c-ortalamalar yöntemi için üyelik matrisi ..... 82
Çizelge 4.16 Gustafson-Kessel yöntemi için doğruluk indeksi sonuçları ..... 83
Çizelge 4.17 Gustafson-Kessel yöntemi sonucunda elde edilmiş küme merkezleri ..... 84
Çizelge 4.18 Gath-Geva yöntemi için doğruluk indeksi sonuçları ..... 85
Çizelge 4.19 Gath-Geva yöntemi sonucunda elde edilmiş küme merkezleri ..... 86
Çizelge 4.20 Doğruluk indeksi yöntemlerinin kümeleme algoritmaları için başarı sonuçları ..... 87
Çizelge 4.21 Karar ağacı sınıflandırma istatistikleri ..... 89
Çizelge 4.22 Karar ağacı sınıf karışıklık matrisi ..... 89

Çizelge 4.23 Anket veri seti ..... 90
Çizelge 4.24 K-ortalamalar yöntemi ile elde edilmiş küme merkezleri ..... 97
Çizelge 4.25 k-medoids yöntemi ile elde edilmiş küme merkezleri ..... 99
Çizelge 4.26 Bulanık c-ortalamalar yöntemi ile elde edilmiş küme merkezleri ..... 101
Çizelge 4.27 Gustafson-Kessel yöntemi ile elde edilmiş küme merkezleri ..... 102
Çizelge 4.28 Gath-Geva yöntemi ile elde edilmiş küme merkezleri ..... 104

# ÖZET

## VERİ MADENCİLİĞİ İLE MÜHENDİSLİK FAKÜLTESİ ÖĞRENCİLERİNİN OKUL BAŞARILARININ ANALİZİ

Ahmet SAYGILI

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Yrd. Doç. Dr. Songül ALBAYRAK

Kümeleme analizi, veri setindeki nesneler arasındaki benzer özellikleri veya farklı özellikleri kullanarak, aynı küme içerisinde homojen, farklı kümeler arasında ise heterojen gruplar oluşturmayı amaçlamaktadır. Başka bir deyişle bir kümeyi oluşturan nesneler birbirine ne kadar benzerse ve farklı kümeler birbirinden ne kadar ayrık ise kümeleme işlemi o ölçüde başarılı olmuştur.

Bu tez çalışmasında veri madenciliği yöntemleri kullanılarak öğrenci analizi işlemleri gerçekleştirilmiştir. Bu analiz işleminde mühendislik fakültesi öğrencilerinin demografik bilgileri, ÖSYS puanları ve yerleşme sıraları ile kazandıkları bölümdeki ağırlıklı başarı not ortalamaları sistemin giriş verisi olarak kullanılmıştır. Öğrencilerin gelmiş oldukları bölgelere ve okul türlerine göre de bir kümeleme işlemi gerçekleştirilmiştir. Bunun yanında öğrencilerin belli bir bölümüne anket yapılmıştır. Bu anketler sayesinde de öğrencilerin aile bilgilerinin başarısını nasıl etkilediği incelenmiştir.

Kümeleme yöntemleri, bulanık kümeleme ve katı kümeleme olmak üzere iki ana başlık altında karşılaştırmalı olarak ele alınmıştır. Bu bağlamda bulanık kümeleme yöntemlerinden Bulanık C-Ortalamalar, Gustafson-Kessel ve Gath-Geva algoritmaları ile katı kümeleme yöntemlerinden k-ortalamalar ve k-medoids algoritmaları ayrıntılı olarak açıklanmış ve veri seti üzerinde uygulanarak başarıları üzerine bir değerlendirme yapılmıştır. Başarı durumlarının tespitinde küme geçerliliği (cluster validity) yöntemleri ve Box-Plot analizi kullanılmıştır.

**Anahtar Kelimeler:** Kümeleme Yaklaşımları, Bulanık Kümeleme, Bulanık C-Ortalamalar, Gustafson-Kessel, Gath-Geva, K-Ortalamalar, K-Medoids, Küme Geçerliliği.

# ABSTRACT

## INVESTIGATION OF STUDENT SUCCESS AT FACULTY OF ENGINEERING BY USING DATA MINING

Ahmet SAYGILI

Department of Computer Engineering

M.Sc. Thesis

Adviser: Assist. Prof. Dr. Songül ALBAYRAK

Cluster analysis, using the different characteristics or similar properties of objects in the data set, aims at creating in the same cluster homogeneous and between different clusters heterogeneous groups. In other words, the objects that make up a cluster that is so similar to each other and discrete clusters in the clustering process, and different from each other as much as it has been so successful.

In this thesis, using the methods of data mining operations carried out analysis of the student. This is the process of analyzing students' demographic data, and settlement in University Entrance Exam scores success percentages weighted grade point average information gained will be used. A clustering process was carried out according to school types and students regions. In addition, a particular section of the students were surveyed. Through these surveys examined how it affects the success of the students' family information.

Clustering methods, fuzzy clustering and hard clustering to be discussed under two main headings in comparison. In this context fuzzy clustering methods, fuzzy C-means, Gustafson-Kessel, and Gath-Geva algorithms and hard clustering methods, k-means and k-medoids algorithms are explained in detail, and an assessment of the achievements made by applying on data set. Cluster validity and Box-Plot analysis methods were used in determining the success status.

**Keywords**: Clustering Approaches, Fuzzy Clustering, Fuzzy C-Means, Gustafson-Kessel, Gath-Geva, K-Means, K-Medoids, Cluster Validation

# BÖLÜM 1

# GİRİŞ

## 1.1 Literatür Özeti

Bilgisayar teknolojilerinin gelişimi ve kayıt cihazlarının çok büyük verileri depolayabilmesinin ardından veri madenciliği alanında yapılan çalışmalarda hız kazanmıştır. Kayıt altına alınan bu verilerden birisi de yükseköğretime başlayan öğrenci verileridir. Öğrenci verilerini kullanarak gerçekleştirilmiş birçok çalışma bulunmaktadır.

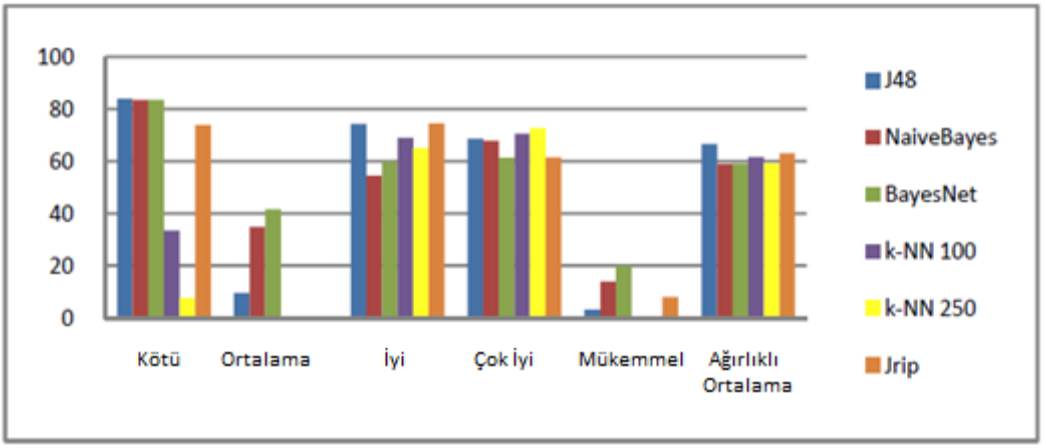
Guruler vd. [1]'de yaptıkları çalışmada veri madenciliği yöntemleri ile öğrenci başarısını etkileyen faktörleri tespit ederek öğrenci performans analizi gerçekleştirmiştir. 2010 yılında gerçekleştirilmiş bu çalışmada 1995-2005 yılları arasındaki Muğla Üniversitesi Ekonomi ve Sosyal Bilimler bölümlerinin öğrencilerinin demografik bilgileri kullanılmıştır. Çalışmada MUSKUP(Mugla University Student Knowledge Discovery Unit Program) ismini verdikleri bir uygulama geliştirmişlerdir. Bu uygulama veriden bilgi keşfi süreçleri ile veri tabanı yönetim sistemini entegre ederek oluşturulmuştur. Öğrenci karakteristiklerini tespit etmek amacıyla sınıflama tekniği olarak karar ağaçları kullanılmıştır. Oluşturulan veri setinin özelliklerinden bazıları; doğum tarihi, nüfusa kayıtlı olduğu il, lise diploma derecesi, lise tipi, yabancı dil bilgisi, üniversiteye giriş sınavı derecesi, üniversite yerleştirme bilgileri, aile yaşam koşulları ve finansal durumu, öğrencinin kaç dönemdir okula devam ettiği ve ağırlıklı not ortalamasıdır. MUSKUP uygulaması Windows Server 2003 işletim sistemi ve SQL server 2000 veri tabanı yönetim sistemi kullanılarak gerçekleştirilmiştir. Veri madenciliği modellerinin oluşturulması için SQL sunucusu içerisinde yer alan analiz servisleri kullanılmıştır. Veriye erişimi sağlamak için de MDAC 2.8 kullanılmıştır. Modellerin ve doğrulama sonuçlarının gösterilmesi içinde Angoss OLEDB (Data Mining Consumer Controls) kullanılmıştır. Program kodlama işlemi ise Visual Basic 6.0 ile ADO(ActiveX Data Objects) ve DSO(Decision Support Objects) kullanılarak

gerçekleştirilmiştir. Veri önişleme işlemlerinin ardından eldeki veriler karar ağaçları ile sınıflandırılmıştır. Bu sınıflandırma yardımıyla hangi demografik bilginin öğrencinin ağırlıklı not ortalamasını (GPA) daha çok etkilediği tespit edilmiştir. Çalışmanın kapsamı dâhilinde birinci model de GPA'sı 2.0'a(mezuniyet için gerekli minimum GPA) eşit veya daha büyük olanlar ile ikinci model de not ortalaması 3.0 veya üstü (onur derecesi) olan öğrencilerin profillerinin belirlenmesi hedeflenmiştir. Ve böylece iki sınıflandırma modeli elde edilmiştir. Birinci modelde üniversiteye kayıt tipi(1. Yerleştirme, yatay geçiş vb.), ikinci model de ise ailenin aylık geliri GPA'yı en fazla etkileyen veriler olmuştur.

Kabakchieva'nın [2]'de 2011 yılında Sofia Üniversitesinde gerçekleştirdiği çalışmada 10330 öğrencinin 20 farklı özelliği kullanılmıştır. Bu özelliklerden bazıları; cinsiyet, doğum yılı ve yeri, yaşadığı yer ve ülke, üniversiteden önce yer aldığı okulun tipi, profili, yeri ve okuldan mezun olma derecesi, üniversiteye kabul edildiği yıl, üniversite sınav notları ve genel not ortalamasıdır. Çalışmanın amacı veri madenciliği yöntemleri ile öğrencinin kişisel bilgileri ve üniversite öncesindeki özellikleri ile üniversite performansının tahmin edilmesidir. Hedef değişkenin alabileceği değerler öğrencilerin genel not ortalaması kullanılarak şu şekilde gruplanmıştır; muhteşem (5.50-6.00), çok iyi (4.50-5.49), iyi (3.50-4.49), ortalama (3.00-3.49), kötü (3.00'ün altı). Veri setinde yer alan 10330 öğrenciden, 539 öğrenci muhteşem, 4336 öğrenci çok iyi, 4543 öğrenci iyi, 347 öğrenci ortalama, 564 öğrenci de kötü kategorisine girmiştir. Çalışma da veri madenciliği yöntemlerini uygulamak için WEKA aracı kullanılmıştır. Birden fazla sınıflayıcı yöntem kullanılmıştır. Ve bu sınıflayıcı yöntemler iki farklı test seçeneği ile gerçeklenmiştir. Bunlardan birisi 10'a katlamalı çapraz geçerlilik testi (10 fold-cross validation) ve bir diğeri de 2/3'ü eğitim ve 1/3'ü test için kullanılan yüzde parçalama (percentage split) yöntemidir. Yapılan sınıflamalar sonucunda elde edilen doğru-pozitif oranı(true positive rate) ve hassasiyet(precision) değerleri Çizelge 1.1'de gösterilmektedir. Şekil 1.1'de ise kullanılan yöntemlerin başarı oranları yer almaktadır.

Çizelge 1.1 Sınıflayıcıların doğruluk değerleri sonuçları [2]

| | **J48** | **Naive Bayes** | **BayesNet** | **kNN, k=100** | **kNN, k=250** | **OneR** | **JRip** |
|---|---|---|---|---|---|---|---|
| **Doğru Pozitif Oranı** | 0.663 | 0.586 | 0.591 | 0.613 | 0.593 | 0.546 | 0.632 |
| **Hassasiyet** | 0.640 | 0.594 | 0.597 | 0.574 | 0.563 | 0.480 | 0.611 |



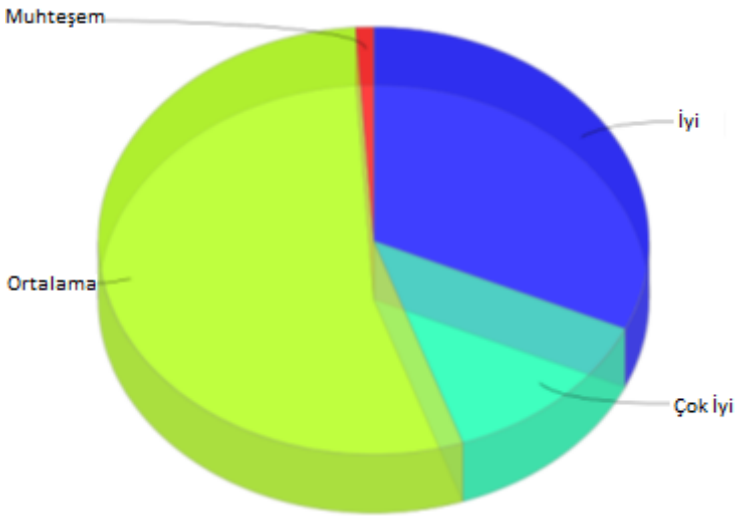

Şekil 1.1 Kullanılan yöntemlerin başarı oranları [2]

Çizelge 1.1'de de görüldüğü gibi çalışma sonucunda en başarılı olan sınıflayıcı bir karar ağacı çeşidi olan J48 yöntemidir. Bayes sınıflayıcılar en düşük başarı oranını vermişlerdir. Yapılan çalışmada genel başarı oranı %70'lerin altında kalmıştır. Bu da verideki hata oranının yüksek olduğunu ve elde edilen sonuçların çokta gerçekçi olmadığını göstermektedir.

Varghese vd. [3]'de performans örüntülerini karakterize etmek için öğrenci verileri üzerinde kümeleme işlemi gerçekleştirmiştir. Çalışmada kümeleme yöntemlerinden k-ortalamalar yöntemi ve bulanık c-ortalamalar yöntemi kullanılmıştır. Veri seti için ise öğrencilerin devam durumu, iç not değerlendirmeleri, seminer değerlendirmeleri, sınıf içi ödev değerlendirmeleri ve üniversite ağırlıklı not ortalamaları kullanılmıştır. Veri seti toplam 8000 kayıt içermektedir. Yapılan çalışma sonucunda dönemler ilerledikçe iç not değerlendirmelerinin düştüğü gözlemlenmiştir. Ve ayrıca öğrenci performansı ile devam etme durumu arasında da pozitif bir ilişki olduğu tespit edilmiştir. Devam azaldığında performans da düşmekte, devam arttığında performans da artmaktadır.

Tair ve Halees [4]'te, 2012 yılında Filistin'de Gazze İslam Üniversitesi'nin Bilgi Teknolojileri fakültesinde öğrenim gören öğrencilerin verileri üzerinde öğrencilerin performanslarını arttırmaya yönelik veri madenciliği yöntemleri uygulamıştır. Veriler 1993-2007 yılları arasındaki 15 yıllık periyodu kapsamaktadır. Çalışma da birliktelik kuralları (association rules), sınıflama (classification), kümeleme (clustering) ve aykırı değer tespiti (outlier detection) teknikleri kullanılmıştır. Veriler Khanyounis Bilim ve Teknoloji Fakültesi öğrencilerinin verilerinden oluşmaktadır. Veri setinde 18 özellik yer almaktadır. Bunlar öğrencinin numarası, adı, cinsiyeti, doğum yeri, adresi, telefon numarası, genel not ortalaması, mezun olduğu lise türü vb. özelliklerdir. Toplam da 3360 kayıt yer almakta iken bu kayıtlardan 51' i eksik veri içerdiğinden dolayı ön

işleme aşamasında çıkartılmıştır. Veri seti özelliklerinden üniversiteye giriş sınavı notu (Matriculation GPA) çok geniş bir yelpaze de olduğu için Şekil 1.2'de görüldüğü gibi önişleme ile 5 farklı gruba kategorize edilmiştir. Böylece öğrencilerin notları Muhteşem, Çok İyi, İyi, Ortalama ve Zayıf'a dönüştürülmüştür.




Şekil 1.2 Kategorize edilmiş üniversiteye giriş notlarının öğrenciler üzerindeki dağılım grafiği [4]

Şekil 1.2'ye bakıldığında 1993-2007 yılları arasındaki öğrencilerin %54'ünün ortalama olarak gruplandığı görülmektedir. Çalışma da kullanılan bir diğer veri madenciliği yöntemi sınıflandırmadır. Sınıflandırma da Naive Bayesian ve Kural Çıkarma metotları kullanılmıştır. Kural çıkarma (Rule induction) yöntemi sonucunda %71.25 doğruluk oranı ile sınıflandırma gerçekleştirilmiştir. Naive bayesian yönteminde de %67.5 doğruluk oranı ile sınıflandırma gerçekleştirilmiştir. Bu çalışmada kullanılan bir diğer veri madenciliği yöntemi de kümelemedir. Çalışmada kümeleme yöntemlerinden k-ortalamalar kullanılmıştır. Çizelge 1.2'de veri setine k-ortalamalar yöntemi uygulanmasının ardından elde edilen kümeler görülmektedir.

Çizelge 1.2 K-ortalamalar yöntemi sonuçları [4]

| **Özellikler** | **Küme-1** | **Küme-2** | **Küme-3** | **Küme-4** |
|---|---|---|---|---|
| Not Ortalaması | 0.414 | 0.862 | 0.500 | 0.308 |
| Cinsiyet | 0.372 | 0.445 | 0.621 | 0.517 |
| Uzmanlık Alanı | 5.108 | 1.316 | 9.900 | 15.335 |
| Şehir | 2.191 | 2.176 | 2.374 | 2.498 |
| Lise Tipi | 0.769 | 0.097 | 0.748 | 0.796 |
| Sınav Notu | 1.895 | 1.511 | 1.886 | 1.929 |
| **Toplam Kayıt Sayısı** | 769 | 1032 | 1033 | 480 |
| **Toplam Yüzde** | 23% | 31% | 31% | 15% |

Elde edilen bu 4 kümenin çeşitli özellikleri Çizelge 1.3'de görülmektedir. Bu çizelge vasıtasıyla hangi kümede hangi özelliklerin var olduğunu gözlemlemek mümkün olmaktadır.

Çizelge 1.3 Özelliklere göre kümelerin karakteristikleri [4]

| **Özellikler / Kümeler** | **Not Ortalaması** | | **Cinsiyet** | | **Lise Tipi** | | **Sınav Notu** | |
|---|---|---|---|---|---|---|---|---|
| **Küme-1** | Zayıf | 73% | Erkek | 63% | İlim | 23% | İyi | 31% |
| | Ortalama | 15% | Kadın | 37% | Adap | 77% | Çok İyi | 8.60% |
| | İyi | 9% | | | | | Muhteşem | 0.40% |
| | Çok İyi | 3% | | | | | Ortalama | 60% |
| **Küme-2** | Zayıf | 48% | Erkek | 56% | İlim | 90% | İyi | 37% |
| | Ortalama | 24.4% | Kadın | 44% | Adap | 10% | Çok İyi | 19% |
| | İyi | 22% | | | | | Muhteşem | 2% |
| | Çok İyi | 5% | | | | | Ortalama | 42% |
| | Muhteşem | 0.60% | | | | | | |
| **Küme-3** | Zayıf | 67% | Erkek | 38% | İlim | 25% | İyi | 29% |
| | Ortalama | 20% | Kadın | 62% | Adap | 75% | Çok İyi | 12.50% |
| | İyi | 9.60% | | | | | Muhteşem | 0.50% |
| | Çok İyi | 3% | | | | | Ortalama | 58% |
| | Muhteşem | 0.40% | | | | | | |
| **Küme-4** | Zayıf | 78% | Erkek | 48% | İlim | 20% | İyi | 30% |
| | Ortalama | 15% | Kadın | 52% | Adap | 80% | Çok İyi | 8.20% |
| | İyi | 5% | | | | | Muhteşem | 0.60% |
| | Çok İyi | 1.50% | | | | | Ortalama | 61% |

Ayrıca çalışmada aykırı değer tespiti yapılarak aykırı değerler değerlendirilmeden çıkartılmıştır. Mesafeye dayalı (Distance based) yaklaşımı ve yoğunluğa dayalı (density based) yaklaşımı çalışmada kullanılan aykırı değer tespiti yöntemleridir.

Yadav ve Pal [5]' te öğrencilerin performanslarını arttırmaya yönelik araştırmalar yapmıştır. Çalışmada C4.5, ID3 ve CART karar ağacı algoritmaları kullanılarak Hindistan Üniversitesi mühendislik fakültesi öğrencilerinin final sınavlarındaki performanslarının tahmin edilmesi amaçlanmıştır. Veri seti 2010 yılında Purvanchal Üniversitesinin mühendislik fakültesinde okumakta olan 90 öğrenciden oluşmaktadır. Yöntemler için WEKA açık kaynak kodlu yazılımı kullanılmıştır. Test işlemi için ise 10'a katlamalı çapraz geçerlilik (10-fold cross validation) testi kullanılmıştır.

Yöntemlerin uygulanmasının ardından sınıflayıcıların doğruluk oranı Çizelge 1.4'de gösterilmektedir.

Çizelge 1.4 Yöntemlerin doğruluk oranları[5]

| Algoritma | Doğru Sınıflandırılmış Örnekler | Yanlış Sınıflandırılmış Örnekler |
|---|---|---|
| **ID3** | 62.2% | 26.6% |
| **C4.5** | 67.7% | 32.2% |
| **CART** | 62.2% | 37.7% |

Çizelge 1.5'deki doğruluk oranlarına bakıldığında doğru-pozitif oranlarının oldukça yüksek olduğu görülmektedir. Bu da yöntemlerin başarısını göstermektedir.

Çizelge 1.5 Yöntemlerin doğruluk oranları[5]

| Algoritma | Sınıf | Doğru-Pozitif | Yanlış-Pozitif |
|---|---|---|---|
| **ID3** | Geçen | 0.714 | 0.184 |
| | Terfi Eden | 0.625 | 0.232 |
| | Kalan | 0.786 | 0.061 |
| **C4.5** | Geçen | 0.745 | 0.209 |
| | Terfi Eden | 0.517 | 0.213 |
| | Kalan | 0.786 | 0.092 |
| **CART** | Geçen | 0.803 | 0 349 |
| | Terfi Eden | 0.31 | 0.18 |
| | Kalan | 0.643 | 0.105 |

Halees [6]'da yaptığı çalışmada 2007-2008 yılının ilk sömestrin de veri tabanı dersini alan üniversite öğrencilerinin verilerini kullanmıştır. Veri setinde toplam da 151 öğrencinin verisi kullanılmıştır. Veriler öğrencilerin kişisel, akademik ve ders bilgilerinden oluşmaktadır. Ayrıca bir açık kaynak sistemi olan Moodle sisteminden de öğrencilerle alakalı çeşitli bilgiler alınmaktadır. Çalışmada veriler 5 farklı veri madenciliği aşamasına tabi tutulmuştur. Bu aşamalar, birliktelik(Assocation), kümeleme (clustering), sınıflama (classification) ve aykırı değer tespitinden (outlier detection) ibarettir. Veri önişleme aşamasında öğrencilerin not bilgileri muhteşem, çok iyi, iyi, zayıf ve başarısız olmak üzere beş farklı gruba ayrılmıştır. Çalışma da sınıflandırma aşamasında bir tür karar ağacı algoritması olan J48 kullanılmıştır.

Kümeleme işlemin de ise beklenti maksimizasyonu (Expectation-Maximization) algoritması kullanılmıştır. Yapılan kümelemenin ardından öğrenciler 5 farklı kümeye ayrılmıştır. Aşağıdaki çizelgede bu kümeler görülmektedir.

Çizelge 1.6 EM yöntemi ile öğrencilerin kümelenmesi [6]

| Özellikler | Küme-1 | Küme-2 | Küme-3 | Küme-4 | Küme-5 |
|---|---|---|---|---|---|
| **Devam** | 87.05 | 65.08 | 39.79 | 15.12 | 67.96 |
| **GPA** | 77.68 | 70.31 | 65.37 | 67.84 | 71.52 |
| **Saatler** | 88.08 | 95.62 | 64.37 | 57.12 | 90.87 |
| **e-kaynaklar** | 0.90 | 0.27 | 1.00 | 1.23 | 3.17 |
| **e-çalışmalar** | 7.81 | 2.17 | 3.37 | 2.76 | 10.70 |
| **e-ödevler** | 9.50 | 6.67 | 3.25 | 1.61 | 6.18 |
| **Ara sınav** | 14.71 | 12.52 | 5.00 | 1.69 | 13.34 |
| **Lab. notu** | 16.35 | 15.33 | 5.00 | 4.76 | 12.96 |
| **Final** | 37.10 | 30.74 | 14.87 | 29.72 | 30.29 |
| **Notlar** | 77.66 | 65.93 | 26.87 | 68.37 | 65.28 |

Son olarak veri setine, başarımı artırmak adına aykırı değer tespiti yapılmış ve toplam 37 adet aykırı veri çıkartılmıştır.

Radaideh vd. [7]'de yaptıkları çalışmayı Ürdün Yarmouk Üniversitesinde gerçekleştirmiştir. Çalışmada Yarmouk Üniversitesinde 2005 yılında C++ dersini alan öğrencilerin verileri kullanılmıştır. Veri setinin oluşturulması için bilgisayar bilimleri fakültesinde öğrenim gören öğrencilere anket yapılmıştır. Bu anket sonucunda öğrencilerle ilişkili toplam 20 özellik elde edilmiştir. Bu özelliklerin bazıları el ile elenmiş ve veri seti 13 özelliğe indirgenmiştir. Bu 13 özellikten bir tanesi hedef sınıfı işaret eden ve öğrencilerin C++ dersinden aldıkları notları gösteren özelliktir. Bir sonraki adım da ise karar ağaçları kullanılarak verileri sınıflandırma işlemi gerçekleştirilmiştir. Karar ağaçlarında bilgi kazancı (Information Gain) metriğine göre sınıflandırma yapılmış ve en yüksek kazanca sahip olan öğrencinin lise not ortalaması kök düğüm olma özelliği kazanmıştır. Diğer kısımlara da bu yöntemler uygulanarak karar ağacı oluşturulmuş ve bu karar ağacına binaen toplamda 41 sınıflandırma kuralı elde edilmiştir. Bu işlemlerin yanı sıra WEKA aracı kullanılarak üç farklı sınıflandırma metodu da veri seti üzerinde uygulanmıştır. Bu metotlar ID3, C4.5 ve Naive Bayes'den oluşmaktadır. Uygulanan bu metotların doğruluk değerleri aşağıdaki çizelgede gösterilmektedir.

Çizelge 1.7 Yöntemlerin doğruluk değerleri [7]

| Algoritma | Hold out yöntemi | 10'lu çapraz sağlama |
|---|---|---|
| **ID3** | 38.46% | 28.31% |
| **C4.5** | 35.89% | 38.05% |
| **Naive Bayes** | 33.33% | 38.05% |

Sembiring vd. [8]'de yaptıkları çalışmanın amacını bir veri madenciliği yöntemi olan kernel yöntemini kullanarak öğrencilerin davranışları arasındaki ilişkilerin ve öğrencilerin başarılarının analiz edilmesi olarak tanımlamaktadırlar. Bunun için destek vektör makinaları sınıflandırma yöntemi ve kernel k-ortalamalar kümeleme yöntemi kullanılmıştır. Çalışmaya konu olan veri seti Malaysia Pahang Üniversitesinde 2007-2008 yıllarının üçüncü sömestrinde veri tabanı yönetim sistemleri dersini alan öğrencilere yapılmış olan anketler vasıtasıyla oluşturulmuştur. Ankette kullanılan değişkenler ilgisi, çalışması, meşguliyeti, inanması ve aile desteği'nden oluşmaktadır. Bu anket bilgisayar sistemleri ve yazılım mühendisliğinde okumakta olan 1000 öğrenciye yapılmıştır. Toplanan veriler öğrencilerin kişisel kayıtları, akademik kayıtları ve kurs kayıtlarından oluşmaktadır. Elde edilen verilere çeşitli önişleme aşamaları uygulanmıştır. Örneğin kategorize edilebilecek durumda olan not verisi için muhteşem, çok iyi, iyi, orta ve zayıf olmak üzere 5 farklı grup tanımlanmıştır. Veriler normal dağılım gösteren özellikler için yüksek, orta ve düşük değerleri verilerek kategorize edilmiştir. Verilerin önişlemden geçirilmesinin ardından kernel k-ortalamalar kümeleme yöntemi uygulanmıştır. Ve beş farklı küme elde edilmiştir. Kümelemenin ardından verilere J48 karar ağacı sınıflandırma yöntemi uygulanmıştır. Oluşturulan karar ağacından elde edilmiş olan sınıflandırma kurallarının küçük bir kısmı Çizelge 1.8'de görülebilmektedir.

Çizelge 1.8 Karar ağacı sınıflandırma kuralları [8]

| | **İlgisi** | **Çalışması** | **Meşguliyeti** | **İnanması** | **Aile Desteği** | **Performans Tahmini** |
|---|---|---|---|---|---|---|
| **Eğer** | Y | Y | Y | Y | Y | Muhteşem |
| | Y | O | O | Y | Y | Çok İyi |
| | O | O | O | O | O | İyi |
| | D | O | O | D | O | Orta |
| | D | D | D | D | D | Zayıf |
| **Y=Yüksek** | | | **O=Orta** | | | **D=Düşük** |

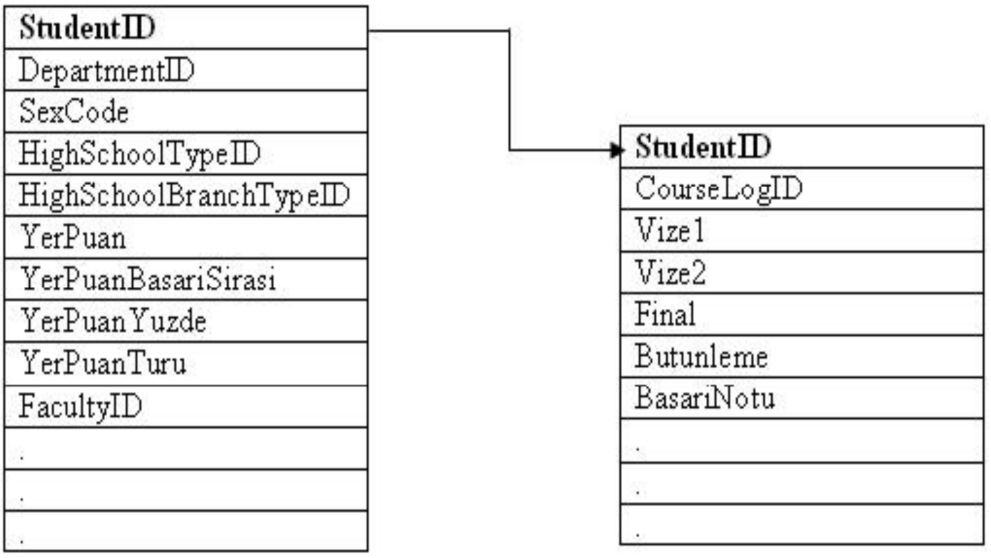
Yukardaki kuralların oluşturulduğu veri setinde toplamda 300 öğrenci yer almaktadır. Her öğrencinin de 10 karakteristik özelliği bulunmaktadır. Bu 10 özelliğin 5 tanesi performans tahminleyicisinden ve diğer 5 tanesi de öğrencilerin karakteristik demografik bilgilerinden oluşmaktadır. Veri setinin %90'ı eğitim %10'u ise test için kullanılmıştır. Daha sonra bu eğitim ve test verilerine 10'a katlamalı çapraz geçerlilik (10 fold cross validation) yöntemi uygulanmıştır. Sonuç olarak elde edilen doğruluk oranları Çizelge 1.9'da yer almaktadır.

Çizelge 1.9 Doğruluk sonuçları [8]

| Performans Tahmini | Eğitim | | Test | |
|---|---|---|---|---|
| | En iyi doğruluk (%) | Ortalama Doğruluk (%) | En iyi doğruluk (%) | Ortalama Doğruluk (%) |
| **Muhteşem** | 100.00 | 99.67 | 100.00 | 92.00 |
| **Çok İyi** | 100.00 | 100.00 | 93.33 | 75.67 |
| **İyi** | 100.00 | 100.00 | 73.33 | 61.00 |
| **Orta** | 100.00 | 99.70 | 80.00 | 69.33 |
| **Zayıf** | 100.00 | 99.70 | 96.67 | 93.67 |

Shannaq vd. [9]'da 2010 yılında yaptıkları çalışmada yükseköğrenim gören öğrencilerin başarısını etkileyen faktörleri tespit etmeyi amaçlamışlardır. Gerçekleştirilen uygulama VB.NET ortamında ve karar ağaçları ile oluşturulmuştur. Kullanılan veri tabanında toplam 19012 öğrenci kaydı bulunmaktadır. Çok fazla eksik veri içeren bu kayıtların sayısı veri önişleme aşaması ile 2106'ya düşmüştür. Çalışmada verilerin hangi üniversiteye ait olduğu belirtilmemiş fakat Arap coğrafyasında yer aldığı belirtilmiştir. Kullanılan veriler 2003-2007 yılları arasını kapsamaktadır. Verilerin %20'si doğrulama için kullanılmıştır. Yapılan bu çalışma ile öğrencilerin üniversiteye kaydolma ya da üniversiteyi seçme durumlarının tahmin edilmesi işlemi gerçekleştirilmiştir. Öğrencilerin profiline hâkim olan bir üniversite, kendini yeni kaydolacak öğrencilere hazırlayabilecek ve böylece başarı oranı artacaktır.

Erdoğan ve Timor [10]'da öğrencilerin üniversiteye giriş sınavı bilgileri ile üniversitedeki dersleri ile ilgili bilgileri kullanarak kümeleme yapmışlardır. Çalışma, veri madenciliği yöntemlerinden k-ortalamalar kümeleme algoritması kullanılarak gerçekleştirilmiştir. Öğrenci verileri ise Microsoft SQL Server 2000 ortamından elde edilmiştir. Veri tabanındaki farklı tablolardan gereken veriler alınarak çalışma için Şekil 1.3'de gösterilen tablolar ve özellikler oluşturulmuştur.




Şekil 1.3 Öğrenci ve öğrenci notları [10]

Yukarda gösterilen veriler üzerinde farklı küme sayıları denenmiş ve en uygun küme sayısı 5 olarak belirlenmiştir. Elde edilen kümeler aşağıdaki çizelgede gösterilmiştir.

Çizelge 1.10 Elde edilen kümeler [10]

| Küme | Alanı | Başarı Notu | Cinsiyet | Lise Tipi | Fakülte No |
|---|---|---|---|---|---|
| **1** | 12.4774 | 89.5350 | 8.1070 | 6.4650 | 2.5844 |
| **2** | 16.3113 | 59.0472 | 6.9811 | 5.1981 | 3.0189 |
| **3** | 46.7851 | 77.1240 | 6.9008 | 5.7190 | 3.2149 |
| **4** | 80.1095 | 78.0146 | 6.6788 | 3.2774 | 3.8321 |
| **5** | 79.3565 | 44.8870 | 6.3043 | 3.1391 | 4.1304 |

## 1.2 Tezin Amacı

Bu çalışmanın amacı, öğrencilerin çeşitli özellikleri (ÖSYS puanı ve sırası, demografik bilgileri, üniversite başarı notu vb.) göz önüne alınarak başarılı olma durumlarının tespit edilmesidir. Veri madenciliğinde kullanılan kümeleme yöntemleri ile başarılarını etkileyen faktörlerin neler olduğunun tespit edilmesi de tezin kapsamı dâhilindedir. Yapılan bu çalışmanın ardından mevcut veriler dahilinde N.K.Ü. Çorlu Mühendislik Fakültesi öğrencilerinin başarılarında rol oynayan özellikleri tespit edilmiş olacaktır. Aynı zamanda bu tez konuyla alakalı sonraki yapılacak çalışmalara da ışık tutacaktır.

## 1.3 Hipotez

Veri madenciliği veri yığınları arasından faydalı ve gözle görülemeyen bilgileri ortaya çıkarma sürecidir. Veri madenciliği yöntemlerinden birisi olan kümeleme ise birbirine

benzerlik gösteren nesneleri bir araya getirerek onları daha iyi kavramamızı ve benzer nesnelerin ortak karakteristik özellikleri hakkında yorum yapabilmemizi sağlayan veri analizi tekniğidir.

Günlük hayatta insanlar birbirine benzer özellikler taşıyan nesneleri kümelere ayırarak düşünür ve algılamalarını kolaylaştırırlar. Örneğin canlıları, bitkiler, hayvanlar ve insanlar şeklinde üç kümeye ayırarak benzer karakteristiklerine göre gruplara bölmüşüzdür. Doğamızdan gelen, bu kümeleme ihtiyacımızdan dolayı ve son yıllarda veri tabanlarında tutulan verilerdeki çok büyük artışlar nedeniyle, çok değişkenli bir teknik olan küme analizi teknikleri ortaya çıkmıştır. Kümeleme teknikleri sayesinde elde ki mevcut veriler içerisinden kullanıcılar için anlamlı ve faydalı sonuçlar ortaya çıkmaktadır.

Kümeleme analizi teknikleri bu çalışmada bulanık kümeleme ve katı kümeleme olmak üzere iki ana başlık altında incelenmiştir. Bulanık kümeleme yöntemlerinden BCO (Bulanık c-ortalamalar), GK (Gustafson-Kessel) ve GG (Gath-Geva) yöntemleri, katı kümeleme yöntemlerinden ise k-ortalamalar ve k-medoids yöntemleri teorik olarak incelenmiştir. Ayrıca bahsi geçen yöntemler uygulama kısmında N.K.Ü. Çorlu Mühendislik Fakültesi öğrenci verilerinin analiz edilmesi ve kümelenmesinde de kullanılmıştır. Literatürde sıkça kullanılan küme geçerlilik indekslerinden, Bölünme Katsayısı, Sınıflandırma Entropisi, Bölümlendirme İndeksi, Ayırma İndeksi anlatılmış ve uygulamada küme sayılarının tespit edilmesinde ve kullanılan yöntemlerin başarılarının analiz edilmesinde kullanılmıştır. Ayrıca yöntemlerin başarılarının analizinde Box-Plot aykırı değer analizi de kullanılmıştır. Kümeleme analizi tekniklerinin yanı sıra sınıflama yöntemlerinden karar ağaçları da verileri sınıflandırmak için kullanılmıştır.

Tez çalışmasının birinci bölümünde tez konusu, yapılacak olan çalışma, çalışmanın amaç ve hedefleri ve bu alanda daha önce yapılmış olan çalışmalardan bahsedilmektedir. İkinci bölümde, veri madenciliğinin ne olduğu, nerelerde kullanıldığı ve veri madenciliği sürecinin adımları detaylı bir şekilde ele alınmıştır. Üçüncü bölümde, tez kapsamında yapılan uygulamada kullanılan kümeleme ve sınıflandırma yöntemleri yapılan uygulamanın anlaşılabilirliğini artırmak adına detaylı bir şekilde ele alınmıştır. Ayrıca yine üçüncü bölümde literatürde sıkça kullanılan küme geçerlilik yöntemlerinden bahsedilmiştir. Dördüncü bölümde ise üçüncü bölümde incelenen yöntemlerin çalışmaya konu olan veri setine uygulanması sonucunda elde edilen çıktılar yer almaktadır. Uygulama bölümü olarak adlandırdığımız bu bölümde, bir önceki

bölümde bahsedilen yöntemler Namık Kemal Üniversitesi Çorlu Mühendislik Fakültesinde 2011 sonu itibariyle kayıtlı olan lisans öğrenci verilerine uygulanmıştır. Öğrencilerin demografik verileri, ÖSYS puanları ve sıraları, üniversite genel başarı notları gibi bilgiler kullanılarak verilerden anlamlı sonuçlar elde edilmeye çalışılmıştır. Yine uygulama bölümünde Çorlu Mühendislik Fakültesi bilgisayar Mühendisliği bölümünde okuyan öğrencilere, mevcut veri setine katkı sağlaması açısından gerçekleştirilmiş olan anket sonuçları da yer almaktadır. Son olarak beşinci bölümde, yapılan uygulamanın sonuçlarına ve bundan sonra yapılabilecek konulara değinilmiştir.

# BÖLÜM 2

# VERİ MADENCİLİĞİ

Günümüzde bilgisayar teknolojilerindeki gelişmelerin ne kadar hızlı gerçekleştiğini gözlemleyebiliyoruz. Bilgisayar teknolojilerindeki bu gelişmeler kayıt altına alınan verilerde büyük bir artış sağlamaktadır. Marketlerde yaptığımız alışveriş işlemlerinde, bankalarda, belediyelerde, hastanelerde ve daha birçok kurumda yaptığımız bütün işlemler bir veri tabanında kayıt altına alınmaktadır. Bu olağanüstü gelişmeyi güçlü veri analizi araçları olmadan anlayabilmek insanoğlunun yeteneklerini aşan bir durum halini almış ve bu durum karar vericiyi *veri zengini fakat bilgi fakiri* konumuna sokmuştur [11]. Tam da bu nokta da veri madenciliği devreye girmektedir. Veri madenciliği için "Kayıtlı veriler içerisinden, ilk bakışta gözle görülemeyecek, geleceğe dönük tahmin yapılmasını sağlayan, anlamlı ve faydalı bilgilerin elde edilmesi işlemidir." tanımını yapmak doğru olacaktır. Bunun yanında veri madenciliği için yapılmış çeşitli tanımlamalar aşağıda yer almaktadır:

- Veri Madenciliği, büyük miktardaki veri yığını içerisinden gelecekle ilgili tahmin yapmamızı sağlayacak, bağıntı ve kuralların bilgisayar programları kullanılarak aranması olarak ifade edilebilir [12].
- Yakın geleceğin geçmişten çok fazla farklı olmayacağı varsayılırsa, geçmiş veriden çıkarılmış olan kurallar gelecekte de geçerli olacak ve ilerisi için doğru tahmin yapılmasını sağlayacaktır [13].
- Büyük veri tabanlarında gizli kalmış örüntüleri çıkarma sürecine veri madenciliği adı verilmektedir. Geleneksel yöntemler kullanılarak çözülmesi çok zaman alan problemlere veri madenciliği süreci kullanılarak daha hızlı bir şekilde çözüm bulunabilir [14].

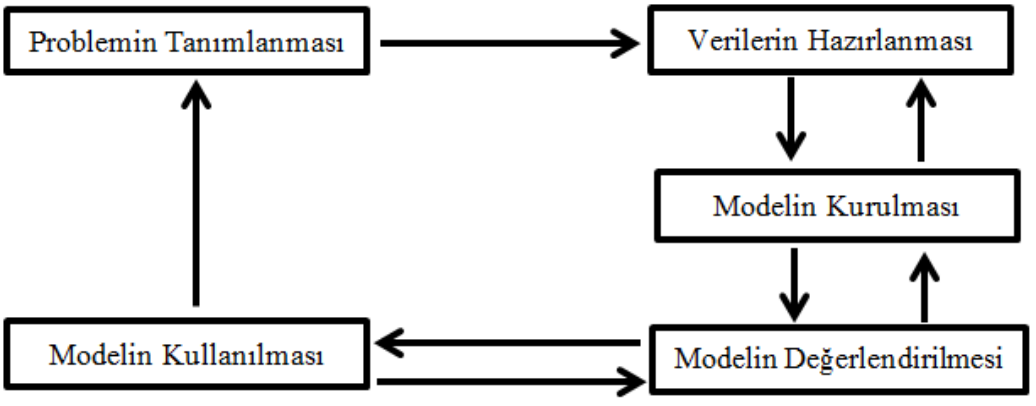
- Veri madenciliği, ham verinin tek başına sunamadığı bilgiyi çıkaran veri analizi sürecidir [15].
- Veri madenciliği, veri tabanlarından, geçerli, önceden bilinmeyen, güvenilir, potansiyel olarak kullanışlı ve nihayette anlaşılabilir örüntülerin çıkarılması işlemidir [16].

Yapılan bu tanımlamaları temel alarak veri madenciliğinin faydalarını şu şekilde sıralayabiliriz:

- Veri tabanlarında depolanmış olan verilerden yeni ve gerekli olan anlamlı bilgileri üretme,
- Verinin özelliklerinden ve elde edilen bilgilerden faydalanarak eğilimlerini anlama,
- Geleceğe yönelik tahminlerde bulunarak bilgiyi gelecekteki çalışmalarda kullanmak amacıyla değerlendirme.

## 2.1 Veri Madenciliği Süreci

Veri Madenciliği büyük ölçekli verilerden anlamlı bilginin elde edilebilmesinde gerekli olan işlemlerin modellenmesi için yöntemler ve algoritmalar sunar. Veri madenciliği pek çok alanda kullanılabilen bir yöntemdir. Ancak bu yöntem bazı aşamaların bir araya gelmesiyle oluşmuştur. Kullanılan veri madenciliği yönteminin doğru sonuçlar üretebilmesi için alt işlemler doğru olarak yerine getirilmelidir. Alt işlemlerden oluşan bu süreç 1996 yılı sonlarında "The Cross- Industry Standart Process for Data Mining (CRISP – DM)" konsorsiyumu tarafından oluşturulmuştur [17]. Bu süreç çalışmaların her adımını detaylı bir şekilde ele almaktadır. Söz konusu süreç şekildeki gibi özetlenebilir:




Şekil 2.1 Veri Madenciliği Süreci [17]

### 2.1.1 Problemin Tanımlanması (Problem Definition)

Problemin tanımlanması aşaması veri madenciliği sürecinin en önemli aşamasıdır. Bu aşama yapılan veri madenciliği çalışmasının amacını, durum değerlendirmesi yapılmasını ve planlama sürecinin belirlenmesini kapsamaktadır.

“Bu adımda ihtiyaç duyulan şeyin tanımlanması için cevaplanması gereken sorular, neyin otomatize edilmeye değer olduğu ve neyin insan içeren süreçlere bırakılması gerektiği, amacın ne olduğu ve hangi performans kriterlerinin daha önemli olduğu, sürecin sonucunda elde edilecek çıktının keşif, sınıflandırma, özetleme gibi şeyler için kullanılıp kullanılmayacağı olabilir.” [18].

Problemin tanımlanması, veri madenciliği süreci sonunda tam olarak ne elde edilmek istendiğinin belirlenmesi anlamına gelmektedir. İyi tanımlanmış bir problem istenilen sorulara ve beklentilere doğru cevaplar verilmesini sağlar.

### 2.1.2 Verinin Hazırlanması (Data Preparation)

Verinin hazırlanması aşaması, modelleme için kullanılacak olan veri kümesinin oluşturulmasına ilişkin tüm süreçleri içermektedir. Veri temizleme, veri indirgeme, veri dönüştürme, veri bütünleştirme, kullanılacak yöntemin belirlenmesi ve değerlendirilmesi faaliyetleri bu aşamada gerçekleştirilmektedir. “Veri ön işleme” olarak da tabir edilebilecek bu süreç, veri üzerindeki problemlerin çözümü ve anlamlı veri analizlerinin yapılabilmesi amacını taşımaktadır ve veri madenciliğinde en fazla zaman ile kaynağın ayrıldığı aşamadır [19].

Veri ön işleme süreci temelde iki farklı türde işlem gerektirir. Bunlardan birincisi veri kümesinin belirlenmesi ve birleştirilmesi, ikincisi ise verilerin daha yararlı bir hale getirilmesi amacıyla işlenmesidir [20].

#### 2.1.2.1 Veri Temizleme (Data Cleaning)

Hiçbir seri seti tamamıyla kusursuz değildir. Bu sebepten ötürü hatalı olan bu verilerin başarıyı olumsuz etkilememesi için veri temizleme işlemleri gerçekleştirilir. Veri temizleme, diğer aşamalarda kullanılacak veri madenciliği modelinde veri kalitesini artırmak üzere yapılır [21].

Tutarsız ve hatalı veriler, veriler üzerinde yapılacak analizlerde doğru sonuç almamızı etkileyeceği için bu “gürültü” olarak tabir edilen verilerden veri setinin temizlenmesi

gerekir. Verilerin gürültülerden temizlenmesi için birçok yöntem kullanılmaktadır. Örneğin, gürültülü verinin sisteme etkisini azaltmak için bu tip veriler analiz kümesinden çıkarılabilir veya bu tip verilerin yerine sabit bir değer koyulabilir ya da hatalı değerler yerine, doğru verilerin ortalaması yazılabilir. Bunun yanı sıra tüm değerlerin ortalaması yerine hatalı kaydın yapısına benzer örneklerin ortalaması alınarak işlem yapılması başarıyı daha da arttıracaktır.

#### 2.1.2.2 Veri Bütünleştirme (Data Integration)

Veri madenciliğinde kullanılan veri setlerinin birçoğu tek bir kaynaktan değil de birçok kaynaktan gelen verilerin bir araya gelmesi ile oluşmuştur. Bu yüzden farklı kaynaklardan gelen verilerin tutarlılık sağlaması gerekir. Örneğin, veri setinde "medeni hal" diye bir özellik olduğunu varsayalım. Medeni hal özelliği bazı kaynaklarda Evli-Bekâr olarak geçebilir. Bunun yanında başka kaynaklarda ise 1 veya 0 değerleri kullanılmış olabilir. Ya da evli için E harfi bekâr için de B harfi kullanılmış olabilir. İşte böyle bir durum da farklı kaynaklardan gelen ancak aynı anlamı ifade eden özelliklerin tek bir çatı altında toplanması gerekir. Bu tip farklı bakış açıları veriler üzerinde çalışmayı imkânsız hale getirir. Bu nedenle bu tip verilerin analiz aşamasından önce ortak bir türe dönüştürülmesi yani veri bütünleştirmesi yapılması gerekir [22].

#### 2.1.2.3 Veri İndirgeme (Data Reduction)

Veri madenciliğinde verinin çözümlenmesi bazen çok uzun sürebilir. Bu zaman kaybını önlemek adına veri de eğer mümkünse indirgemeye gidilebilir. Veri kümesi içerisinde aynı tipte çok fazla kayıt olduğu biliniyorsa ve bu kayıtların bazılarının çıkarılmasının sonucu değiştirmeyeceği düşünülüyorsa, verilerin sayısı azaltılabilir. Veri indirgeme için veri sıkıştırma kullanarak verilerin daha az yer kaplamaları sağlanabilir. Veya benzer karakteristiğe sahip birçok özellik yerine daha az özellik kullanılarak işlemlerin daha hızlı yapılması sağlanabilir [22].

#### 2.1.2.4 Veri Dönüştürme (Data Transformation)

Veri Madenciliğinde bazen verileri olduğu gibi kullanmak yerine çeşitli işlemlerden geçirerek kullanmak gerekebilir. Örneğin, bazı özeliklerin ortalamasının, diğer özelliklerin ortalamasından çok büyük veya çok küçük olması durumunda, bu büyük farkı oluşturan özelliklerinin etkisi diğer özelliklerden daha çok olur ve onların rollerini

önemli ölçüde azaltır. Bu durumda veri setinin aynı standartlara sahip olması için Min-Max normalleştirmesi veya Z-score standartlaştırması gibi yöntemler kullanılabilir. Bunların yanında seçilen veri madenciliği modeline göre değişkenlerin aralık değerleri değiştirilerek gruplanabilir. Örnek vermek gerekirse, 0'dan 100'e kadar olan yaş verisi bebek, çocuk, genç, orta-yaş ve yaşlı şeklinde beş farklı gruba bölünerek sistem için daha uygun hale getirilebilir [22].

### 2.1.3 Modelin Kurulması ve Değerlendirilmesi (Evaluation of the Model)

Verilerin hazırlanma aşamasından sonra, analize uygun hale getirilen veriler üzerinden, önceden tespit edilen problemin optimal çözümü için çeşitli modellemeler yapılır. Hangi modelin ve algoritmanın daha iyi performans vereceğinin kestirilmeye çalışıldığı bu süreçte mevcut olan tüm modeller kurulup, karşılaştırılır [23].

Model kurma süreci denetimli (supervised) ve denetimsiz (unsupervised) öğrenmenin kullanılmasına göre farklılık göstermektedir. Denetimli öğrenme sürecinde veri kümesinde önceden tanımlanmış sınıflara ilişkin özelliklerin belirlenmesi ve bu özelliklerin kural cümleleri ile ifade edilmesi hedeflenmektedir. Süreç tamamlandığında bu kural cümleleri yeni verilere uygulanmakta ve bu verilerin hangi sınıfa ait olduğu öngörülmektedir. Denetimsiz öğrenmede ise verilerin benzerliklerinden ya da uzaklıklarından hareketle ait oldukları sınıfların üretilmesi amaçlanmaktadır. Denetimli öğrenme sürecinde modelin belirli bir veri kümesi üzerinde oluşturulması ve kalan verilerle modelin geçerliliğinin sınanması yöntemi kullanılmaktadır. Bunun için veri kümesi en az iki alt gruba ayrılmalı ve birinci grupta model parametrelerinin kestirimi, ikinci grupta ise modelin sınanması sağlanmalıdır. Örneğin sınıflandırma algoritmaları kullanılırken tüm veri kümesi öğrenme ve test kümesi olarak ikiye ayrılmalı; modelin verilerden öğrenerek oluşturulması öğrenme kümesi, doğruluğunun kontrolü ise test kümesi ile gerçekleştirilmelidir [17].

Kurulan modellerde birbiri ile ilişkili veya anlamsız olan değişkenlerin elenmesine dikkat edilmelidir. Amaç bilgi çıkarımı olduğundan ve birbiri ile ilişkili olan değişkenler bize ekstra bilgi vermediğinden, diğerine göre daha anlamlı olan değişkeni modele katmak faydamıza olacaktır.

### 2.1.4 Modelin Kullanılması (Using the Model)

Veri madenciliği modelinin kullanılabilmesi için yukarıda bahsedilen yöntemlerden gerekli olanlar uygulanır ve veri hazır hale geldikten sonra konuyla ilgili veri madenciliği algoritmaları uygulanır.

Veri madenciliği algoritması veriler üzerinde uygulandıktan sonra, sonuçlar düzenlenerek sunumu yapılır. Yapılan sunumlarda çoğu kez grafiklerle desteklenir. Örneğin bir hiyerarşik kümeleme modeli kullanılmış ise sonuçlar dendogram adı verilen özel grafiklerle sunulur [24].

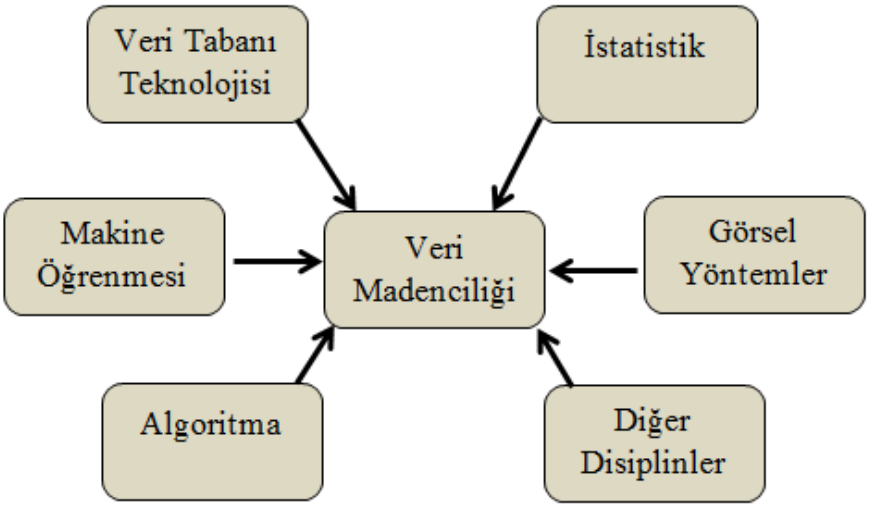
Zaman içerisinde sistemlerin özelliklerinde meydana gelen değişikliklerden dolayı oluşturulan modellerin sürekli olarak izlenmesi ve gerekiyorsa yeniden düzenlenmesi gerekmektedir [25]. O yüzden hazırlanmış bir modelin ya da çalışmanın uzun yıllar güncel kalması ancak iyi gerçekleştirilmiş bir izleme ile olabilir.

## 2.2 Veri Madenciliğinin Farklı Disiplinlerle İlişkisi

Veri madenciliği, önceden bilinmeyen ilişkilerin bulunması için büyük miktarlardaki veriyi tahlil eden bir yolculuktur. Yüksek performanslı bilgisayarlara ve gereken yazılımlara kolay ve düşük fiyatlarla ulaşılabilmesi bu teknolojinin işlenmesini olanaklı kılmıştır.

VM aracılığıyla, büyük veri kümelerini içerisinde barındıran veri tabanı sistemlerinde gizli kalmış bilgilerin çekilmesi sağlanır. Bu işlem, istatistik, matematik, modelleme teknikleri, veri tabanı teknolojisi gibi birçok disiplinin bir araya gelmesi ile yapılır [26].

Makine öğrenmesi, istatistik ve VM arasında sıkı bir ilişki vardır. Bu üç disiplinde veri içindeki örüntüleri tespit etmeyi amaçlar. Makine öğrenmesinde yer alan yöntemler, VM algoritmalarında kullanılan yöntemlerin çekirdeğini teşkil eder. Bunun yanında veri tabanı teknolojisi ve örüntü tanıma yöntemleri de veri madenciliğinin bağlantılı olduğu alanlardır. Şekil 2.2’de VM’ nin ilişkili olduğu disiplinler gösterilmektedir [27].




Şekil 2.2 Veri madenciliğinin ilişkili olduğu disiplinler [27]

## 2.3 Veri Madenciliği Uygulama Alanları

Günümüzde telekomünikasyon firmaları ve bankalar başta olmak üzere birçok organizasyon veri madenciliği disiplininden yararlanmaktadır. Veri madenciliği, pazarlama, üretim, tıp, astronomi, biyoloji, finans, mühendislik bilimleri vs. gibi pek çok bilim dalında uygulanmaktadır.

Veri madenciliğinin pazarlamadaki öncelikli uygulama alanı, farklı tüketici gruplarını belirlemek ve onların davranışlarını tahmin etmek için, tüketici veri tabanlarını analiz eden veri tabanı market sistemleridir [16]. Yaygın olarak kullanılan veri madenciliği uygulama alanları aşağıda listelenmektedir:

a) **Bankacılık, Sigortacılık ve Telekomünikasyon**
   - Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
   - Müşteri dağılımında,
   - Usulsüzlük tespiti,
   - Hatların yoğunluk tahmini,
   - Hisse tespitleri,
   - Kalite ve iyileştirme analizleri,
   - Sahtekârlık tespiti,
   - İlişki tespiti,
   - Risk yönetimi ve risk analizleri [28],
   - Hisse senedi fiyat tahmini,
   - Piyasa analizleri,

- Müşteri analizi,
- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,
- Riskli müşteri örüntülerinin belirlenmesi [25].

**b) Pazarlama**

- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması [28],
- Reklam politikalarının etkinliğinin ölçümü,
- Hedef pazar bulma,
- Tüketici davranışlarını inceleme,
- Tüketici profili ve ihtiyaçlarını belirleme,
- Sepet analizleri,
- Mağaza yerleşimi,
- Promosyon ve kuponların etkisini ölçme,
- Müşteri ilişkileri yönetimi,
- Satış tahmini,
- Satış kampanyalarının verimlilik analizlerinin yapılması [29],
- Pazar ve sektör özelliklerini belirleme,
- Bağımsız firma ve stok performansını tahmin etme.
- Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında

**c) Tıp**

- Cerrahi işlemlerin, tıbbi test ve ilaçların etkisini tahmin etme,
- Hastalık teşhisi,
- Ürün geliştirme
- Tedavi sürecinin belirlenmesi
- İlaç geliştirme,
- Semptomlara göre hastalık tespiti [28].

**d) Üretim Yönetimi ve İmalat**

- Kalite kontrol,
- Üretim süreçlerinin optimizasyonu,
- Sorun giderme,

- Bakım-onarım,
- Lojistik [29].

**e) Mühendislik uygulamaları**

**f) Astronomi ve gökyüzü nesnelerinin tespiti**

**g) Web uygulamaları**

Yukarıda sayılan veri madenciliği uygulama alanlarının sektörler bazında tespiti için 2003 yılında yapılmış bir araştırmanın sonuçları Çizelge 2.1'de [30] yer almaktadır. Parantez içerisinde belirtilen sayılar veri madenciliğinin o sektördeki kullanım sayısını göstermektedir.

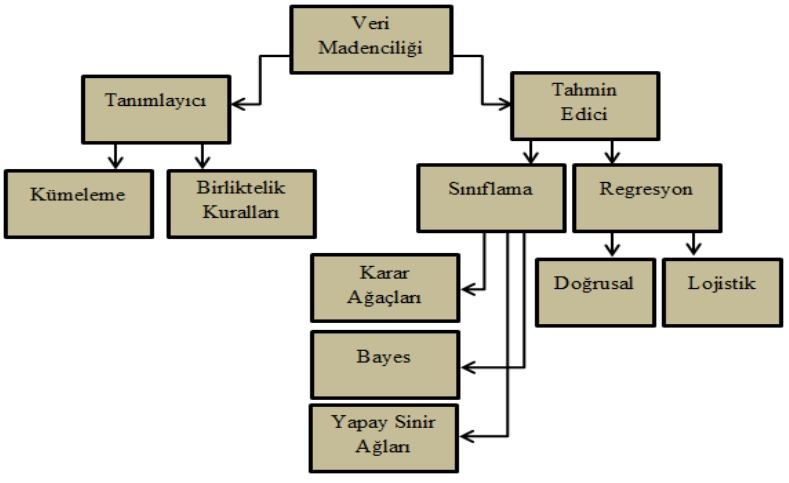
Çizelge 2.1 Veri madenciliğinin sektörler bazında kullanımı [30]

| **Veri Madenciliğinin Sektörler Bazında Kullanımı** | |
|---|---|
| Bankacılık(51) | 12% |
| CRM(52) | 12% |
| Kredi Skorlama(35) | 8% |
| Doğrudan Pazarlama(34) | 8% |
| Sahtekârlık Tespiti(31) | 7% |
| Sigortacılık(24) | 6% |
| Telekomünikasyon(23) | 5% |
| İmalat(19) | 5% |
| Bilim(17) | 4% |
| Sağlık/İK(15) | 4% |
| Kamu Uygulamaları(12) | 3% |
| Tıp/Farmakoloji(12) | 3% |
| Bioteknoloji/Genetik(11) | 3% |
| E-Ticaret(11) | 3% |
| Web(9) | 2% |
| Seyahat(8) | 2% |
| Yatırım/Hisse Senedi(5) | 1% |
| Junk E-mail/Anti-Spam(5) | 1% |
| Güvenlik/Anti-Terörizm(5) | 1% |
| Şans Oyunu(0.01) | 0.01% |
| Diğer(11) | 1% |

## 2.4 Veri Madenciliği Modelleri

Veri madenciliği çalışmalarında birçok teknik ve yöntem kullanılmaktadır. Kullanılacak yöntemin belirlenmesi problemin tanımına ve verinin yapısına göre değişir. Bu yüzden veri madenciliğinde kullanılacak yöntemler için en iyi yöntem, en başarılı yöntem veya en iyi algoritma gibi bir kavramdan söz edilemez.

Veri madenciliği yöntemleri, tahmin edici ve tanımlayıcı olmak üzere iki ana başlık altında toplanmaktadır. Tahmin edici modeller, keşfe dayalı modellerdir. Sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edilmesi amaçlanmaktadır [25]. Sınıflama ve regresyon veri madenciliğinin tahmin edici modellerinden ikisidir. Tanımlayıcı modellerde ise karar vermeye yardım edebilecek, mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. X/Y aralığında geliri ve iki veya daha fazla arabası olan çocuklu aileler ile çocuğu olmayan ve geliri X/Y aralığından düşük olan ailelerin satın alma örüntülerinin birbirlerine benzerlik gösterdiğinin belirlenmesi tanımlayıcı modellere bir örnektir [25]. Kümeleme, birliktelik kuralı ve ardışık örüntü veri madenciliğinin tanımlayıcı tekniklerinden bazılarıdır. Şekil 2.3'de veri madenciliği modellerinin hiyerarşisi yer almaktadır:




Şekil 2.3 Veri Madenciliği Modelleri

### 2.4.1 Sınıflama ve Regresyon (Classification and Regression)

Sınıflama ve regresyon, gelecek veri eğilimlerini tahmin eden modelleri kurabilen veri analiz yöntemleridir. Sınıflama; belli özelliklere göre gruplanmış değerleri tahmin ederken, regresyon; mevcut değerleri kullanarak diğer değerlerin ne olacağını tahmin etmeye çalışır.

Sınıflama algoritması, ortak özelliklere sahip kayıtların farklı sınıflar içine aktarılmasını belirleyen algoritmadır. Sınıf olmak için her kaydın sınıf içinde yer alan diğer kayıtlarla belirlenmiş bir ortak özelliği olması gerekir [31].

Sınıflama kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır. Sınıflama, verinin önceden belirlenen çıktılara uygun olarak ayrıştırılmasını sağlayan bir tekniktir. Çıktılar, önceden bilindiği için sınıflama, veri kümesini denetimli olarak öğrenir.

Sınıflama veri madenciliğinde sıkça kullanılan bir yöntem olup, veri tabanlarındaki gizli örüntüleri ortaya çıkarmakta kullanılır. Verilerin sınıflandırılması için belirli bir süreç izlenir. Öncelikle var olan veri tabanının bir kısmı eğitim amacıyla kullanılarak sınıflandırma kurallarının oluşturulması sağlanır. Daha sonra bu kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir.

Sınıflama sorgusu kullanılarak bir kaydın daha önceden nitelikleri belirlenmiş bir sınıfa girmesi amaçlanmaktadır Sınıflama algoritması öğrenme verilerini kullanarak hangi sınıfların var olduğu ve bu sınıflara girebilmek için kayıtların hangi özelliklere sahip olması gerektiğini otomatik olarak keşfeder.

Sınıflandırma ile ilgili bir uygulama kısaca şöyledir:
"Genç kadınlar küçük araba satın alır, yaşlı, zengin erkekler büyük, lüks araba satın alır." Amaç bir malın özellikleri ile müşteri özelliklerini eşlemektir. Böylece bir müşteri için ideal ürün veya bir ürün için ideal müşteri profili çıkarılabilir. Örneğin bir otomobil satıcısı şirket, geçmiş müşteri hareketlerinin analizi ile yukarıdaki gibi iki kural bulursa genç kadınların okuduğu bir dergiye reklam verirken küçük modelinin reklamını verir [13]. Örneğin sınıflandırma modeli banka kredi uygulamalarının güvenli veya riskli olmalarını sınıflamak amacıyla kullanılırken, regresyon modeli geliri ve mesleği bilinen potansiyel müşterilerin herhangi bir alışveriş esnasında yapacakları harcama davranışlarını tahmin etmek için kullanılabilir [32],

Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler şunlardır [25]:

- Genetik algoritmalar
- K-en yakın komşu
- Bayes sınıflandırıcıları,
- Karar ağaçları

- Yapay sinir ağları

### 2.4.2 Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler (Association Rules and Sequential Patterns)

Birliktelik kuralları, veri kümesindeki nesneler arasındaki ilişkiyi ortaya koymaktır. Ayrıca birliktelik kuralları, veri kümesi içindeki sıklıkla bir arada görülen örüntülerin tespit edilmesi işlemidir.

Birbiri ardına yapılmış alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, ürün satışlarını artırabilecek önemli bir bilgidir. Satın alma eğilimlerinin belirlenmesini sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, market sepeti analizi adı altında veri madenciliğinde yaygın olarak kullanılmaktadır. Aynı zamanda bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi gereken durumlarda önem taşımaktadır [36]. Pazar sepet analizleri yardımıyla bir müşteri herhangi bir ürünü aldığında, sepetine başka hangi ürünleri de koyduğu bir olasılığa göre tahmin edilir. Birlikte satın alınan ürünler belirlendiğinde, mağazalarda raflar ona göre düzenlenerek müşterilerin bu tür ürünlere daha kolayca erişmeleri sağlanabilir. Bu modelin en bilinen temsilcisi "Apriori" algoritmasıdır. Birliktelik kuralları aşağıda sunulan örneklerde görüldüğü gibi eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılır[25].

- Müşteriler bira satın aldığında, % 75 ihtimalle çocuk bezi de alırlar,
- Az yağlı peynir ve yağsız yoğurt alan müşteriler, %85 ihtimalle diyet süt de satın alırlar.

Birliktelik kuralı, bir ilişkide bir niteliğin aldığı değerler arasındaki bağımlılıkları, anahtarda yer almayan diğer niteliklere göre gruplama yapılmış verileri kullanarak bulur. Keşfedilen örüntüler veri kümesinde sıklıkla birlikte geçen nitelik değerleri arasındaki ilişkiyi gösterir [37]. Örneğin bir süper markette ekmek ve peynir satın alınan satış hareketlerinin % 75 'inde zeytin de satın alınmıştır. Bu tür birliktelik örüntüleri ancak örüntüde yer alan öğelerin birden fazla harekette tekrarlanması ile gerçekleşebilir [37].

Ardışık örüntü ise birbirleri ile ilişkisi olan ve birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır. Belli bir olayın gerçekleşmesinden sonra aynı

olayla ilintili bir başka olayın gerçekleşmesi, ardışık örüntü kavramını oluşturur. Örneğin;

- Otomobil alan bir kişinin kısa vadede otomobili için aksesuar alması ardışık bir örüntüdür [32].
- IMKB endeksi düşerken A hisse senedinin değeri % 15'den daha fazla artacak olursa, üç iş günü içerisinde B hisse senedinin değeri % 60 ihtimalle artacaktır[32].

# BÖLÜM 3

# UYGULAMADA KULLANILAN KÜMELEME VE SINIFLANDIRMA YÖNTEMLERİ

## 3.1 Kümeleme (Clustering)

Verileri, diğer verilerle olan benzerliklerini veya farklılıklarını tarif eden bilgileri kullanarak gruplara ayırma işlemine kümeleme denir. Kümelemede amaç; grup içindeki nesneleri, diğer gruplardaki nesnelerden olabildiğince bağımsız, kendi aralarında ise birbirine bağımlı olacak şekilde oluşturmaktır [29]. Bu noktada tanımlayıcı bir veri madenciliği yöntemi olan kümeleme devreye girmekte ve veriyi çeşitli tekniklerle önceden sayısı bilinmeyen kümelere bölmektedir [27].

Veriler, veri tabanındaki veri grubuna göre farklı kümelerde yer alabilir. Veriler, sadece taşıdıkları değerlere göre değil veri tabanındaki diğer verilerin değerlerine göre de değerlendirildikleri için kümeleme sonuçları dinamiktir ve bu özellik kümelemeyi sınıflamadan ayırır.

Kümeleme yönteminde bir çıktı değişkeni yoktur. Bu sebeple denetimsiz öğrenme metodu olarak bilinmektedir. Bu noktada kümeleme, veri kümelerinde verileri birbirinden ayıran başka bir yöntem olan diskriminant analizinden ayrılmaktadır. Zira kümelemede küme sayısı bilinmemekte ve analiz sonucunda veriden elde edilmektedir. Bununla birlikte kümelemede herhangi bir fonksiyon elde edilerek sonrasında diğer veriler için kullanılma durumu yoktur; çünkü ayırma işlemi tamamen o verilerin özellikleri kullanılarak yapılır.

Sınıflama ile kümelemeyi birbirinden ayıran en önemli etken, kümeleme işleminin sınıflama işleminde olduğu gibi önceden belirlenmiş bir takım sınıflara göre bölme yapmamasıdır. Sınıflamada her bir veri, önceden sınıflandırılmış bir takım sınıflar üzerinde yapılan bir eğitim neticesinde ortaya çıkan modele göre bir sınıfa atanmaktadır. Kümeleme işleminde ise önceden tanımlanmış sınıflar ya da örnek

sınıflar bulunmamaktadır. Verilerin kümelenmesi işlemi, verilerin birbirlerine olan benzerliklerine göre yapılmaktadır. Oluşan sınıfların hangi anlamları taşıdığının belirlenmesi tamamen çözümlemeyi yapan kişiye kalmıştır.

Kümeleme süreci, her bir verinin farklı özelliklerini ifade eden değerlerinin, diğer bir veriden farklı ya da aynı olmasına göre işler. Farklılık ya da aynılık kümeleme algoritmaları için uzaklıkları ifade etmektedir. Kümeleme algoritmaları, Öklid uzaklığı, Minkowski uzaklığı, Manhattan uzaklığı gibi farklı uzaklıkları kullanabilirler.

### 3.1.1 Kümeleme Yöntemlerinde Uzaklık Ölçütleri

Kümeleme yöntemlerinin büyük çoğunluğu veriler arasındaki uzaklıkların hesaplanması esasına dayanmaktadır. Bu nedenle iki veri arasındaki uzaklığın hesaplanmasını sağlayan bağıntılara ihtiyaç duyulmaktadır. Çeşitli değişkenlerden oluşan bir veri setini X matrisi biçiminde göstermek mümkündür. Örneğin dört veri ve bu verilere ait dört özellikten oluşan bir veri seti matris ile şu şekilde gösterilir [24]:

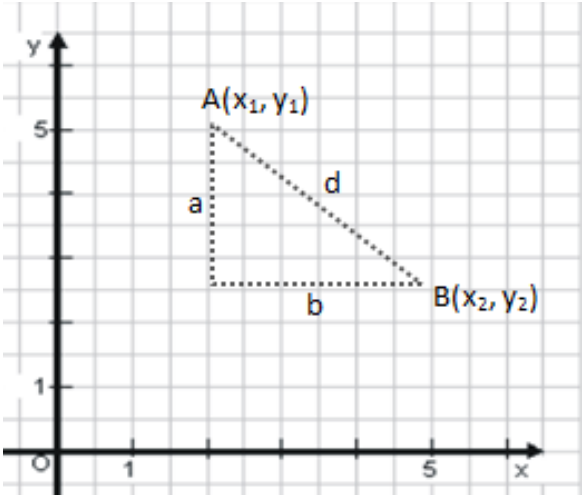
$$X=\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{bmatrix} \tag{3.1}$$

Yukarıda ki matriste birinci verinin konumu ($x_{11}$, $x_{12}$, $x_{13}$, $x_{14}$) biçimindedir. İkinci verinin konumu ise ($x_{21}$, $x_{22}$, $x_{23}$, $x_{24}$) olarak ifade edilmektedir. Bu iki verinin birbirlerine olan uzaklıkları ise d(1,2) biçiminde gösterilir. Yukardaki matriste her bir verinin diğerine olan uzaklığı d(i, j) biçiminde gösterilir. Ve D uzaklık matrisi aşağıdaki biçimde gösterilir [24]:

$$D=\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} \tag{3.2}$$

### 3.1.1.1 Öklid Uzaklığı

Uygulamalarda sıklıkla kullanılan yöntem Öklid uzaklığı yöntemidir. Bu uzaklık aşağıdaki şekilde gösterildiği gibi, iki boyutlu uzayda Pisagor teoreminin bir uygulaması olarak karşımıza çıkmaktadır [24].




Şekil 3.1 İki nokta arasındaki d uzaklığı

A ve B noktaları arasındaki Öklid uzaklığı şu şekilde olacaktır:

$$d(A, B)= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (3.3)$$

Bu bağıntı p boyutlu i ve j noktaları için genelleştirilirse:

$$d(i,j)= \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \quad (3.4)$$

### 3.1.1.2 Manhattan Uzaklığı

Diğer bir uzaklık ölçüsü Manhattan uzaklığıdır. Bu uzaklık, veriler arasındaki mutlak uzaklıkların toplamı alınarak hesaplanır. Söz konusu uzaklık şu şekilde ifade edilir [24]:

$$d(i,j)=\sum_{k=1}^{p}|x_{ik} - x_{jk}| \quad i, j=1,2\ldots.n; \quad k=1,2\ldots.p \quad (3.5)$$

### 3.1.1.2 Minkowski Uzaklığı

p sayıda değişken göz önüne alınarak veriler arasındaki uzaklığın hesaplanması söz konusu ise Minkowski uzaklık bağıntısı kullanılabilir. Bağıntı p boyutlu i ve j noktaları için şu şekilde hesaplanır [24]:

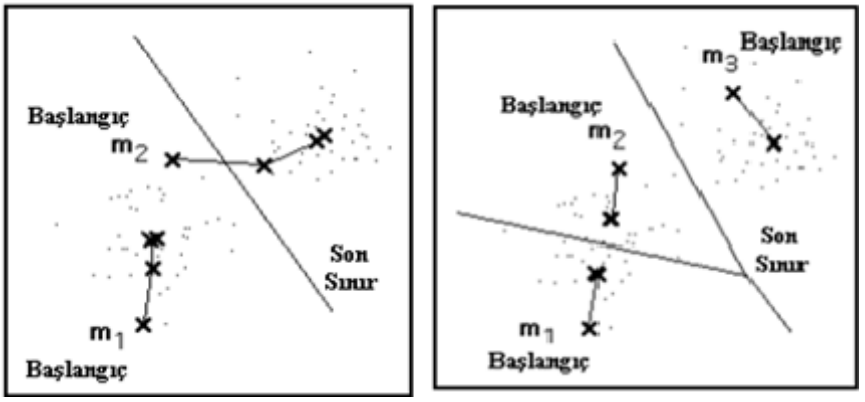
$$d(i,j)= \left[ \sum_{k=1}^{p} |x_{ik} - x_{jk}|^{m} \right]^{1/m} \quad i, j=1,2\ldots.n; \; k=1,2\ldots.p \qquad (3.6)$$

### 3.1.2 Katı Kümeleme Algoritmaları

#### 3.1.2.1 K-ortalamalar Algoritması

K-ortalamalar kümeleme algoritması 1967 yılında J.B. MacQueen tarafından geliştirilmiş en eski kümeleme algoritmalarından birisidir [38]. En yaygın kullanılan denetimsiz öğrenme yöntemlerinden biridir. K-ortalamaların atama mekanizması her verinin sadece bir kümeye ait olabilmesine izin verir [39].

K-ortalamalar algoritması, merkez noktanın kümeyi temsil etmesi ana fikrine dayalı bir metottur [16]. Eşit büyüklükte küresel kümeleri bulmaya eğilimlidir [16]. K-ortalamalar yönteminde her biri farklı bir kümenin merkezini işaret eden k tane nesne belirlenir. Kalan diğer nesneler, belirlenen bu küme merkezlerine olan uzaklıklarına göre kümelere dâhil olurlar. Daha sonra oluşturulan bu kümelere göre yeni küme merkezlerini tespit etmek için her bir kümenin ortalama değerleri ayrı ayrı hesaplanır. Bulunan bu yeni küme merkezlerine göre nesnelerin tekrardan her bir kümeye olan uzaklığına bakılır. Şekil 3.2'de görüldüğü gibi kümelerde yer alan nesnelerde herhangi bir değişim olmayıncaya kadar algoritma çalışmaya devam eder.




**(a)** **(b)**

Şekil 3.2 k-ortalamalar algoritmasının (a) k=2 için; (b)k=3 için ötelenişi [47]

Başlangıç küme merkezlerinin seçimi k-ortalamalar'ın sonucunu önemli oranda etkiler. Başlangıç noktalarının belirlenmesinde çeşitli teknikler vardır. Bu tekniklerden bazıları aşağıdaki gibidir [40]:

- k sayısı kadar rastgele veri seçilip küme merkezleri olarak atanır.

- Veriler rastgele k tane kümeye atanır ve küme ortalamaları alınarak başlangıç küme merkezleri belirlenir.
- En uç değerlere sahip veriler küme merkezleri olarak seçilir.
- Veri setinin merkezine en yakın noktalar başlangıç noktaları olarak seçilir.

K-ortalamalar kümeleme algoritmasının değerlendirilmesi yapılırken en yaygın kullanılan yöntem karesel hata değeridir. En düşük karesel hata değerine (SSE) sahip kümeleme sonucu en iyi sonucu verir. Nesnelerin bulundukları kümenin merkez noktasına olan uzaklığının karelerinin toplamı (3.7)'ye göre hesaplanmaktadır [41].

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x) \quad (3.7)$$

x: $C_i$ kümesinde bulunan nesne,
$m_i$: $C_i$ kümesinin merkez noktası
k: Küme sayısı

Burada, dist iki nesne arasındaki standart Öklid Uzaklığı, x değeri $C_i$ kümesinde bulunan bir nesne, $m_i$ değeri $C_i$ kümesinin merkez noktasıdır. Yukarıda açıklanan k-ortalamalar algoritması, iki boyutlu veriler üzerinde Öklid Uzaklık ölçütüne göre çalışmaktadır ve hiçbir nesne kümesini terk etmeyene kadar ötelenmektedir.

K-ortalamalar algoritmasının avantajı kolay uygulanabilir olması ve büyük veri setlerinde hızlı sonuç üretmesidir. K-ortalamalar algoritması birbirinden uzak ancak küme içerisinde birbirine yakın nesnelerden, yani yoğunlaşmış nesnelerden oluşuyorsa iyi sonuçlar üretir. K-ortalamalar yönteminin zayıf yönlerinin başında küme sayısının belirlenmesi gelmektedir. Yöntemin en başında kullanıcı tarafından k küme sayısının belirlenmiş olma zorunluluğu vardır. Bir diğer zayıf noktası kategorik özelliklerin yer aldığı kimi uygulamalarda gerçekleştirilememektedir. Ayrıca k-ortalamalar yöntemi küresel olmayan, farklı yoğunluklara ve büyüklüklere sahip kümeler için uygun bir yöntem değildir [42].

#### 3.1.2.1 K-medoids Algoritması

K-medoids algoritması, verinin farklı yapısal özelliklerini temsil eden k tane temsilci nesne bulma esasına dayanır [61]. Temsilci nesne yönteme ismini veren medoid olarak adlandırılır ve kümenin merkezine en yakın nokta olma özelliği taşır. Amaç k tane nesneyi bulmak olduğundan dolayı, k-medoids metodu olarak adlandırılmaktadır.

K-medoids kümeleme yönteminin temel stratejisi ilk olarak n adet nesnede, merkezi temsili bir medoid olan k adet küme bulmaktır[11]. Geriye kalan nesneler, kendilerine en yakın olan medoide göre k adet kümeye yerleşirler. Bu bölünmelerin ardından kümenin ortasına en yakın olan nesneyi bulmak için temsilci nesne, temsilci olmayan her nesne ile yer değiştirir. Bu işlem en verimli medoid bulunana kadar devam eder [11].

**Algoritmanın Adımları:**

**Adım 1:** Veri seti içerisinden medoid olarak rastgele k tane birey seçilir.

**Adım 2:** *i* → medoid olarak seçilmiş bireyler, *h* → medoid olarak seçilmemiş bireyler olmak üzere bunların yerlerinin değiştirildiğini varsayalım. i ve h arasındaki uzaklık $d(x_i, x_h)$ ile gösterilir.

Toplam yer değiştirme katsayısı $T_{ih}$ şöyle hesaplanır

$$T_{ih} = \sum_j c_{jih} \tag{3.8}$$

Burada $c_{jih}$ j. birey için i 'nin h'a katkısıdır ve aşağıdaki gibi 4 farklı şekilde hesaplanabilir.

- j. bireyin i. medoid tarafından tanımlanan kümeye girdiğini ve j. ve h. Bireyler arasındaki uzaklığın d($x_j$-$x_h$) olduğunu varsayalım. Bu durumda; eğer h j'ye ikinci en iyi medoid *i* 'sinden daha uzaksa:
  $$C_{jih} = d(x_j - x_i) - d(x_j - x_h) \tag{3.9}$$
- Eğer h j'ye *i* 'den daha yakınsa
  $$C_{jih} = d(x_j - x_h) - d(x_j - x_i) \tag{3.10}$$
- Eğer j, k. kümeye giriyorsa $k \neq i$ j. birey ile h. birey arasındaki uzaklığın kontrol edilmesi gerekir.
- Eğer h ile j arasındaki uzaklık k. medoid ile j arasındaki uzaklıktan daha büyükse o zaman j. bireyin katkısı $C_{jih} = 0$
- Eğer h ile j arasındaki uzaklık k. medoid'le j arasındaki uzaklıktan daha küçükse o zaman j. bireyin katkısı
  $$C_{jih} = d(x_j - x_h) - d(x_j - x_k) \tag{3.11}$$

**Adım 3:** $(i^*, h^*) = argmin_{i^*,h^*} T_{ih}, T_{ih} < 0$ ise i* ile h* yer değiştirir.

Adım 2'ye geri dön.

**Adım 4:** Medoid olarak seçilmemiş bireyleri en yakın medoid'e göre kümelere yerleştir.

Bu yöntemi kullanmanın dezavantajları ise şu şekildedir:

- Bu algoritmaya girdi değeri olarak k küme sayısının verilmesi gerekmektedir. Bu nedenle iyi bir kümeleme elde etmek için k sayısının ne olduğuna karar vermek gerekir. Bu seçimin kullanıcıya bağlı olması bu yöntemin dezavantajıdır.
- Zaman karmaşıklığı, algoritmanın her yinelemesi için $O(n^2)$'dir.
- Medoid'ler kullanıldığından, kümelerin başka herhangi bir şekilde tanımlanmasını sağlamaz[56].

### 3.1.3 Bulanık Kümeleme Algoritmaları

#### 3.1.3.1 Bulanık C-Ortalamalar Algoritması

Bulanık c-ortalamalar(BCO) algoritması 1973 yılında Dunn tarafından ortaya atılmış ve 1981'de Bezdek tarafından geliştirilmiştir [43]. BCO algoritması, bulanık bölünmeli kümeleme yöntemlerinden en sık kullanılanıdır. BCO yöntemi nesnelerin katı kümeleme yöntemlerinde olduğu gibi tek bir kümeye ait olmak yerine iki veya daha fazla kümeye ait olabilmesine izin verir[44]. Bulanık mantık yaklaşımı uyarınca her veri, kümelerin her birine [0,1] arasında değişen bir üyelik değeri ile bağlıdır. Ve bu üyelik değerleri toplamı her zaman 1 olmalıdır. Bir nesnenin üyelik değeri hangi küme için en büyükse o kümeye ait olacaktır. Çoğu bulanık kümeleme algoritması amaç fonksiyon tabanlıdır. Kümelemenin sonlanması algoritmanın amaç fonksiyon için belirlenmiş minimum değişim değerine yakınsaması gerekmektedir[43].

BCO algoritmasının üyelik matrisi özelliği belirsiz durumların tanımlanmasını kolaylaştırır [45]. Ayrıca aykırı değerlerin üyelik dereceleri düşük olacağından kümeleme işlemine etkisi azdır. Bunun yanında esnek bir yapıya sahip olması ve benzeşen kümeleri bulma yeteneği diğer bölünmeli yöntemlere göre daha fazladır.
Üyelik fonksiyonu işlemsel yükü ve karmaşıklığı arttırdığı için zaman açısından diğer yöntemlere göre maliyetli bir bölünmeli kümeleme algoritmasıdır. Temel olarak k-ortalamalar algoritmasına çok benzemekle beraber BCO'nun, k-ortalamalar'dan en önemli farkı verilerin her birinin sadece bir sınıfa dâhil edilmesi yerine her bir sınıfa belli bir üyelik değeri ile bağlı olmasıdır.

BCO algoritması (3.12)'deki amaç fonksiyonunu öteleyerek minimize etmek için çalışır [46];

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad 1 \leq m < \infty \tag{3.12}$$

BCO Kümeleme Algoritması için Gerekli Adımlar:

X veri seti olmak üzere, 1<c<N sayıda küme sayısı seçilir. Bulanıklaştırma katsayısı m>1, işlem sonlandırma kriteri ε >0 olacak şekilde belirlenir.

**Adım 1:** U üyelik matrisi rastgele atanarak algoritma başlatılır.

**Adım 2:** İkinci adımda ise merkez vektörleri hesaplanır. Merkezler 3.13' deki formüle göre hesaplanır [46].

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{3.13}$$

**Adım 3:** Hesaplanan küme merkezlerine göre U matrisi 3.14'deki formül yardımıyla yeniden hesaplanır. Eski U matrisi ile yeni U matrisi karşılaştırılır ve fark ε'dan küçük olana kadar işlemler devam eder [46].

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left( \frac{\left\| x_i - c_i \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \tag{3.14}$$

Kümeleme işlemi neticesinde U üyelik matrisi kümelemenin sonucunu yansıtır. İstenirse, berraklaştırma yapılmak suretiyle bu değerler yuvarlanarak 0 ve 1'lere dönüştürülebilir.

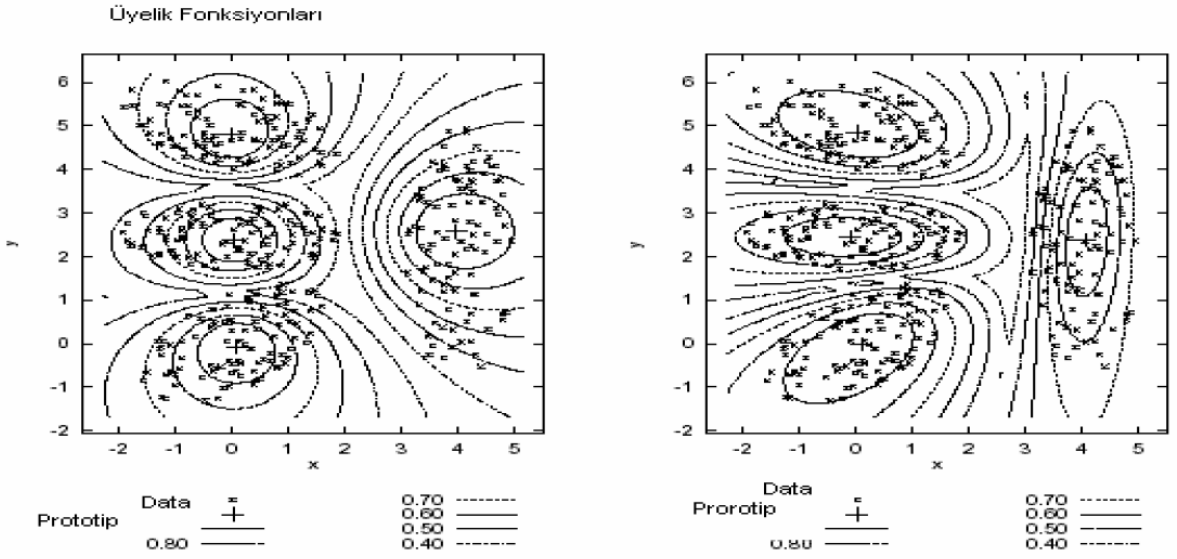
Bu adımların tümü yapıldıktan sonra nesneler, kümeleme sonucu elde edilen üyelik dereceleri ile beraber, ilgili bulanık kümelere atanır. Yapılacak bazı hesaplamalarda buradan da kesin üyeliklere geçmek istenebilir. Bunun için her bir kümede en fazla üyelik derecesine sahip olan elemanın üyeliği 1'e, diğerlerinin ki de 0'a eşitlenerek durulaştırma yapılabilir. Durulaştırma, bulanık durumdan kesin hale geçmektir.

Bulanık kümelemenin kesin kümeleme tekniklerine göre avantajı, veri hakkında daha detaylı bilgi ortaya koyabilmesidir. Diğer taraftan dezavantajları da vardır. Çok sayıdaki birey ve küme durumunda çok fazla çıktı olacağından, özetlemek ve bilgiyi tasnif etmek

zordur. Ayrıca bulanık kümeleme algoritmaları genellikle karmaşıktır ve daha çok belirsizlik söz konusu olduğunda kullanılır[55].

### 3.1.3.2 Gustafson-Kessel algoritması

Gustafson ve Kessel veri setleri içerisindeki farklı geometrik şekilleri tespit edebilmek için standart BCO algoritmasını, adaptif bir uzaklık normu kullanarak genişletmişlerdir [58]. Bulanık C-Ortalamalar algoritması küresel yapıya sahip kümeler için uygunken elipsoidal kümeler için uygun değildir. Gustafson ve Kessel algoritması ise yapısı itibariyle bu tip kümeler için uygundur. Bu algoritmanın, BCO algoritmasından farkı küme merkezlerine ek olarak her bir kümenin simetrik ve pozitif tanımlı bir A matrisine sahip olmasıdır. Bu matris her küme için $\|x\|_A = \sqrt{x^T A x}$ normuna sebep olur. Burada A matrisi optimizasyon değişkeni olarak kullanılmıştır, böylece her bir kümenin, verinin yerel topolojik yapısının uzaklık normuna adapte olmasını sağlamıştır [58]. Matrislerin isteğe bağlı olarak seçilmesinden dolayı uzaklıklar küçük değerler alabilmektedir. GK algoritmasında, bulanık kovaryans matrisinin sınırlandırılması gerekmektedir. Sıfır (ya da yaklaşık sıfır) değerli matrislerin amaç fonksiyonunu minimize etmesini önlemek için, det(A) = 1 olacak şekilde sabit hacimli kümeler gerekmektedir. Böylece küme büyüklükleri sabit kalmakla beraber, küme şekilleri değişken hale gelmektedir. A matrisinin, determinantıyla oluşturulan varyasyonlarla, hacimleri sabit kalacak şekilde en uygun küme şekillerinin belirlenmesi amaçlanmaktadır[43].




Şekil 3.3 BCO Algoritması ve GK Algoritması [56]

Yukarıdaki şekilde BCO ve GK algoritmalarının bir veri seti üzerindeki kümeleme biçimleri verilmiştir. Şekillere bakıldığında BCO algoritmasının elipsoidal yapıya sahip bu veri setinde GK algoritması kadar başarılı olamadığı gözükmektedir[56].

Gustafson-Kessel algoritmasının yapısı aşağıdaki gibidir;

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_i\|}{\|x_i - c_k\|}\right)^{2/m-1}} \tag{3.15}$$

$d^2$:DXC $\rightarrow$ R,(x,(v,A)) $\rightarrow$ $(x\text{-}v)^T A\ (x\text{-}v)$ (Mahalanobis Uzaklığı) (3.16)

$$J(X, u, v) = \sum\sum u_{ik}^{m} (x_k - v_i)T\ A_i\ (x_k - v_i) \tag{3.17}$$

$$A_i = \sqrt{\det(Si)}\ S_i^{-1} \tag{3.18}$$

$$S_i = \sum u_{ij}^{m}\ (x_j - v_i)(x_j - v_i)^T \tag{3.19}$$

Gustafson-Kessel Algoritması $S_i$ yerine, bulanık kovaryans matrisleri olarak adlandırılan aşağıdaki eşitliğe göre hesaplanan matrisleri kullanır.

$$F^i = \frac{\sum_{j=1}^{n} u_{ij}^{m}(x_j - v_i)(x_j - v_i)^T}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{3.20}$$

Ancak $\frac{1}{\sum_{j=1}^{n} u_{ij}^{m}}$ sonucu etkilemez. Çünkü matrisler determinantları 1 olacak şekilde ayarlanmıştır. J(X, u, v) amaç fonksiyonu minimize edildiğinde;

$$v_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{3.21}$$

GK Kümeleme Algoritması için Gerekli Adımlar:

X veri seti olmak üzere, 1<c<N sayıda küme sayısı seçilir. Bulanıklaştırma katsayısı m>1, işlem sonlandırma kriteri ε >0 ve norm matrisi A olacak şekilde belirlenir. U üyelik matrisi rastgele seçilerek başlatılır.

**Adım 1:** Bulanık küme merkezlerini her küme için hesapla (3.21)

**Adım 2:** Bulanık kovaryans matrisini her küme için hesapla (3.20)

**Adım 3:** (3.16) formülünü kullanarak her bir birey için Mahalanobis uzaklığını hesapla
**Adım 4 :** (3.15) formülünü kullanarak yeni üyelik değerleri matrisini hesapla
**Adım 5:** Yeni üyelik değerleri ile eski üyelik değerlerini karşılaştır.$\left\|U^{(t)} - U^{(t-1)}\right\| < \epsilon$ ise iterasyonu durdur. Aksi takdirde Adım 1'ye geri dön[57].

### 3.1.3.3 Gath-Geva algoritması

Gath ve Geva algoritması, kümelerin yoğunluğunu ve boyutlarını da dikkate alan Gustafson-Kessel algoritmasının geliştirilmiş bir versiyonudur. Yoğunluk odaklı çalıştığı için aykırı özellikler taşıyan veriler içeren yapılar için daha uygun bir algoritmadır [58]. Gerçekte bu yaklaşım, amaç fonksiyonunu optimize etmeye dayanmaz. GG algoritmasının temel prensibi, veri noktalarının p-boyutlu normal dağıldığını varsaymaktır [56]. Gath ve Geva, bulanık maksimum benzerlik kümeleme algoritmasının farklı küme şekillerini, boyutlarını ve yoğunluklarını saptayacağını söylemişlerdir [58]. Bu sebepten GK algoritmasından farklı olarak, bu uzaklık ölçüsü üstel terim içermektedir ve kümeler hacimle sınırlandırılmamıştır. Bununla birlikte bu algoritma, üstel uzaklık ölçü biriminin kullanılması sebebiyle başlangıç koşullarına çok duyarlıdır ve iyi bir başlangıç yapılmasını gerektirir. Aksi takdirde yerel optimumlar üretecektir [59].

$x_k$ veri noktaları, k={1,2,..,n}, $N_i$ ,i={1,2,…,c} normal dağılımlar, klasik bir üyelik fonksiyonu düşünecek olursak $u_{ik} \in \{0,1\}$. Bu durumda istatistikten bilinen i. normal dağılımın ortalama değeri;

$$v_i = \frac{\sum_{k=1}^{n} u_{ik} x_k}{\sum_{k=1}^{n} u_{ik}} \tag{3.22}$$

ve uygun kovaryans matrisi;

$$A_i = \frac{\sum_{k=1}^{n} (x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^{n} u_{ik}} \tag{3.23}$$

Bu eşitlikler Gustafson-Kessel algoritmasından elde edilen eşitliklerle benzerdir. Bulanık üyelik dereceleri için olasılık teorisinden sonuçların nasıl elde edileceği açıktır. Verilere göre elde edilen önsel olasılıklara $P_i$ dersek;

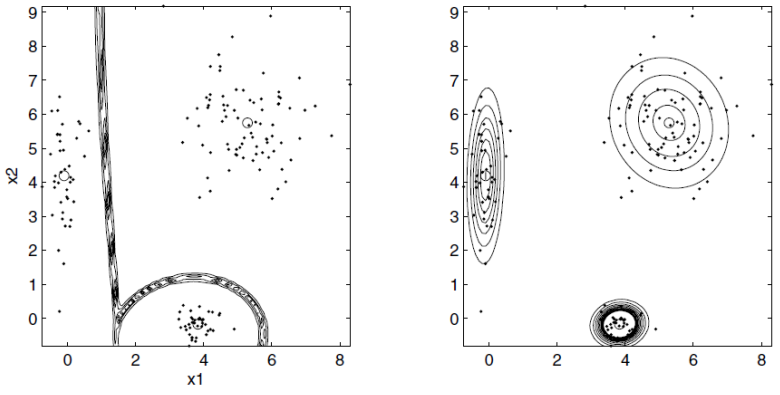
$$P_i = \frac{\sum_{j=1}^{n} u_{ij}^m}{\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^m} \tag{3.24}$$

$x_j$ verisi için normalleştirilmemiş sonsal olasılık (olabilirlik) $N_i$ normal dağılımı tarafından aşağıdaki gibi oluşturulur;

$$S_i = \frac{P_i}{(2\pi)^{\frac{p}{2}}\sqrt{\det(A^i)}} \exp\left(-\frac{1}{2}\left(x_j - v_i\right)^T A^{-1}\left(x_j - v_i\right)\right) \tag{3.25}$$

Beklenen değer ve kovaryans matrisi için istatistiksel tahmin edicilerin genelleştirilmesi, doğrudan küme merkezi $v_i$ ve kovaryans matrisi $A_i$ için hesaplama işlemlerine yol gösterir. $P_i$ üyelikler için önsel olasılıkları tahmin eder. Normalleştirilmiş sonsal olasılıklar, i. normal dağılım tarafından oluşturulan bir verinin olasılığını belirtir. Uzaklık fonksiyonu da sonsal olasılıklara orantılı olarak seçilir. Eğer uzaklık küçük ise bunun anlamı üyelikler için yüksek olasılık, uzaklık büyük ise düşük olasılıktır[60].

Eğer Gath-Geva algoritması, BCO ve GK algoritmalarının kullanıldığı tekniğine göre uygulanırsa yani J amaç fonksiyonu minimize edilmeye çalışılırsa, prototipler için elde edilecek eşitlikler analitik olarak çözülemez. Gath-Geva algoritması, BCO ve GK algoritmalarının teşhis edebileceği bütün örnekleri başarıyla kümeler. Ayrıca, düşük yoğunluklu ve geniş bir alana yayılmış kümeleri de kolayca teşhis eder. Uzaklık fonksiyonu içinde eksponansiyel fonksiyonun yer almasından dolayı, tüm uzaklıklar hemen hemen yakın ve uzak iki aralığa bölünür. Uzaklıklar, üyeliklerin 0 ile 1 den başka herhangi bir değer olması durumunda eksponansiyel fonksiyondan dolayı çok fazla artacaktır[60].




Şekil 3.4 Gath-Geva algoritmasının sentetik veri seti için kümeleme sonuçları [58]

GG Kümeleme Algoritması için Gerekli Adımlar:

X veri seti olmak üzere, 1<c<N sayıda küme sayısı seçilir. Bulanıklaştırma katsayısı m>1, işlem sonlandırma kriteri ε >0 ve U üyelik matrisi rastgele seçilerek başlatılır.

**Adım 1:** (3.22) formülü kullanılarak küme merkezlerinin hesaplanması

**Adım 2:** (3.23) formülü ile kovaryans matrisinin hesaplanması

**Adım 3:** (3.24) formülü ile önsel olasılıkların hesaplanması

**Adım 4:** (3.25) ile uzaklıkların hesaplanması

**Adım 5:** (3.15) ile üyelik değerlerinin güncellenmesi

**Adım 6:** Yeni üyelik değerleri ile eski üyelik değerlerini karşılaştır.$\left\| U^{(t)} - U^{(t-1)} \right\| < \epsilon$ ise iterasyonu durdur. Aksi takdirde Adım 1'ye geri dön[56].

### 3.1.4 Küme Geçerliliği (Cluster Validity)

Küme geçerliliği, elde edilen bulanık bölümlemenin verilerin tümü için uygun olup olmadığını inceler. Kümeleme algoritmaları belirli bir küme sayısı için en iyi çözüme ulaşmayı hedefler. Fakat bazı durumlarda en iyi çözümler bile anlamlı değildir. Ya küme sayısı hatalı olabilir ya da algoritmalarda oluşturulan küme şekilleri verinin yapısına uygun olmayabilir [58]. Küme sayısının belirlenmesi konusunda son yıllarda yoğun çalışmalar olmakla birlikte halen çok da güvenilir olmayan ve 1970'lerde kullanılan teknikler kullanılmaktadır. Anlamlı ve sağlıklı sonuçlara ulaşabilmek için en uygun küme sayısının belirlenmesi önemlidir [59]. En uygun küme sayısını belirlemek için farklı teknikler bulunmaktadır. Bu teknikler iki ana yaklaşım çerçevesinde toplanmaktadır. Birinci yaklaşımda kümeleme, mümkün olan en büyük küme sayısı ile gerçekleştirilir, önceden belirlenen kriterlere göre benzer kümeler birbirleriyle birleştirilerek küme sayısı azaltılır. Bu yaklaşım 'uygun kümeleri birleştirme' (compatible cluster merging) olarak adlandırılmaktadır. İkinci yaklaşımda ise kümeleme işlemi farklı küme sayılarıyla gerçekleştirilip elde edilen sonuçlar küme geçerlilik endeksleri kullanılarak değerlendirilir [58].

Literatürde, birçok bulanık küme geçerlilik yöntemi bulunmasına rağmen bu yöntemlerden hiçbiri mükemmel sonuçlar vermemektedir. Karmaşık yapılardan oluşan veri setlerinde, küme geçerlilik yöntemlerinin birbirleriyle çelişkili yanıtlar çıkarabildiği görülmektedir. Hangi yöntemin küme sayısını belirlemede daha başarılı olduğunu

ortaya koyan ölçüler de bulunmamaktadır. Bu sebeple başarılı bir sonuç elde etmek için tek bir yöntemle değerlendirme yapmak yarına birden fazla yöntemin kullanılması ve karşılaştırılması daha uygun olmaktadır [58]. Aşağıda literatürde en sık kullanılan küme geçerlilik yöntemleri yer almaktadır.

### 3.1.4.1 Bölünme Katsayısı (Partititon Coefficient) (PC)

1974 yılında Bezdek tarafından geliştirilen bu yöntem, kümeler arası örtüşmenin miktarını tespit etmektedir [58]. Bu yöntemde *c* küme sayısı olmak üzere küme geçerliliği 1/*c* ile 1 arasında değerler alır. 1/*c*, bölünme katsayısı'nın alabileceği minimum değer, 1 ise alabileceği maksimum değerdir. PC değerinin 1'e yakın olması en uygun küme sayısının seçilmiş olması anlamına gelmektedir. 1/c değerine yakın sonuçlar elde edilmiş kümelerde ise sonuçların başarısız olduğu söylenmektedir [59]. Bezdek tarafından geliştirilmiş olan bu yöntemin formülü (3.26)'de verilmektedir:

$$\mathrm{PC(c)} = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^2 \tag{3.26}$$

N: Toplam eleman sayısı

c: Küme sayısı

$u_{ij}$= j. nesnenin i. kümeye üyelik derecesi

Bu yöntemin zayıf yönleri:

- Nesnelere ait verilerin doğrudan değerlendirmeye alınmaması,
- Küme sayısı *c'nin* yüksek olduğu durumlarda PC değerinin düşmesidir.

### 3.1.4.2 Sınıflandırma Entropisi (Classification Entropy) (CE)

Bu yöntem de Bezdek tarafından 1974 yılında geliştirilmiştir. Sadece küme bölümlemesinin bulanıklığını ölçmekte olup PC ile benzeşmektedir [58].

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}\log(u_{ij}) \tag{3.27}$$

N: Toplam eleman sayısı

c: Küme sayısı

$u_{ij}$= j. nesnenin i. kümeye üyelik derecesi

En uygun küme sayısı, CE değerini minimize eden küme sayısıdır. CE değerinin 0'a yakın olması başarılı olduğu anlamına gelmektedir. Bu yöntemin de PC ile benzer zayıf yönleri bulunmakta, veri setini direk dikkate almamakta ve küme sayısı *c* arttıkça, CE endeksi değeri monoton bir şekilde artmaktadır[59].

### 3.1.4.3 Bölümlendirme İndeksi (Partition Index) (SC)

Bölümlendirme indeksi (SC) kümelerin sıklıklarının toplamının, kümelerin ayrıklıklarının toplamına olan oranını verir. SC yöntemi, eşit küme sayısına sahip farklı bölümlendirmeleri karşılaştırırken faydalıdır. SC değerinin küçük çıkması iyi bir bölümlendirme olduğunu gösterir [58].

$$SC(c) = \sum_{i=1}^{c} \frac{\sum_{k=1}^{N} u_{ik}^{m} \|x_k - v_i\|^2}{\sum_{k=1}^{N} u_{ik} \sum_{j=1}^{c} \|v_j - v_i\|^2} \quad (3.28)$$

N: Toplam eleman sayısı

c: Küme sayısı

$u_{ik}$: k. nesnenin i. kümeye üyelik derecesi

$v_i$: i. küme merkezi

$v_j$: j. küme merkezi

$x_k$: veri setindeki k. eleman

### 3.1.4.4 Ayırma İndeksi ( Seperation Index) (S)

Bölümlendirme yöntemi SC'nin aksine ayırma indeksi geçerlilik yöntemi, değerlendirme için küme merkezleri arası minimum uzaklığı kullanmaktadır [58]. Şu şekilde tanımlanır:

$$S(c) = \frac{\sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{2} \|x_k - v_i\|^2}{N min_{i,j} \|v_j - v_i\|^2} \quad (3.29)$$

N: Toplam eleman sayısı

c: Küme sayısı

$u_{ik}$: k. nesnenin i. kümeye üyelik derecesi

$v_i$: i. küme merkezi

$v_j$: j. küme merkezi

$x_k$: veri setindeki k. eleman

$min_{i,j}\|v_j - v_i\|^2$ küme merkezleri arasındaki minimum uzaklık

Kümeler birbirinden ne kadar ayrıysa, uzaklık o kadar büyük çıkacak ve S İndeksi de aynı şekilde azalacaktır. S değerini minimize eden küme sayısı uygun küme sayısıdır. Bölümlendirme Endeksi (SC) ve Ayırma Endeksi (S), aynı küme sayısı ile farklı kümeleme tekniklerini karşılaştırma durumunda daha kullanışlıdır [59].

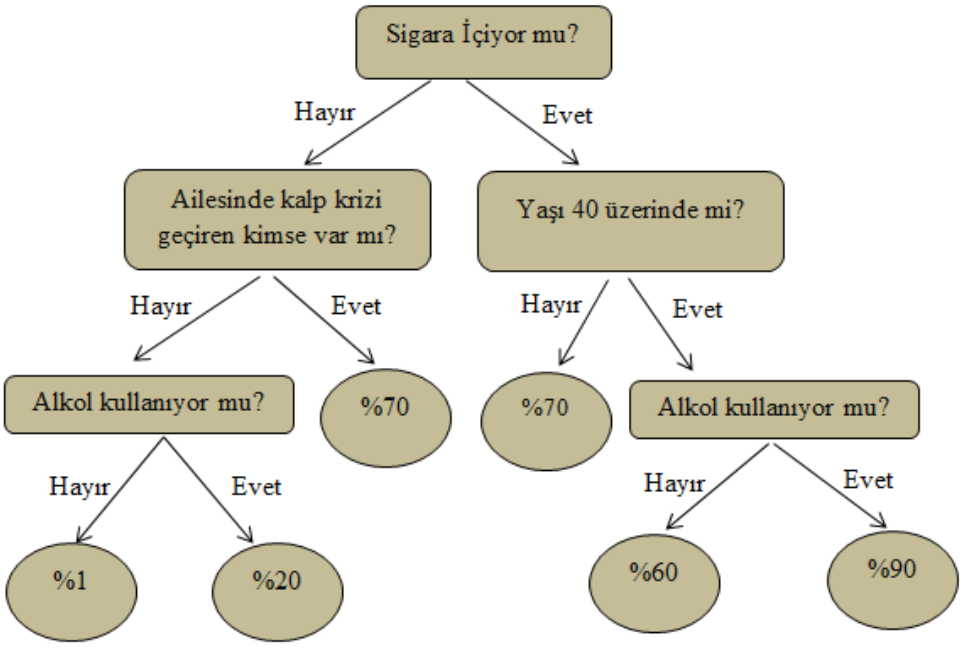
### 3.2 Karar Ağacı (Decision Trees) Sınıflandırma Yöntemi

Tahmin edici özelliklere sahip olan karar ağaçları, veri madenciliğinde yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre olabilmeleri ve güvenilirliklerinin iyi olması sebepleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahiptir [49].

Karar ağaçları, karar düğümlerinden, dallardan ve yapraklardan oluşurlar. Karar düğümü, gerçekleştirilecek testi işaret eder. Bu testin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olur.

Karar ağacı işlemi kök düğümünden başlar ve yukarıdan aşağıya doğru yaprağa ulaşana dek ardışık düğümleri takip ederek gerçekleşir [50]. Karar ağacı algoritmalarından en sık kullanılanları ID3, C4.5, C5.0, M5P ve J48 algoritmalarıdır. İyi bir karar ağacı algoritması ayrık ve sürekli verilerin her ikisi üzerinde de çalışabilmelidir. Sürekli veriler üzerinde oluşturulan modeller regresyon ağacı olarak adlandırılır [51].

Karar ağacı algoritmalarında her zaman sorgulanan kategorilerde en uygun dala ayrılmaya çalışılır. Bu ayrılma işleminin hesaplanmasında da farklı hesaplama türleri kullanılır. Sınıflama ve regresyon(Classification and Regression) ağaçları çeşitlilik indeksini(index of diversity) kullanırken, ID3 algoritması entropi değerine göre düğümlerin ayrılmasını hesaplar[52]. Örnek bir karar ağacı yapısı Şekil 3.3'de gösterilmiştir. Şekilde kalp krizi geçiren kişilerin çeşitli özelliklerine bakılarak veriler üzerinden öğrenme işlemi gerçekleştirilmektedir. Başka bir ifade ile, iç düğümlerde girdilere göre ağacın dallanması söz konusudur. Yapraklarda ise bu dallanmalar sonucunda elde edilen değerler gösterilmiştir [53].




Şekil 3.5 Örnek bir karar ağacı [53]

Karar ağaçlarının, diğer sınıflama tekniklerine göre üstün yanı, bilgiden elde edilen çıkarımın anlaşılır bir şekilde yazılabilmesidir. Ayrıca, karar ağaçlarının kuralları kesinlik gösterirken, diğerleri yaklaşımsal sonuçlar üretmektedirler [54].

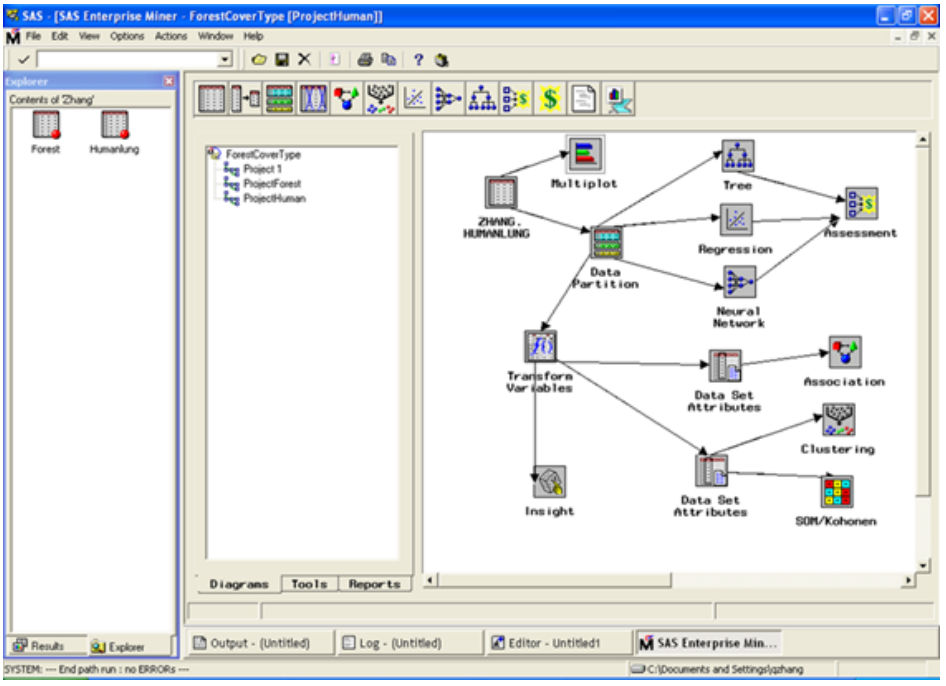
# BÖLÜM 4

# UYGULAMA

## 4.1 Uygulamada Kullanılan Veri Madenciliği Programları

### 4.1.1 SAS Enterprise Miner

SAS firmasının veri madenciliği aracıdır. Enterprise Miner karar ağaçları, yapay sinir ağları, regresyon çözümlemesi, kümeleme, zaman serileri, birliktelik kurallarının bulunması gibi veri madenciliği sorgularını ele alabilmektedir. Grafiksel ara yüzü sayesinde kullanım kolaylığı sağlar. Diğer veri madenciliği programlarına oranla istatiksel ve regresyon açısından fazla sayıda araca sahiptir. Algoritma derinliği ve görsel arabirimi güçlü olduğu yönlerindendir [48]. Şekil 4.1'de Enterprise Miner yazılımının kullanıcı arayüzü yer almaktadır.

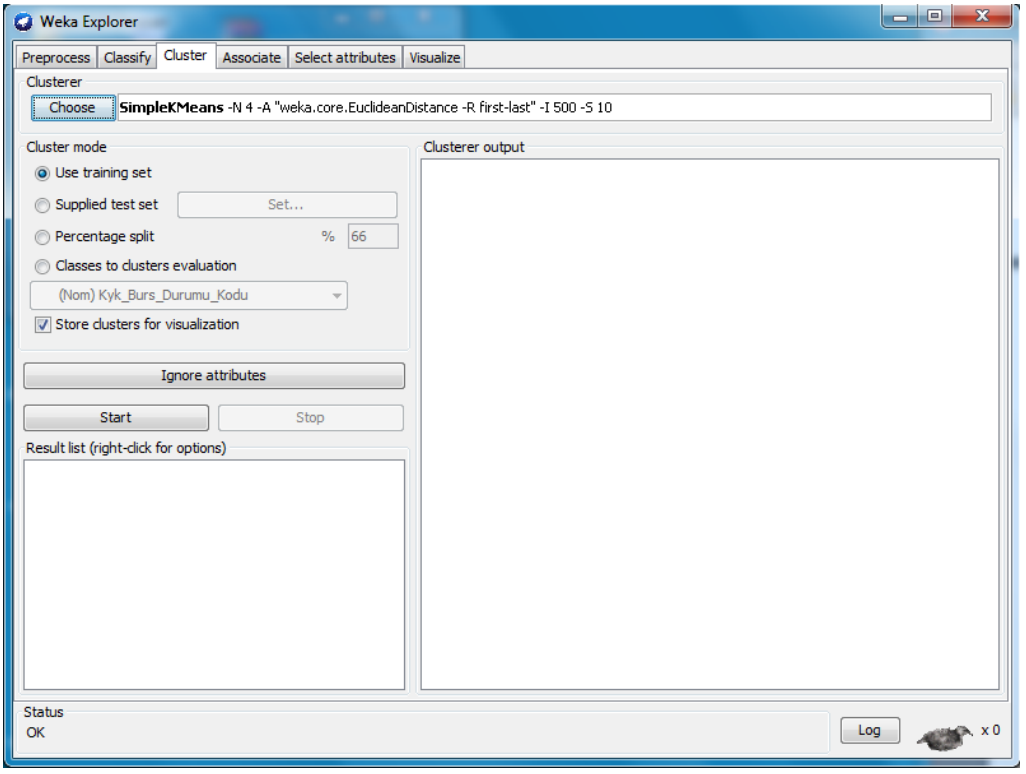



Şekil 4.1 SAS Enterprise Miner kullanıcı arayüzü

### 4.1.2 WEKA

Yeni Zelanda Waikato Üniversitesinde, Java programlama dili ile geliştirilmiş açık kaynak kodlu bir veri madenciliği yazılımıdır. Java ile geliştirilmiş olması, Linux, Unix, Windows ve Macintosh gibi başlıca işletim sistemlerinde kullanılabilir olmasını sağlamıştır.

Weka veri madenciliği ve makine öğrenmesi için birçok farklı algoritmayı içinde barındırmaktadır. Kullanımı kolay bir yazılımdır. *.arff* formatını ve *.csv* formatını kullanabilmektedir. Bir SQL veri tabanından veya bir URL'den de veri okuyabilmektedir. Kullanıcı ara yüzü Şekil 4.2'de gösterilmektedir.




Şekil 4.2 Weka kullanıcı arayüzü

### 4.1.3 MATLAB

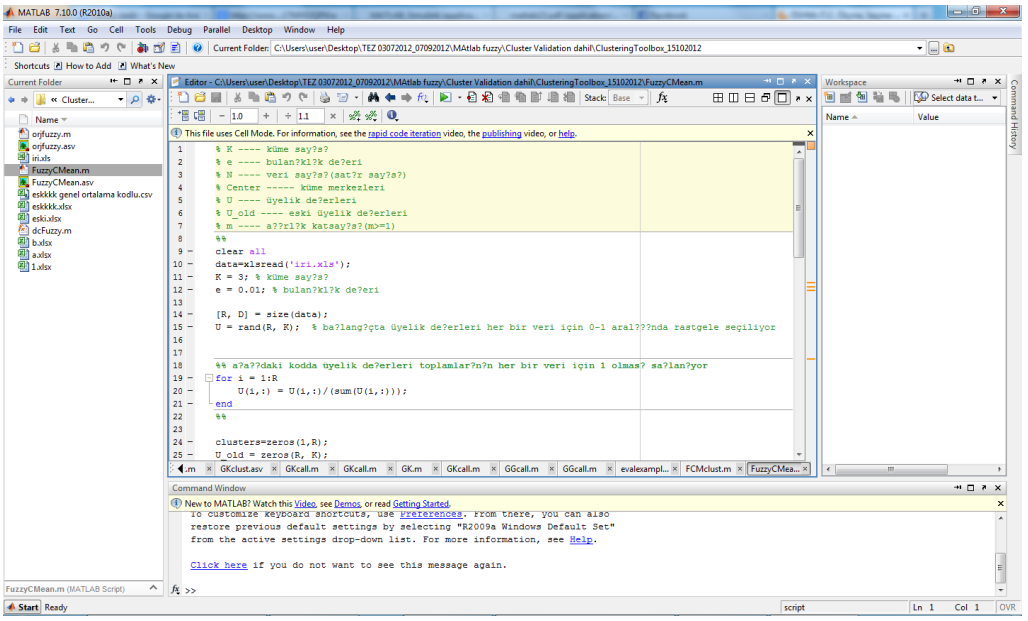
MATLAB temel de bilimsel hesaplamalar yapmak için yazılmış yüksek performanslı bir yazılımdır. 1970'lerin sonunda Cleve Moler tarafından yazılan Matlab programının tipik kullanım alanları:

- Hesaplama işlemleri.
- Modelleme
- İki ve üç boyutlu grafiklerinin çizimi

- Veri analizi

şeklinde özetlenebilir.

MATLAB, matematik, istatistik, yapay sinir ağları, bulanık mantık, işaret ve görüntü işleme, kontrol tasarımları, yöneylem çalışmaları, tıbbi araştırmalar, finans ve uzay araştırmaları gibi birçok alanda kullanılmaktadır. Şekil 4.3'de MATLAB kullanıcı ara yüzünü görmek mümkündür.



Şekil 4.3 MATLAB kullanıcı arayüzü

## 4.2 Veri Setinin Tanımlanması

Çalışma da Namık Kemal Üniversitesi Çorlu Mühendislik Fakültesinde 2011 yılı sonu itibariyle halen kayıtlı olan lisans öğrencilerinin verileri kullanılmıştır. Veri seti toplamda 1934 mühendislik öğrencisinin kaydını içermektedir.

Veriler ilk elde edildiğinde Excel formatında tutulmaktaydı. WEKA'da analizlerin gerçekleştirilebilmesi için veriler virgülle ayrılmış CSV formatına dönüştürülmüştür. Ayrıca WEKA veri madenciliği aracının yanında SAS Enterprise Miner veri madenciliği platformu da analizlerde kullanılmıştır.

Veri setinde yer alan 25 farklı özellik ve bu özelliklerin açıklamaları Çizelge 4.1'de yer almaktadır.

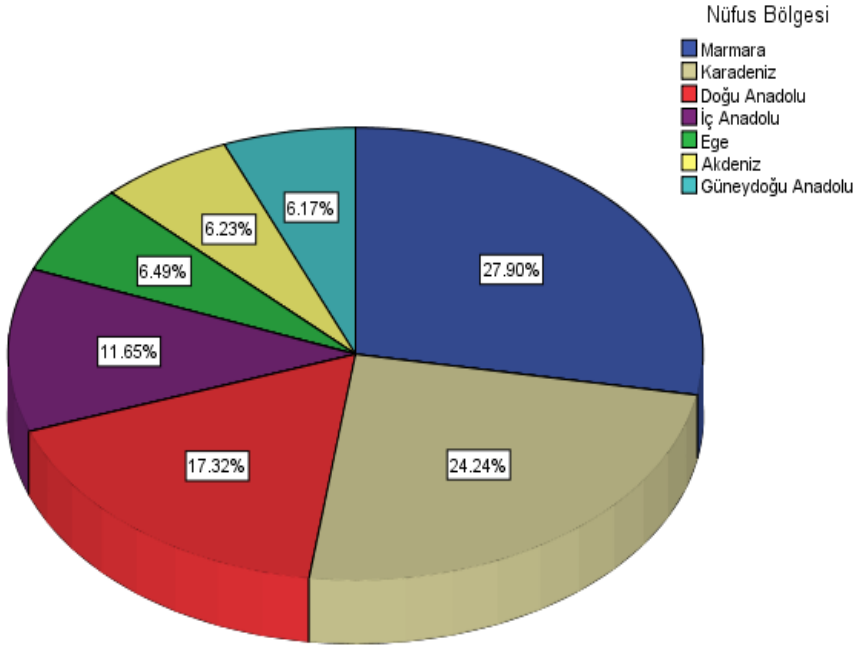
Çizelge 4.1 Veri Seti Özellikleri

| **Özellik** | **Açıklaması** | **Değeri** |
|---|---|---|
| Öğrenci No | Öğrencinin Numarası | Nümerik |
| Bölüm | Öğrencinin Bölüm bilgisi | Kategorik (Çevre, İnşaat, Bilgisayar vb.) |
| Yerleştiği Kontenjan Tipi | Üniversiteyi kazanma şekli (1. yerleştirme, ek yerleştirme, yatay geçiş vb.) | Kategorik (1.Yerleştirme, DGS, Yatay Geçiş vb.) |
| Sınıfı | Öğrencinin öğrenimine devam ettiği sınıf bilgisi | Nümerik |
| İntibak | İntibak öğrencisi olup olmadığı bilgisi | Kategorik(Evet, Hayır) |
| Kayıt Yılı | Öğrencinin üniversiteye kaydolduğu yıl bilgisi | Nümerik |
| Eğitim Yılı | Öğrencinin üniversitede kaçıncı yılında olduğu bilgisi | Nümerik |
| Öğretim Türü | Örgün öğretim-İkinci öğretim bilgisi | Kategorik(Normal, Örgün) |
| Disiplin Cezası | Öğrencinin disiplin cezası bilgisi | Kategorik(Evet, Hayır) |
| Cinsiyet | Öğrencinin cinsiyeti | Kategorik(Kadın, Erkek) |
| Yaşı | Öğrencinin yaşı | Nümerik |
| Medeni Durum | Öğrencinin medeni durumu | Kategorik(Evli, Bekâr) |
| Türk Asıllı | Öğrencinin milliyeti | Kategorik(Evet, Hayır) |
| Nüfus İli | Öğrencinin nüf. kayıtlı olduğu il | Kategorik |
| Adres İli | Öğrencinin ikamet ettiği il | Kategorik |
| Okul Türü | Öğrencinin ortaöğretim kurum tipi | Kategorik(Resmi lise, teknik lise, Anadolu lisesi vb.) |
| Orta Öğrenim Mezuniyet Yılı | Orta öğrenimden mezuniyet yılı | Nümerik |
| Yerleştirme Puanı | Üniversiteye yerleştirme puanı | Nümerik |
| Yerleştirme Sırası | Üniversiteye yerleştirme sırası | Nümerik |
| Tercih Sırası | Öğrencinin kazandığı bölümü tercih sırası | Nümerik |
| Genel Ortalama | Öğrencinin 2011 bahar dönemi itibariyle üniversitedeki genel not ortalaması | Nümerik |
| Katkı Kredisi Durumu | Katkı kredisi bilgisi | Kategorik(Evet, Hayır) |
| Öğrenim Kredisi Durumu | Öğrenim kredisi bilgisi | Kategorik(Evet, Hayır) |
| Kyk Burs Durumu | Burs durum bilgisi | Kategorik(Evet, Hayır) |
| Yüzde On | Yüzde ona girip girmediği bilgisi | Kategorik(Evet, Hayır) |

### 4.2.1 Veri Ön İşleme

Veriler elde edildikten sonra çalışmanın başında anlatılan ön işleme teknikleri uygulanmış ve ön işleme aşamasında çok fazla eksik veri içeren, çalışmanın sonucunu olumsuz etkileyecek kayıtlar silinmiştir. Sonuç da 1588 öğrenci verisi elde edilmiştir. Ayrıca özelliklerden bazıları yeteri kadar veri olmayışı sebebiyle veri setinden çıkartılmıştır. Örneğin, 1934 öğrenci içerisinden yaklaşık 20 öğrenci intibak ile yerleşmiş olduğundan, intibak özelliği bize yeterli doğruluk oranını veremeyeceğinden veri setinden çıkartılmıştır. Bunun yanında adres ili ve orta öğrenim mezuniyet yılı da veri setinden çıkartılan diğer özelliklerdir.

Veri setinde bazı özellikler kategorize edilmiştir. Bunlardan bir tanesi öğrencinin nüfusa kayıtlı olduğu il bilgisidir. 82 ilin bilgileri ayrı ayrı bir anlam ifade etmeyeceği düşüncesiyle, iller bölgelere kategorize edilmiştir. Ve özellik nüfus bölgesi olmak üzere 7 farklı değer almaktadır. Şekil 4.4'de öğrencilerin bölgelere göre dağılımı gözükmektedir.




Şekil 4.4 Öğrencilerin bölgelere göre dağılımı

Öğrenci yüzdelerine göre bölgeler şu şekildedir;

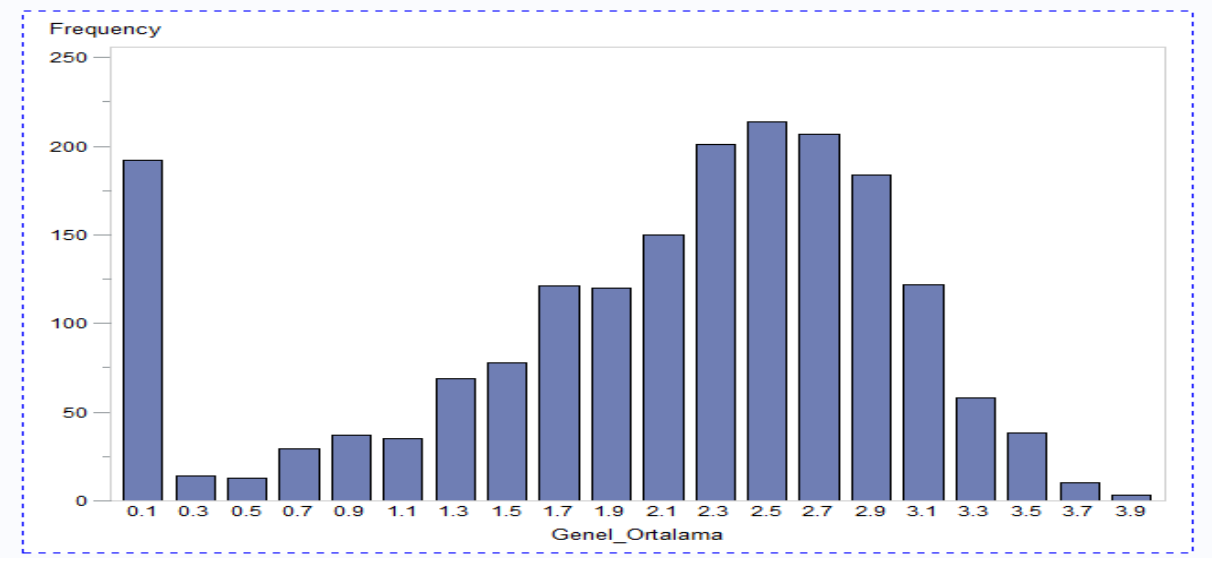
- Marmara Bölgesi %27,90
- Karadeniz Bölgesi %24,24

- Doğu Anadolu Bölgesi %17,32
- İç Anadolu Bölgesi %11,65
- Ege Bölgesi %6,49
- Akdeniz Bölgesi %6,23
- Güneydoğu Anadolu Bölgesi %6,17

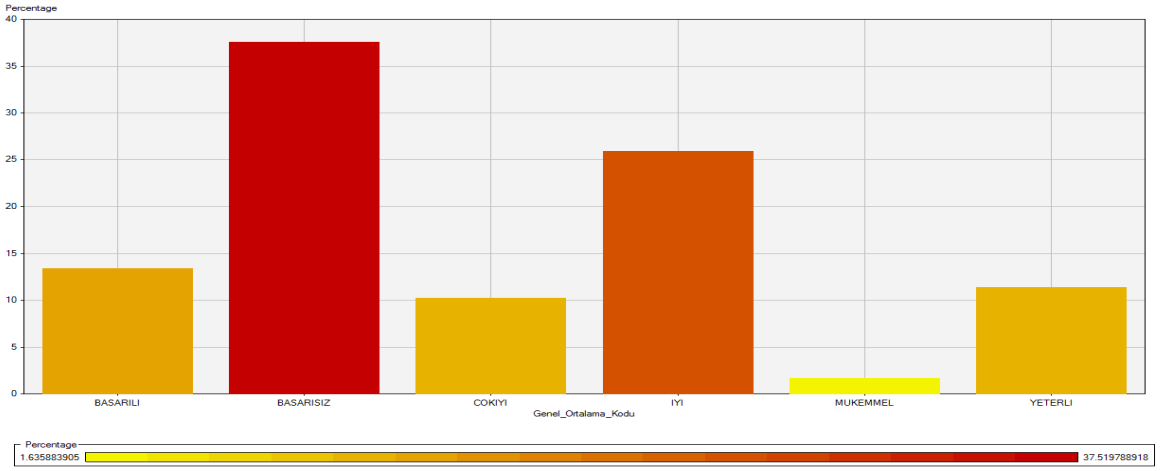
Nüfus bölgelerinin yanı sıra öğrencilerin ağırlıklı not ortalamaları bilgisini tutan özellik de kategorize edilmiştir. Öğrencilerin ağırlıklı not ortalamaları veri setinde 0.0 ile 4.0 aralığında ki sayısal değerlerden oluşmaktadır. Ön işleme aşamasının ardından veri setindeki hedef değişken (target class) olan Genel Not Ortalamasının alabileceği değerler şu şekilde gruplanmıştır;

- “Başarısız” (0.0-2.0),
- “Yeterli” (2-2.25),
- “Başarılı” (2.25-2.5),
- “İyi” (2.5-3.0)
- “Çok İyi” (3-3.5).
- “Mükemmel” (3.5-4.0).

Öğrencilerin ağırlıklı not ortalamalarının dönüşüme uğramadan önce ve dönüşüme uğradıktan sonraki hali Şekil 4.5 ve 4.6’dan görülebilmektedir.




Şekil 4.5 Notların dönüştürülmeden önceki dağılımı



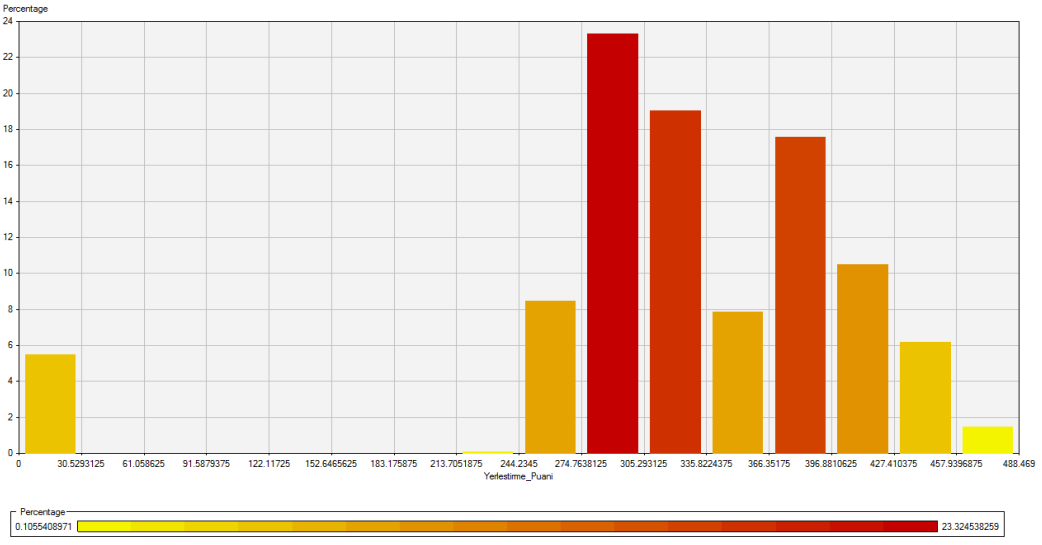

Şekil 4.6 Notların dönüştürüldükten sonraki dağılımı

## 4.3 Veri Setinin Analiz Edilmesi

Çalışmanın bu kısmında SAS Enterprise Miner yazılımı kullanılarak elde edilmiş çeşitli sonuçlar ve grafikler yer alacaktır. Enterprise Miner'ın Multiplot özelliği ile ikili karşılaştırma grafikleri, Distribution Explorer özelliğiyle ise 3'lü karşılaştırma grafikleri elde edilmiştir.

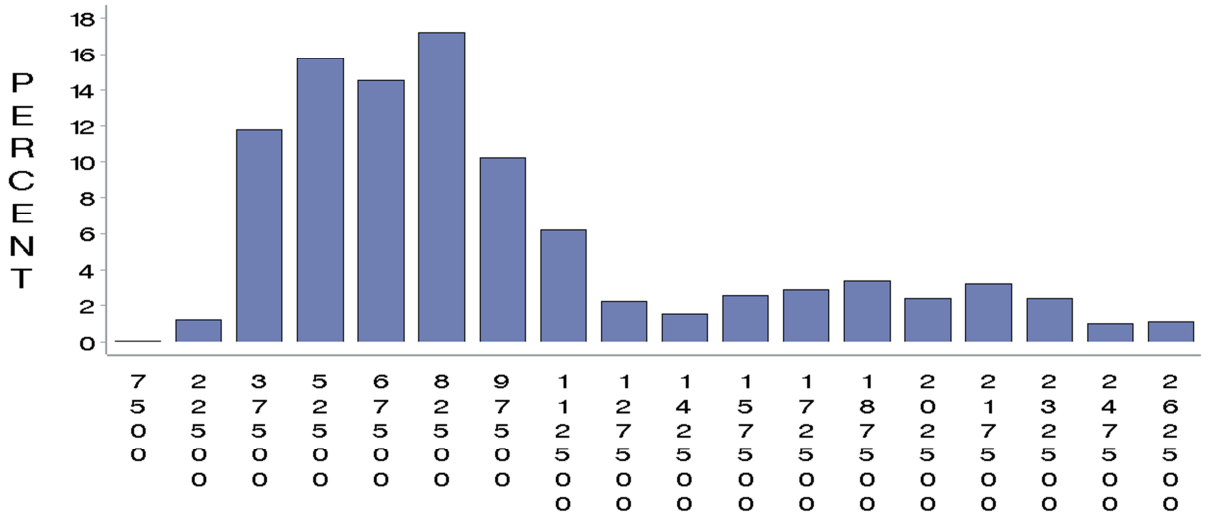
### 4.3.1 Yerleştirme Puanlarına ve Sıralarına Göre Öğrencilerin Dağılımı

Şekil 4.7'de ön işleme aşamasının ardından öğrencilerin puanlarına göre nasıl dağılım gösterdikleri görülmektedir. Grafiğe bakıldığında 290 ile 335 puan aralığında en fazla öğrencinin yer aldığı görülmektedir. Ayrıca çok yüksek puana sahip öğrencilerin bu fakülteyi pek tercih etmedikleri de şekilden görülebilmektedir.




Şekil 4.7 Yerleştirme puanlarına göre öğrencilerin dağılımı

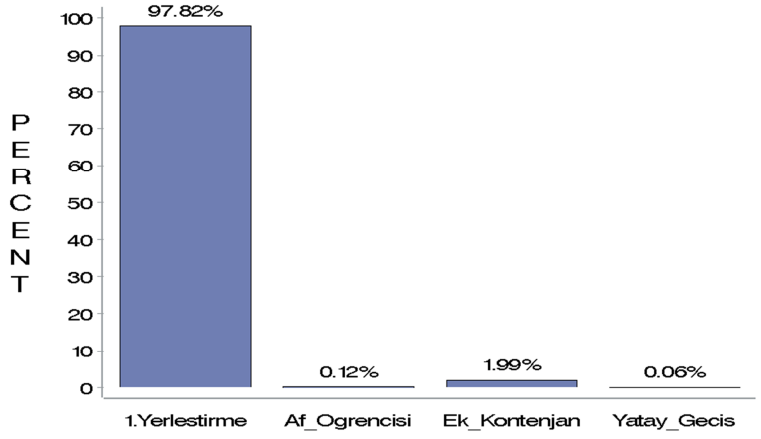
Şekil 4.8’de görüldüğü üzere fakülteyi kazanan öğrencilerin yerleşme sıralarının büyük çoğunluğunun 40 ile 100 bin arasında olduğu gözlemlenmektedir.




Şekil 4.8 Yerleştirme sıralarına göre öğrenci dağılımları

### 4.3.2 Yerleştikleri Kontenjan Tipine Göre Öğrencilerin Dağılımı

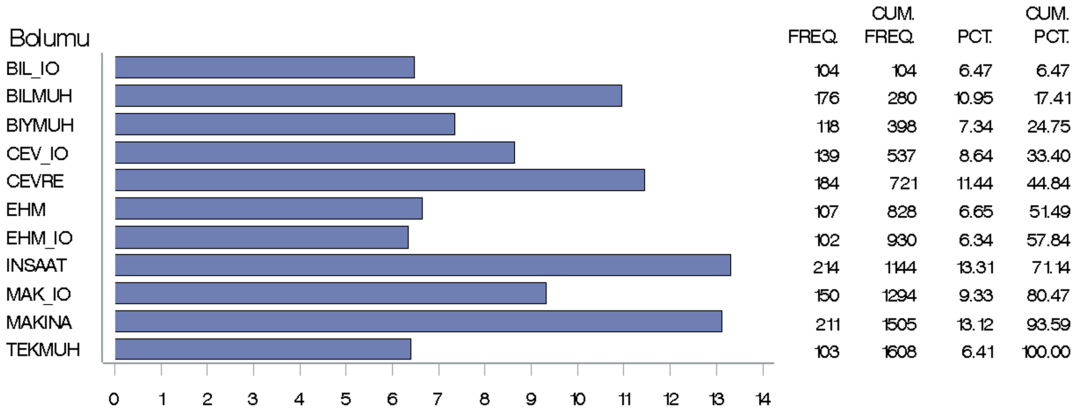
Şekil 4.9’da görüldüğü gibi öğrencilerin %97’si 1.yerleştirme neticesinde fakülteyi kazanmışlardır. %2’lik bir bölüm ise ek kontenjan ile yerleştirme ile kayıt yaptırmıştır. Af ile gelen öğrencilerle yatay geçiş ile kazanan öğrencilerin oranının çok düşük olduğu da görülebilmektedir.




Şekil 4.9 Öğrencilerin yerleştikleri kontenjan türleri

### 4.3.3 Bölümlere Göre Öğrencilerin Dağılımı

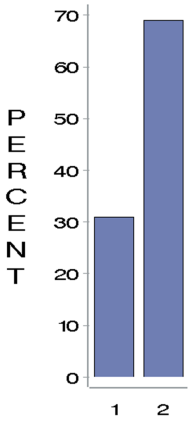
Şekil 4.10'da bölümlere göre öğrenci yüzdeleri görülmektedir. %13 ile inşaat mühendisliği bölümü en yüksek öğrenci sayısına sahiptir. İnşaat mühendisliğini yine yakın bir yüzde ile makine mühendisliği izlemektedir. İnşaat ve makine mühendisliğinin sayılarının diğer bölümlere göre yüksek olmasının sebebi en eski kurulmuş bölüm olmaları ve bölüm kontenjanlarının yüksek olmasıdır.




Şekil 4.10 Bölümlere göre öğrenci dağılımları

### 4.3.4 Cinsiyete Göre Öğrencilerin Dağılımı

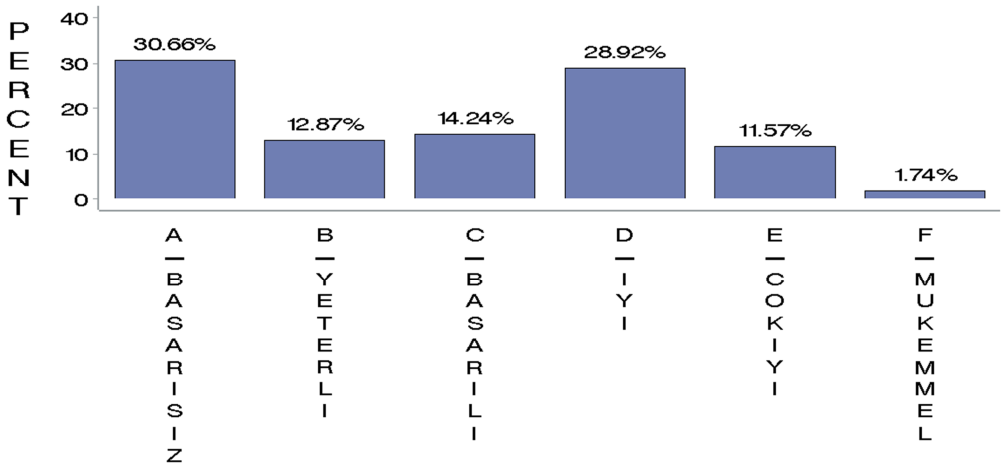
Şekil 4.11'de erkek öğrenci sayısının kız öğrenci sayısının iki katından daha fazla olduğu görülmektedir. Bu şekil üzerinden kız öğrencilerin mühendislik fakültelerini erkek öğrencilere göre daha az tercih ettikleri sonucuna ulaşabiliriz.




Şekil 4.11 Cinsiyete göre öğrenci dağılımları

### 4.3.5 Genel Not Ortalamasına Göre Öğrencilerin Dağılımı

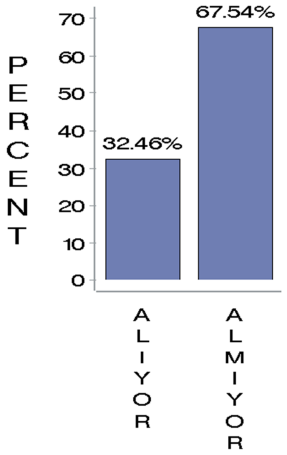
Şekil 4.12'ye bakıldığında fakülte genelinde öğrenci profilinin başarısız ve iyi öğrencilerden oluştuğu görülmektedir. Aradaki durumların oranlarının düşük olduğu ve ayrıca çok başarılı öğrenci sayısının da düşük olduğu görülmektedir.




Şekil 4.12 Genel not ortalamasına göre öğrenci dağılımları

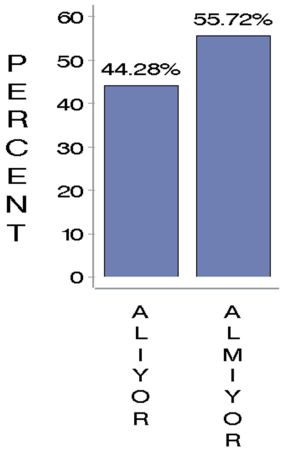
### 4.3.6 Kredi ve Burs Durumlarına Göre Öğrencilerin Dağılımı

Katkı ve öğrenim kredileri devlet tarafından öğrencilere verilen borçlardır. Öğrenci okulunu bitirdikten sonra bu kredileri devlete geri ödemektedir. Burs ise maddi durumu zayıf olan ve başarılı öğrencilere verilen karşılıksız yardımdır. Bu açıklamalar doğrultusunda aşağıdaki şekillere bakıldığında öğrencilerin %32'sinin katkı kredisi aldığı ve %67'sinin ise almadığı görülmektedir.



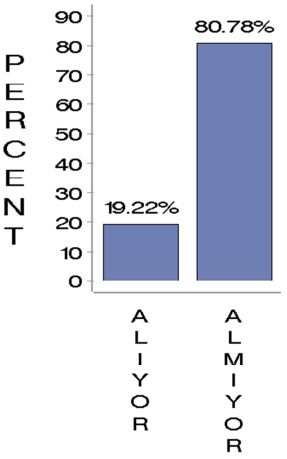

Şekil 4.13 Katkı kredisi durumuna göre öğrenci dağılımları

Şekil 4.14'e göre öğrencilerin %44'ü öğrenim kredisi alırken %55'i almamaktadır.




Şekil 4.14 Öğrenim kredisi durumuna göre öğrenci dağılımları

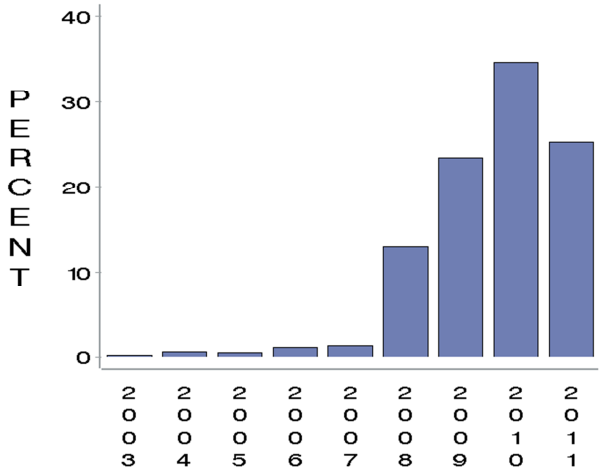
Öğrencilerin %19'u Kredi Yurtlar Kurumundan burs alırken %80'i almamaktadır.




Şekil 4.15 Burs durumuna göre öğrenci dağılımları

### 4.3.7 Kayıt Yılına Göre Öğrencilerin Dağılımı

Şekil 4.16'da yıllara göre öğrenci sayılarının yüzde olarak durumları görülmektedir. Öğrenci sayıları 2008 yılından itibaren artışa geçmiştir. Bu artışın en önemli nedeni başlangıçta iki bölümden oluşan fakültenin bölüm sayısını 7'ye çıkarmış olmasıdır. 2010 yılındaki yüksek artışın sebebi ise YÖK'ün üniversitelerin almış olduğu öğrenci kontenjanlarını bu yıl artırmış olmasından dolayıdır.



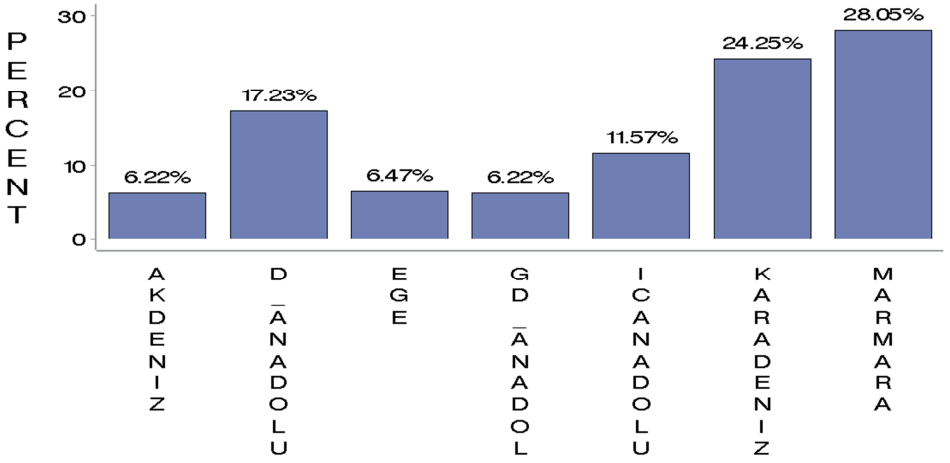

Şekil 4.16 Kayıt yılına göre öğrenci dağılımları

### 4.3.8 Nüfusa Kayıtlı Oldukları Bölgelere Göre Öğrencilerin Dağılımı

Şekil 4.17'de kazanan öğrencilerin geldikleri coğrafi bölgeler yer almaktadır. Bu grafiği analiz ederken Türkiye'nin bölgelere göre nüfus yoğunluğunu bilmekte fayda vardır. Türkiye'nin nüfusu en çok olan bölgeden en aza göre sıralanışı şu şekildedir;

1. Marmara= 22.387.173
2. İç Anadolu= 11.965.642
3. Ege= 9.687.692
4. Akdeniz= 9.620.240
5. Karadeniz= 7.508.293
6. Güney Doğu Anadolu= 6.923.256
7. Doğu Anadolu= 6.631.973

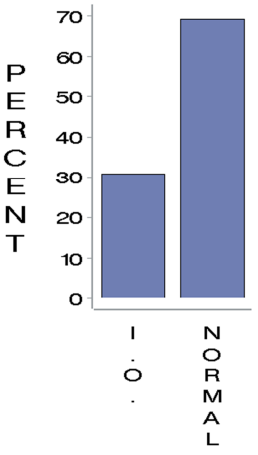
Bölgeye yakın olması ve nüfus yoğunluğunun en fazla olması sebebi ile en yüksek oran Marmara bölgesindedir. Marmara bölgesini %24'lük oranla Karadeniz bölgesi izlemektedir.




Şekil 4.17 Nüfusa kayıtlı olduğu bölgeye göre öğrenci dağılımları

### 4.3.9 Öğretim Türüne Göre Öğrencilerin Dağılımı

Şekil 4.18'de görüldüğü üzere öğrencilerin%70'i örgün öğretim, %30'u ise ikinci öğretimdir.



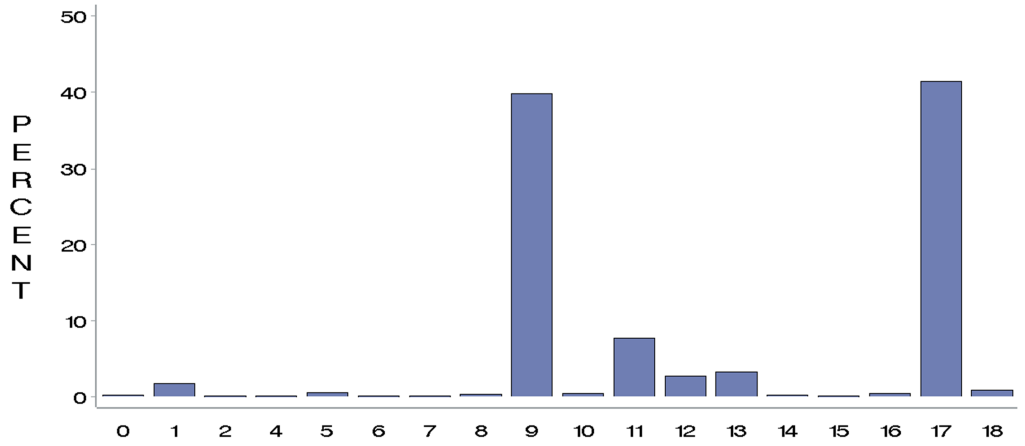

Şekil 4.18 Öğretim türüne göre öğrenci dağılımları

### 4.3.10 Okul Türlerine Göre Öğrencilerin Dağılımı

Çizelge 4.2'de okul türlerinin kodları görülmektedir. Bu kodlar ışığında öğrencilerin orta öğretim türlerine göre dağılımını gösteren Şekil 4.19'a bakıldığında ilk sırayı Anadolu liselerinin aldığı, sonrasında ise resmi ve gündüz öğretim yapan liselerin geldiği görülmektedir. Fen lisesi öğrencilerinin fakülteyi tercih oranlarının yok denecek kadar az olduğu da şekilden görülmektedir.

Çizelge 4.2 Okul Türü Kodları

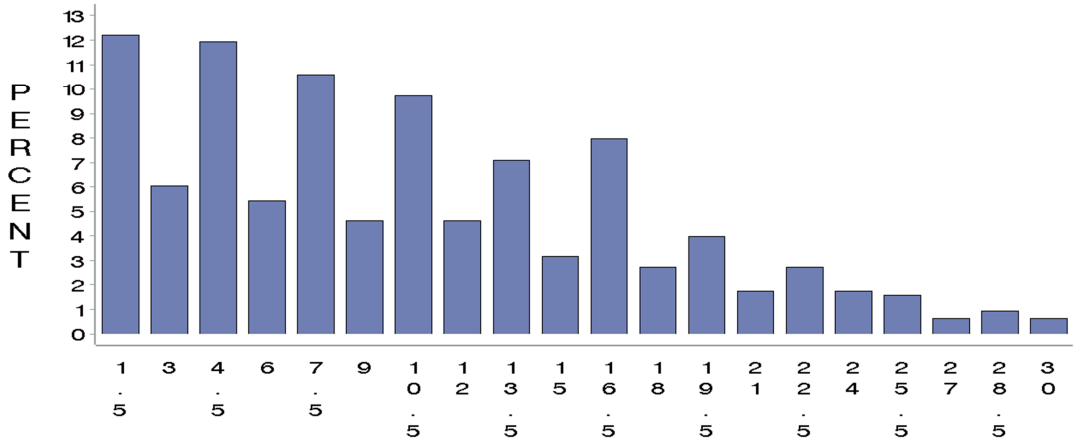
| No | Okul Türü Kodu | | |
|---|---|---|---|
| 1 | Lise Programı | 10 | Özel Lise |
| 2 | Endüstri Meslek Lisesi | 11 | Anadolu Güzel Sanatlar Lisesi |
| 3 | Anadolu Ticaret Meslek Lisesi | 12 | Yabancı Dille Öğretim Yapan Özel Lise |
| 4 | Özel Akşam Lisesi | 13 | Anadolu Öğretmen Lisesi |
| 5 | Anadolu Teknik Lisesi | 14 | Askeri Lise |
| 6 | Anadolu Kız Meslek Lisesi | 15 | Lise Programı/Yabancı Dil Ağırlıklı |
| 7 | Anadolu İmam hatip Lisesi | 16 | Özel Fen Lisesi |
| 8 | Anadolu Meslek Lisesi | 17 | Anadolu Lisesi(Yab.Dil.Öğretim |
| 9 | Lise (Resmi Ve Gündüz | 18 | Fen Lisesi |




Şekil 4.19 Okul türüne göre öğrenci dağılımları

### 4.3.11 Tercih Sıralarına Göre Öğrencilerin Dağılımı

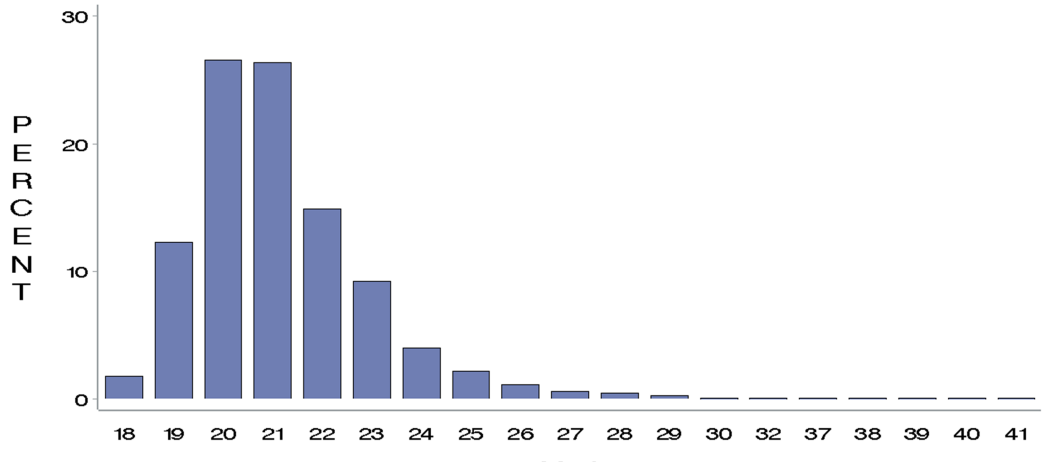
Öğrencilerin fakülteyi tercih etme durumlarını gösteren aşağıdaki şekle bakıldığında genellikle ilk on tercihleri arasında fakülteye yer verdikleri görülmektedir.




Şekil 4.20 Tercih sırasına göre öğrenci dağılımları

### 4.3.12 Yaşa Göre Öğrencilerin Dağılımı

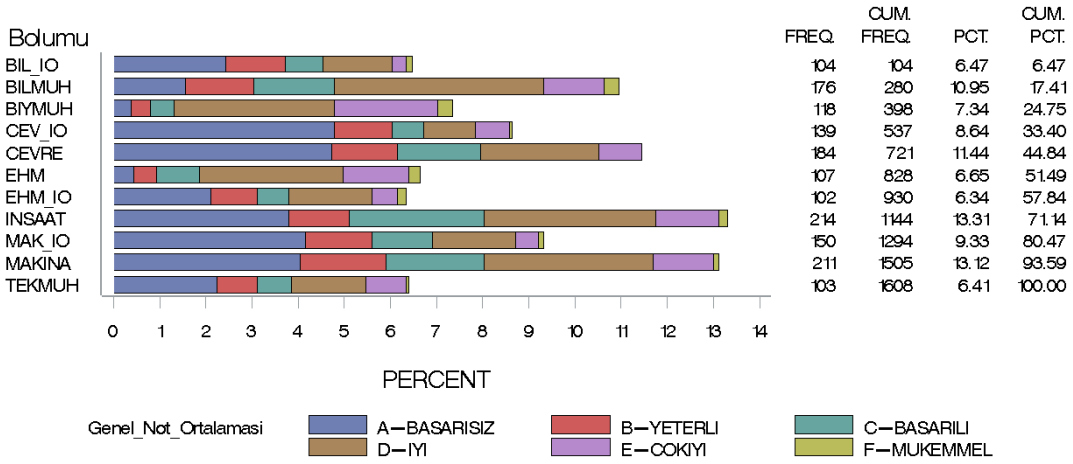
Yaş yüzdelerini gösteren aşağıdaki şekilde beklenildiği üzere liseden mezun olan öğrencilerin yaşlarıyla paralellik gösterecek şekilde 19 ve 22 yaş aralığında yoğunlaşmıştır.




Şekil 4.21 Yaşa göre öğrenci dağılımları

### 4.3.13 Bölüm - Genel Not Ortalamasına Göre Öğrenci Dağılımı

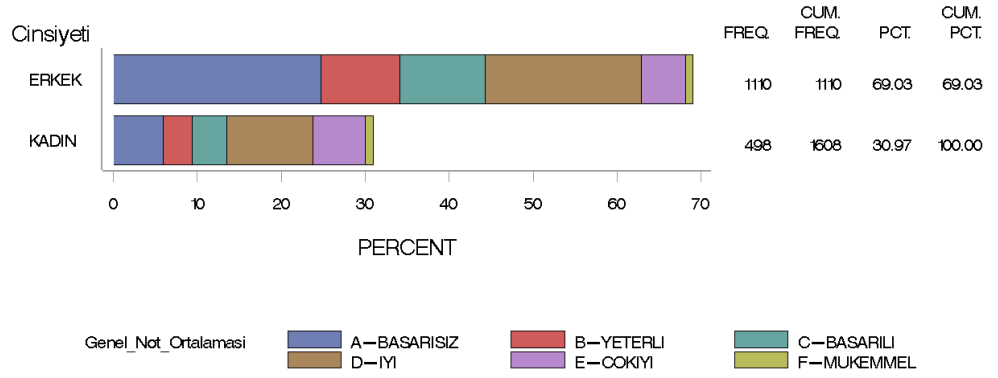
Şekil 4.22'de bölümlere göre öğrencilerin başarı durumları gözükmektedir. Başarı durumlarının en iyi olduğu iki bölüm %90'lık başarı oranı ile biyomedikal mühendisliği ve %84'lük başarı oranları ile elektronik ve haberleşme mühendisliği bölümleridir. İkinci öğretim öğrencilerinin örgün öğretim öğrencilerine göre daha başarısız oldukları gözlemlenmektedir. Bu doğrultuda en başarısız bölümler %30 başarı oranı ile çevre mühendisliği ikinci öğretim ve %40 başarı oranı ile makine mühendisliği ikinci öğretim öğrencileridir. Bu iki bölümü de %42 başarı oranı ile bilgisayar mühendisliği ikinci öğretim öğrencileri izlemektedir.




Şekil 4.22 Bölüm- Genel not ortalamasına göre öğrenci dağılımları

### 4.3.14 Cinsiyet - Genel Not Ortalamasına Göre Öğrenci Dağılımı

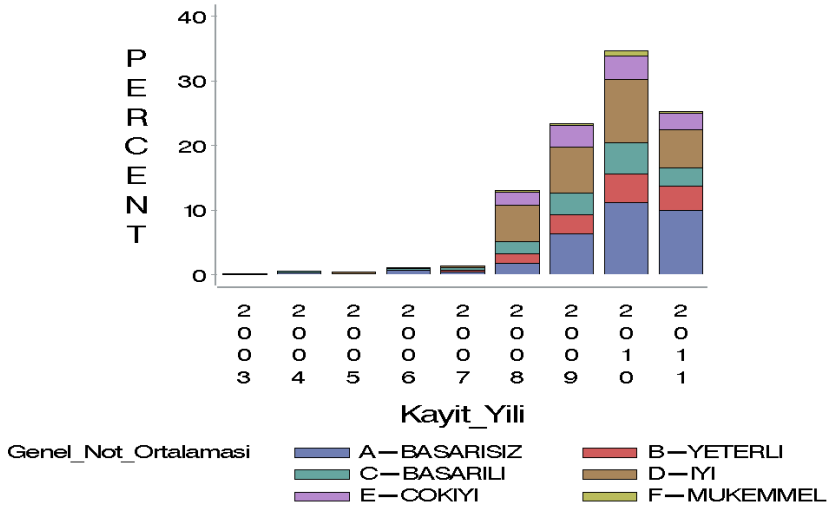
Cinsiyete göre genel not ortalamalarının yer aldığı Şekil 4.23'de bayanların başarı oranlarının %67, erkeklerin başarı oranının ise %50 olduğu görülmektedir. Bu doğrultuda erkeklerin bayanlara göre daha başarısız olduğu görülmektedir.




Şekil 4.23 Cinsiyet-Genel not ortalamasına göre öğrenci dağılımları

### 4.3.15 Kayıt Yılı - Genel Not Ortalamasına Göre Öğrenci Dağılımı

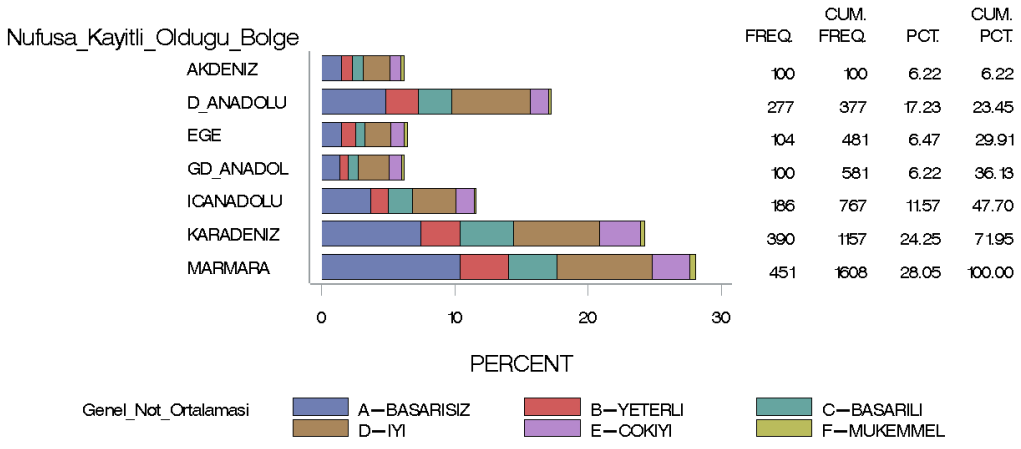
Şekil 4.24'de kayıt yıllarına göre GNO'ların dağılımları gözükmektedir. Şekil'de dikkatleri çeken hususlardan birisi 2010 yılında öğrenci sayılarında ki yüksek artıştır. Bu artışın sebebi de ÖSYM'nin alınan öğrenci kontenjanlarını artırmasından dolayıdır. 2010 yılında başarılı öğrencilerin artışına paralel olarak başarısız öğrencilerin sayısında da artış gözükmektedir.




Şekil 4.24 Kayıt yılı-Genel not ortalamasına göre öğrenci dağılımları

### 4.3.16 Nüfusa Kayıtlı Bölge - Genel Not Ortalamasına Göre Öğrenci Dağılımı

Şekil 4.25'e bakıldığında Karadeniz ve Marmara bölgelerinden en yüksek sayıda öğrenci alındığı ve alınan öğrencilerinde başarılı ve başarısızlık oranlarının birbirine yakın olduğu görülmektedir. Ancak alınan öğrenci sayılarını bölgelerin nüfus yoğunluğuna göre değerlendirdiğimizde ise en yoğun öğrencinin Karadeniz bölgesinden geldiği onu ise Doğu Anadolu bölgesinin izlediği ortaya çıkmaktadır.

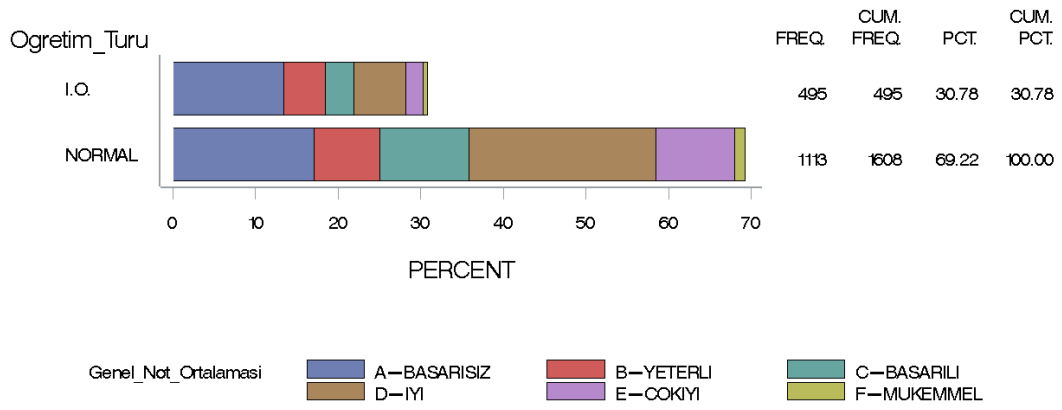



Şekil 4.25 Nüfusa kayıtlı olduğu bölge-Genel not ortalamasına göre öğrenci dağılımları

### 4.3.17 Öğretim Türü - Genel Not Ortalamasına Göre Öğrenci Dağılımı

İkinci öğretim öğrencilerinin %40'ı başarılı iken, örgün öğretim öğrencilerinin %64'ünün başarılı oldukları aşağıdaki şekilden görülmektedir. Bu doğrultuda örgün

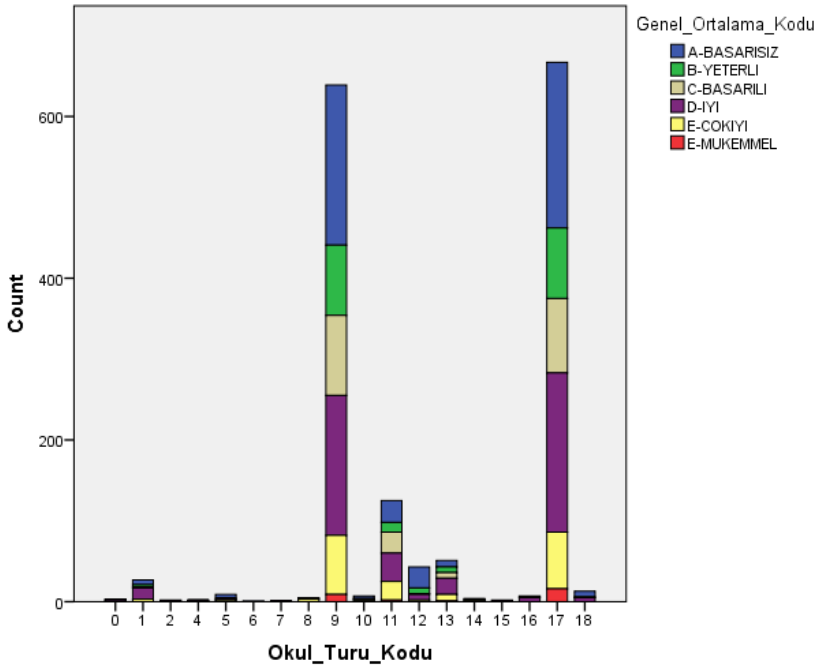
öğretim öğrencileri ikinci öğretim öğrencilerine göre daha başarılıdır demek doğru olacaktır.




Şekil 4.26 Öğretim türü-Genel not ortalamasına göre öğrenci dağılımları

### 4.3.18 Okul Türü - Genel Not Ortalamasına Göre Öğrenci Dağılımı

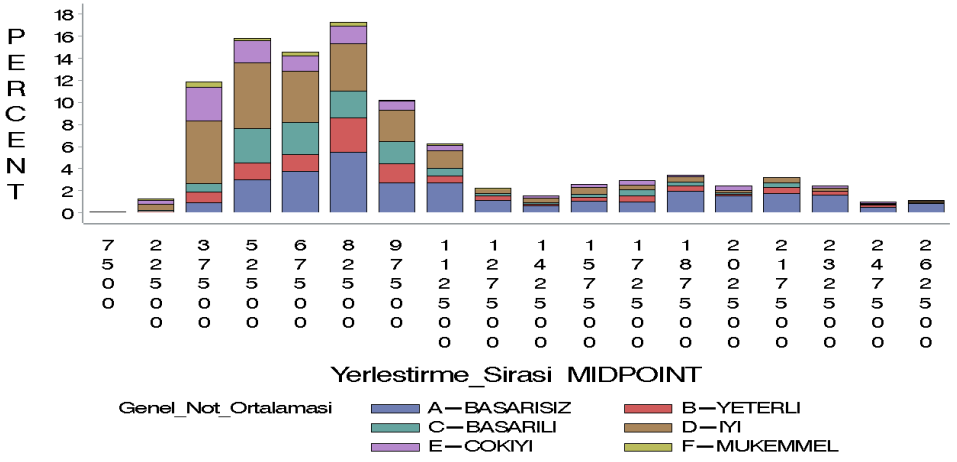
Şekil 4.27'e göre en yüksek oranda öğrenci alımı 17 numaralı Anadolu liselerinden ve 9 numaralı resmi ve gündüz öğretim yapan devlet liselerinden olmuştur. Bu liselerden gelen öğrencilerin %60'ından fazlasının başarılı olduğu grafikten görülmektedir.




Şekil 4.27 Okul Türü-Genel not ortalamasına göre öğrenci dağılımları

### 4.3.19 Yerleştirme Sırası - Genel Not Ortalamasına Göre Öğrenci Dağılımı

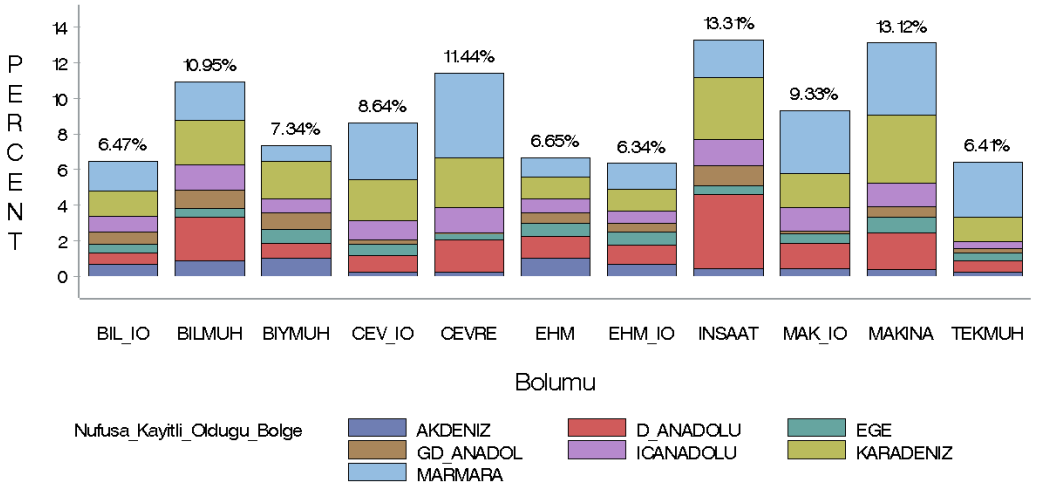
Yerleştirme sıralarına göre öğrencilerinin başarı analizini Şekil 4.28'e göre yapmak mümkün olacaktır. Şekil incelendiğinde 37 bininci sıralardan gelen öğrencilerin, tüm öğrencilerin %12'si olduğu ve bu %12'lik dilimin yaklaşık %85'inin başarılı olduğu görülmektedir. Buna karşın 82 bininci sıralardan gelen öğrencilerin tüm öğrencilerin %17'si olduğu ve bu %17'lik dilimin yaklaşık %50'sinin başarılı olduğu görülmektedir. Bu doğrultuda yerleştirme sıraları arttıkça öğrencilerin başarılı olma durumlarında düşüş olduğu yorumunu yapmak yanlış olmayacaktır.




Şekil 4.28 Yerleştirme sırası-Genel not ortalamasına göre öğrenci dağılımları

### 4.3.20 Bölüm-Nüfusa Kayıtlı Olduğu Bölgeye Göre Öğrenci Dağılımı

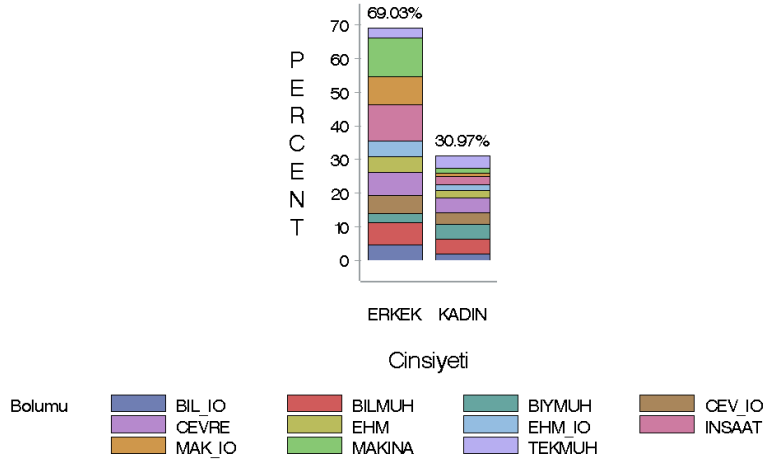
Tüm bölümlere en fazla öğrencinin Marmara bölgesinden geldiği aşağıdaki şekilden görülmektedir. Marmara bölgesini ikinci en yüksek oranla Karadeniz bölgesi takip etmektedir. İnşaat mühendisliği bölümünde ise bu söylenilenlerin aksine doğu Anadolu bölgesinden en fazla öğrenci alımı olmuştur. Güney doğu Anadolu bölgesinden gelen öğrencilerde en fazla inşaat mühendisliği bölümünü tercih etmiştir. Marmara bölgesinden gelen öğrencilerin ağırlıklı olarak makine ve çevre mühendisliği bölümünü tercih ettikleri görülmektedir. Marmara bölgesinin en yüksek sayıda öğrenci göndermesinin sebebi hem fakültenin konumundan hem de nüfus yoğunluğunun fazla olmasından dolayıdır.




Şekil 4.29 Bölüm-Nüfusa kayıtlı olduğu bölgeye göre öğrenci dağılımları

### 4.3.21 Cinsiyet-Bölüme Göre Öğrenci Dağılımı

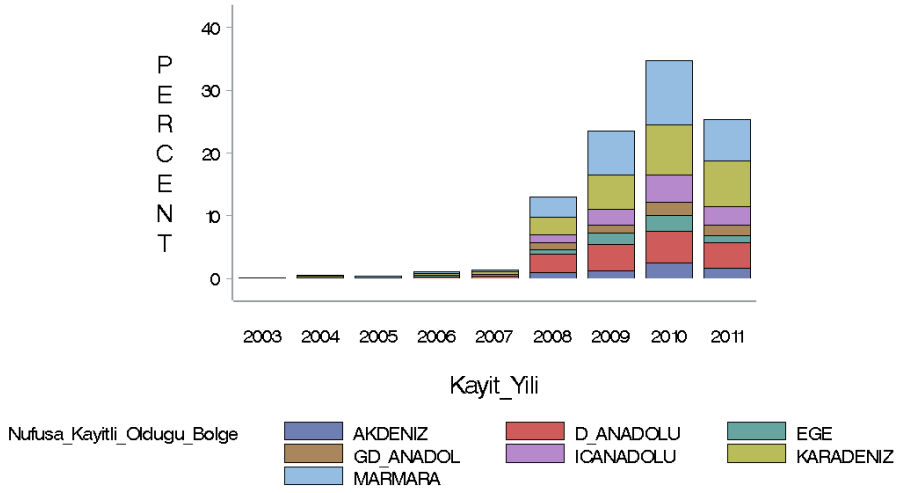
Bayanların tercih durumlarında biyomedikal ve tekstil mühendisliklerine, erkeklerin ise makine ve inşaat mühendisliklerine ağırlık verdikleri görülmektedir. Ve aynı zamanda mühendislik fakültesini tercih eden öğrencilerin %70'inin erkek %30'unun ise bayan olduğu görülmektedir.




Şekil 4.30 Cinsiyet-Bölüme göre öğrenci dağılımları

### 4.3.22 Kayıt Yılı-Nüfusa Kayıtlı Olduğu Bölgeye Göre Öğrenci Dağılımı

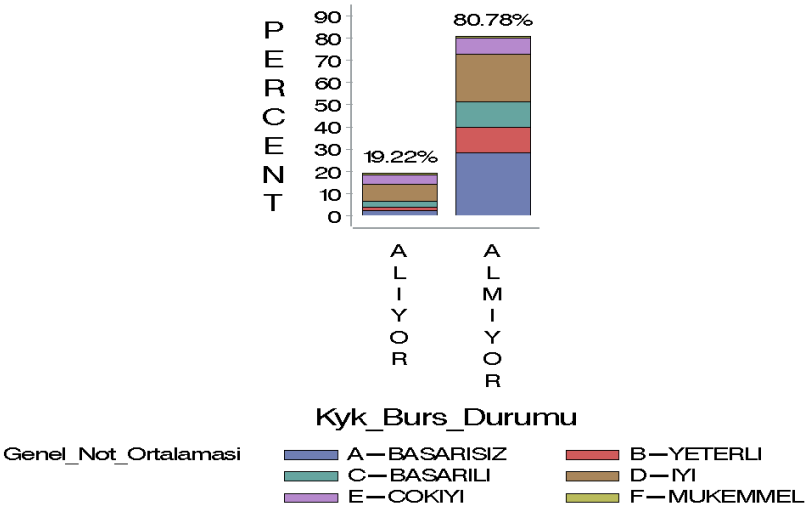
2011 yılına kadar en fazla öğrenci Marmara bölgesinden gelmekteyken, 2011 yılında en fazla öğrenci Karadeniz bölgesinden gelmiştir. 2010 yılında üniversitelere alınan öğrenci kontenjanlarındaki artış neticesinde bölgelerden gelen öğrenci sayılarında artış olduğu görülmektedir.




Şekil 4.31 Kayıt yılı-Nüfusa kayıtlı olduğu bölgeye öğrenci dağılımları

### 4.3.23 Burs Durumu - Genel Not Ortalamasına Göre Öğrenci Dağılımı

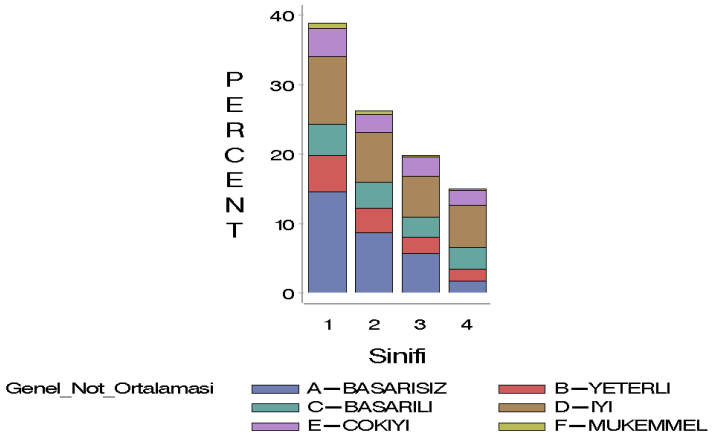
Öğrencilerin yaklaşık %20'sinin burs aldığı ve %80'inin burs almadığı, burs alan öğrencilerin %85'inin başarılı olduğu, burs almayan öğrencilerin ise %50'sinin başarılı olduğu Şekil 4.32'de görülmektedir. Burdan yola çıkarak burs alan öğrencilerin almayan öğrencilere göre daha başarılı oldukları sonucuna ulaşmak mümkündür.




Şekil 4.32 Burs Durumu-Genel not ortalamasına göre öğrenci dağılımları

### 4.3.24 Sınıfı - Genel Not Ortalamasına Göre Öğrenci Dağılımı

Sınıflara göre genel not ortalamasının değişimi Şekil 4.33'de görülmektedir. Şekle bakıldığında 1. Sınıf öğrencilerinin toplam öğrencilerin yaklaşık %39'una tekabül ettiği ve bu %39'luk dilimin %49'unun da başarılı olduğu görülmektedir.




Şekil 4.33 Sınıfı-Genel not ortalamasına göre öğrenci dağılımları

Bu analizin tüm sınıflar için yapılması neticesinde Çizelge 4.3 elde edilmiştir.

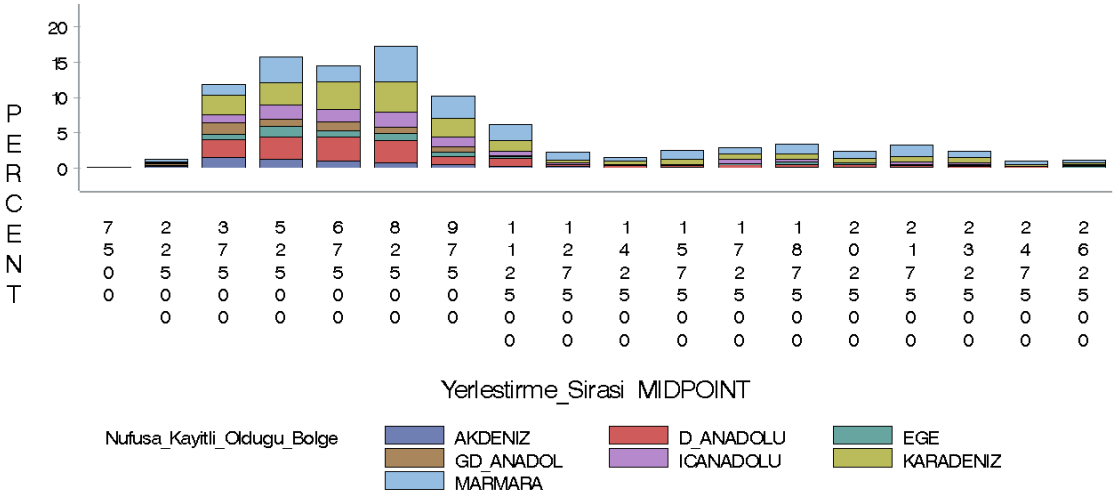
Çizelge 4.3 Sınıflara göre öğrencilerin başarı yüzdeleri

| | **Sınıf Numarası** | | | |
|---|---|---|---|---|
| | **1** | **2** | **3** | **4** |
| **Öğrenci Yüzdesi(%)** | 39 | 26 | 20 | 15 |
| **Kümelerdeki öğrencilerin lisans başarı ortalaması(%)** | 49 | 52 | 60 | 73 |

Çizelgeye bakıldığında öğrencilerin üst sınıflara geçtikçe başarı oranlarının arttığı ve en düşük başarı oranının 1. sınıftayken oluştuğu sonucuna ulaşmak mümkündür.

### 4.3.25 Yerleştirme Sırası – Nüfusa Kayıtlı Olduğu Bölgeye Göre Öğrenci Dağılımı

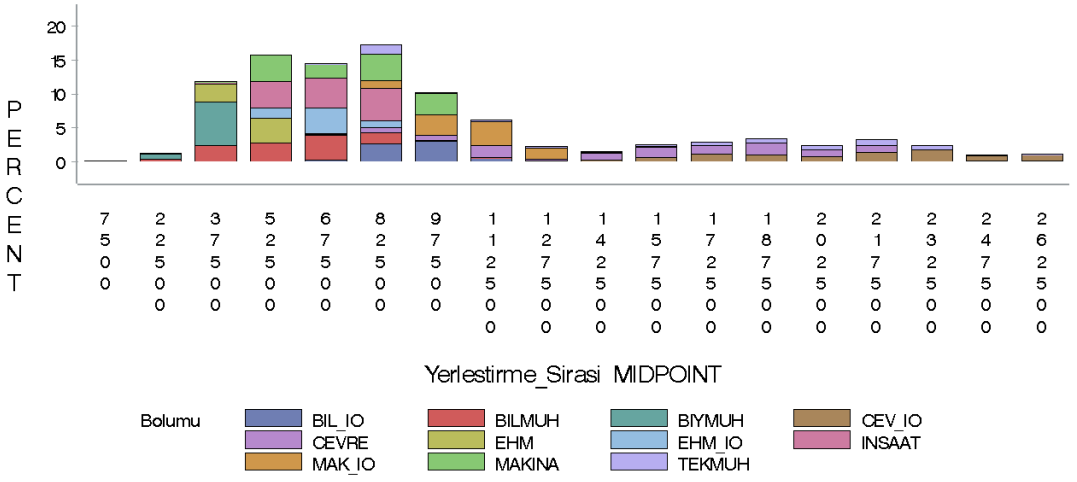
Gelen öğrencilerin yerleştirme sırası ve nüfus bölgeleri Şekil 4.34'de görülmektedir. Şekil'de en fazla öğrencinin Marmara ve Karadeniz'den geldiği görülmektedir. Yerleştirme sıralarına göre nüfus bölgesi dağılımları ise benzer oranlar göstermektedir. Yani 40 bininci sıradan girmiş olan öğrencilerin bölgelere dağılımı ile 100 bininci sıradan yerleşmiş öğrencilerin bölgelere göre dağılımı benzer özellikler göstermektedir.




Şekil 4.34 Yerleştirme sırası-Nüfusa kayıtlı olduğu bölgeye göre öğrenci dağılımları

### 4.3.26 Yerleştirme Sırası - Bölümüne Göre Öğrenci Dağılımı

Bölümlere göre yerleştirme sıraları Şekil 4.35'te yer almaktadır. Bu grafiğe bakıldığında en başarılı yerleştirme sırasına sahip öğrencilerin yoğun olarak biyomedikal mühendisliği bölümünü seçtikleri görülmektedir. Biyomedikal bölümünü sırasıyla elektronik haberleşme ve inşaat mühendisliği bölümleri izlemektedir. En kötü yerleştirme sırasına sahip öğrenciler ise çevre mühendisliği örgün ve ikinci öğretim ile tekstil mühendisliğinde yer almaktadır.

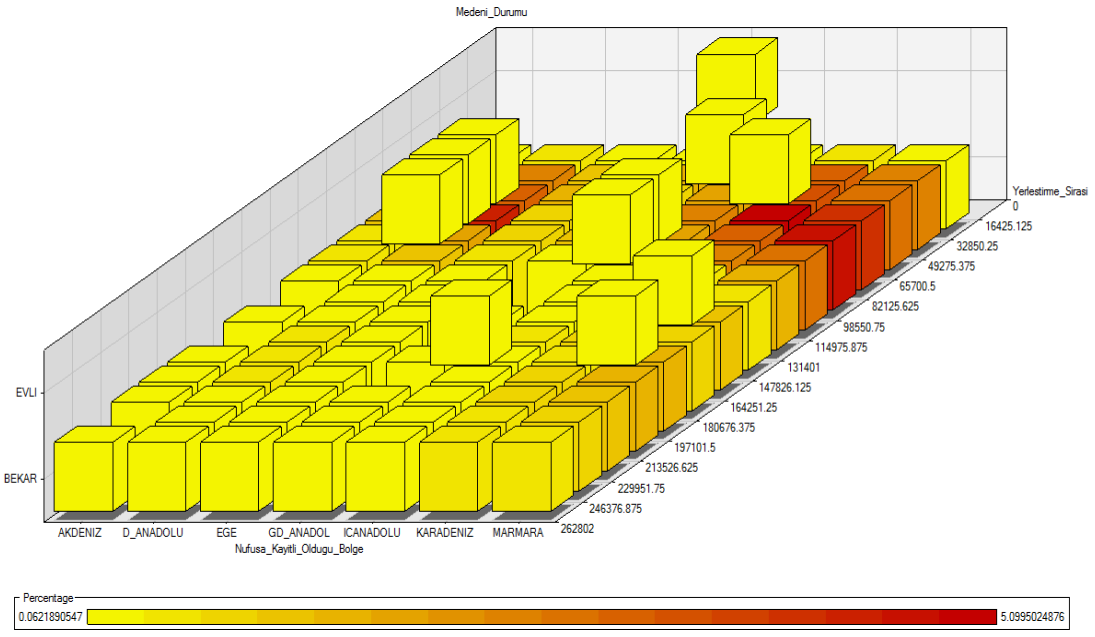



Şekil 4.35 Yerleştirme sırası-Bölüme göre öğrenci dağılımları

### 4.3.27 Yerleştirme Sırası–Medeni Durum-Nüfusa Kayıtlı Olduğu Bölgeye Göre Öğrenci Dağılımı

Şekil 4.36'da öğrencilerin medeni durumları, nüfusa kayıtlı olduğu bölgeler ve yerleştirme sıralarının yer aldığı dağılım grafiği görülmektedir. Şekle bakıldığında evli öğrencilerin Doğu Anadolu, İç Anadolu ve Marmara bölgelerinde yoğunlaştığı ve yerleştirme sıralarının genel tablo içerisinde de iyi olduğu görülüyor. Bölgelerin nüfus yoğunluğunu da düşünerek yorum yapacak olursak Doğu Anadolu bölgesinin en düşük nüfusa sahip olmasına rağmen, en yüksek nüfusa sahip Marmara bölgesi ile eşit oranlarda evli öğrenci olduğu görülmektedir. Bu yüzden evlilik oranlarını sıralarsak ilk sırada Doğu Anadolu, ikinci sırada İç Anadolu ve son sırada Marmara bölgesi yer alacaktır.

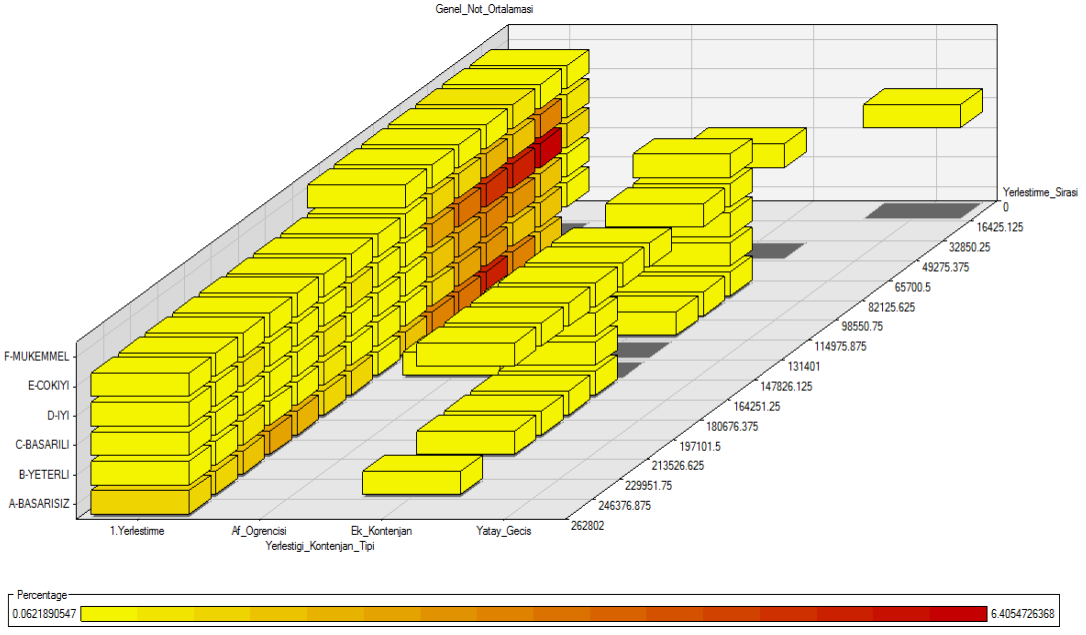
Bekâr öğrencilerin yerleştirme sıralarının 50 bin ile 100 bin aralığında yoğunlaştığı ve Karadeniz ve Marmara bölgelerinin diğer bölgelere oranla sınavda daha başarılı olmuş öğrencilere sahip olduğu görülmektedir. Bunun bir sebebi daha öncede söylenildiği gibi bu iki bölgenin nüfus yoğunluklarının yüksek olmasıdır.




Şekil 4.36 Yerleştirme Sırası – Medeni Durum- Nüfusa Kayıtlı Olduğu Bölgeye Göre Öğrenci Dağılımı

### 4.3.28 Genel Not Ortalaması-Yerleştiği Kontenjan Tipi-Yerleştirme Sırasına Göre Öğrenci Dağılımı

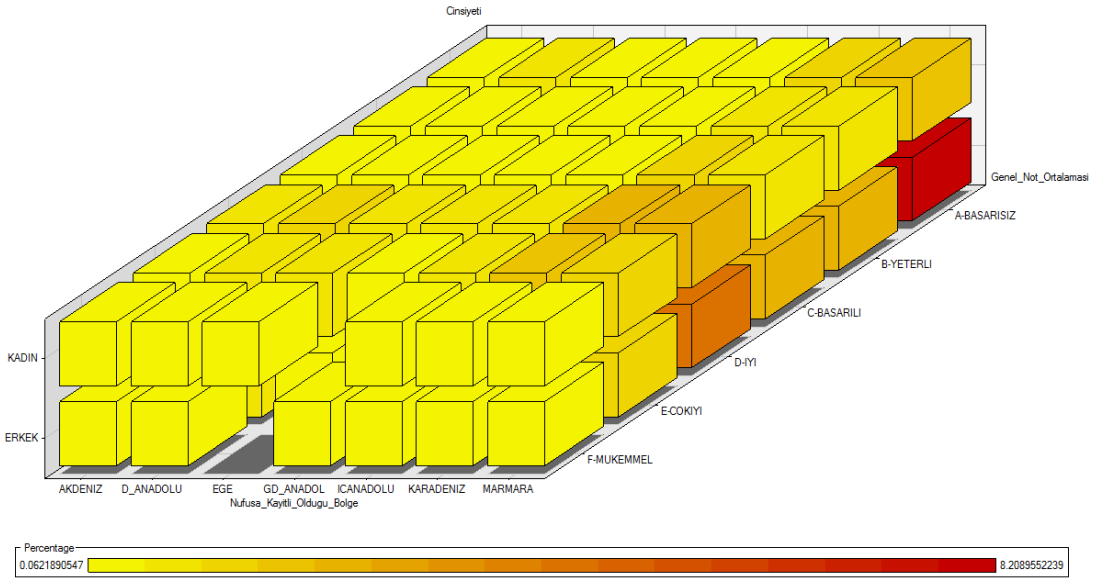
Genel not ortalamaları, yerleştiği kontenjan tipi ve yerleştirme sıralarının yer aldığı Şekil 4.37'ye bakıldığında, sınava yerleştirme sıraları ile üniversite başarıları arasında doğru orantı olduğu görülmektedir. Yani üniversite sınavında iyi bir derece almış olan öğrenci üniversite öğrenimi boyunca da aynı başarıyı gösterme eğilimindedir. 1. Yerleştirme ile üniversiteyi kazanan öğrencilerin yoğunluğu diğer yerleştirme şekillerine göre çok yüksek olduğu için okul başarılarının da daha iyi olduğu görülmektedir.




Şekil 4.37 Genel Not Ortalamasına-Yerleştiği Kontenjan Tipi-Yerleştirme Sırasına Göre Öğrenci Dağılımı

### 4.3.29 Nüfusa Kayıtlı Olduğu Bölge-Cinsiyet-Genel Not Ortalamasına Göre Öğrenci Dağılımı

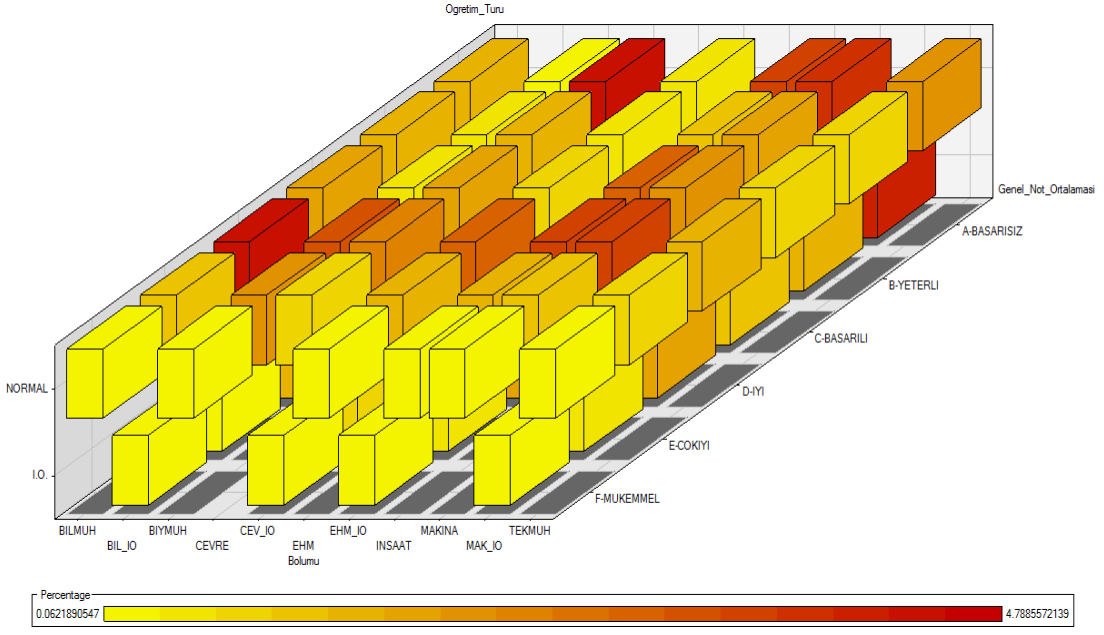
Şekil 4.38'de bölgelere göre cinsiyet ve başarı notları dağılımı gözükmektedir. Grafiğe göre Marmara bölgesinden gelen öğrenci sayısının çok olması sebebiyle başarısız olan öğrencilerin yüzdeleri de fazladır. Ve yine şekilden erkek öğrencilerin başarısızlık oranlarının bayan öğrencilere göre fazla olduğu dikkat çekmektedir. Üst seviye notlara sahip öğrencilerin sayılarının az olduğu ve genel profilin vasatın üzerinde seyrettiği de şekilden görülebilir.




Şekil 4.38 Nüfusa Kayıtlı Olduğu Bölge-Cinsiyet- Genel Not Ortalamasına Göre Öğrenci Dağılımı

### 4.3.30 Öğretim Türü-Bölüm- Genel Not Ortalamasına Göre Öğrenci Dağılımı

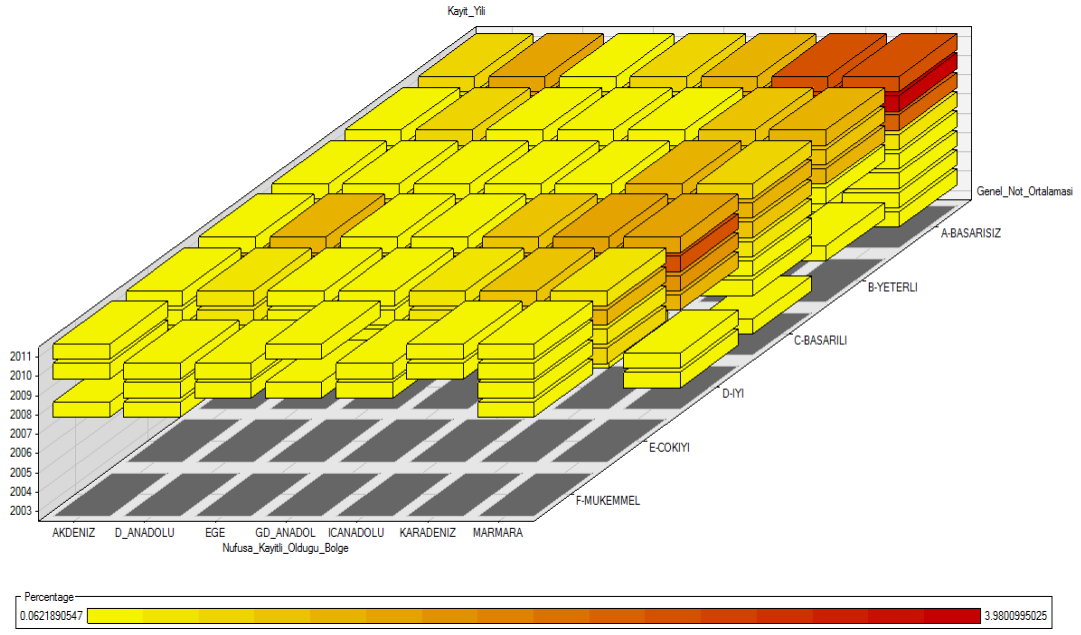
Öğretim türlerine göre bölümlerin başarı oranlarının gösterildiği Şekil 4.39'a bakıldığında başarı oranı en yüksek bölümün biyomedikal mühendisliği bölümü olduğu görülmektedir. En başarısız bölümler ise çevre mühendisliği ve makine mühendisliği bölümleridir. İkinci öğretimdeki başarının da örgün öğretime göre biraz daha düşük olduğu görülebilmektedir.




Şekil 4.39 Öğretim Türü-Bölüm- Genel Not Ortalamasına Göre Öğrenci Dağılımı

### 4.3.31 Kayıt Yılı-Nüfusa Kayıtlı Olduğu Bölge-Genel Not Ortalamasına Göre Öğrenci Dağılımı

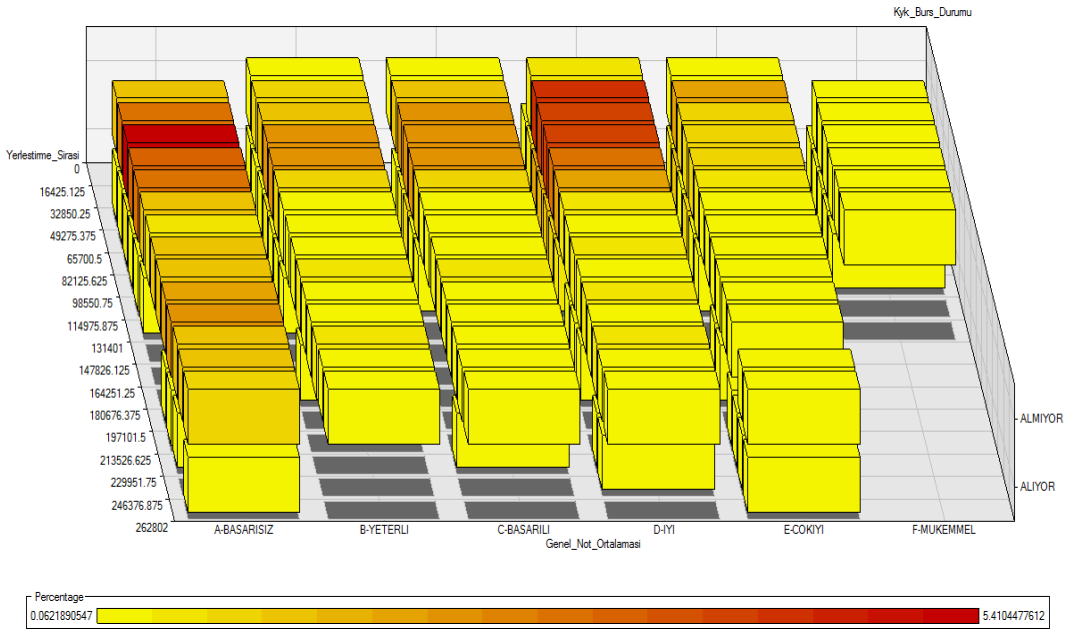
Şekil 4.40 yıllara göre öğrencilerin geldikleri bölgeleri ve üniversitedeki başarı durumlarını göstermektedir. Marmara bölgesinden gelen öğrencilerin sayılarının yüksek ancak başarı oranlarının düşük olduğu görülmektedir. Yıllar ilerledikçe gelen öğrencilerin profillerinde düşme olduğu yani başarılı olma yüzdelerinin geçmişten günümüze geldikçe düştüğü görülmektedir.




Şekil 4.40 Kayıt Yılı-Nüfusa Kayıtlı Olduğu Bölge-Genel Not Ortalamasına Göre Öğrenci Dağılımı

### 4.3.32 Yerleştirme Sırası –Burs Durumu- Genel Not Ortalamasına Göre Öğrenci Dağılımı

Yerleştirme sıralarına göre burs alma durumlarının başarılarına olan etkisini görebileceğimiz Şekil 4.41'de burs alan öğrencilerin daha başarılı oldukları görülmektedir. Bunun bir sebebi başarısız olduklarında burslarının kesilmesi olarak gösterilebilir. Ve burs alan öğrencilerin yerleştirme sıralarının da iyi olduğu görülmektedir. Burs almayan öğrencilerin başarılarının burs alanlara oranla daha düşük olduğu da şekilden görülebilmektedir.




Şekil 4.41 Yerleştirme Sırası –Burs Durumu- Genel Not Ortalamasına Göre Öğrenci Dağılımı

## 4.4 Veri Setine Kümeleme Yöntemlerinin Uygulanması

Verilerle ilgili çeşitli ön işlem aşamaları gerçekleştirildikten sonra verilere kümeleme yöntemlerinin uygulanması aşamasına geçilmiştir. Verileri kümelemek için Matlab ve Weka araçları kullanılmıştır. Kümeleme için katı kümeleme yöntemlerinden k-ortalamalar ve k-medoids ile bulanık kümeleme yöntemlerinden Bulanık c-ortalamalar, Gustafson-Kessel ve Gath-Geva yöntemleri uygulanmıştır. Her bir küme sonucunda elde edilmiş küme merkezleri çizelgeler halinde her yöntemin altında verilmiştir. Aşağıda uygulamada kullanılan ve gerçeklenen yöntemlerin sonuçları yer almaktadır.

### 4.4.1 K-ortalamalar Yönteminin Verilere Uygulanması

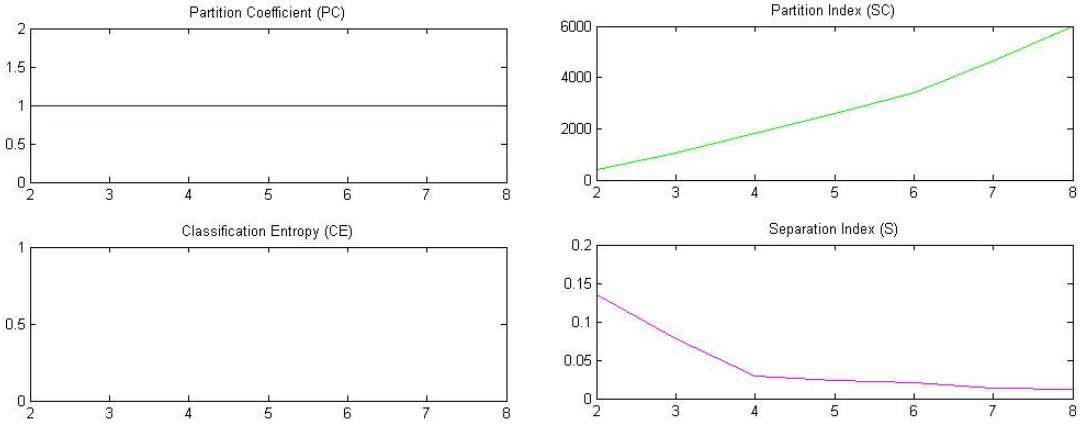
Kümeleme işlemine başlamadan önce küme sayısının kaç seçileceğine karar verilmesi gerekir. Kümeleme yöntemlerinde küme sayısının belirlenmesi başlı başına bir problem olma niteliği taşımaktadır. Küme sayılarının belirlenmesi için çeşitli yöntemler ortaya atılmıştır. Bunlardan birisi Çizelge 4.4'de sonuçları görülen doğruluk indeksi yöntemleridir. Çalışma da dört farklı doğruluk indeksi kullanılmıştır. Küme sayısı için ise 2-8 aralığında küme sayıları denenmiştir. Çizelgeyi incelediğimizde PC ve CE doğruluk indekslerinin k-ortalamalar yöntemi için doğru sonuçlar üretemeyeceği görülmektedir. Bunun sebebi ise PC ve CE yöntemlerinin hesabında kullanılan değişkenlerin değerlerinin k-ortalamalar yöntemi sonucunda elde edilen değerlerle

uyumlu olmamasıdır. PC ve CE yöntemleri hesap edilirken üyelik değeri($u_{ij}$) kullanılır. K-ortalamalar yönteminde de bir nesne bir kümeye ya üyedir ya da değildir. Yani üyelik değeri ya 1'dir ya da 0. Bu yüzden çizelge de PC değeri 1, CE değeri NaN(not a number) olmuştur. Şekil 4.42'de doğruluk indeksi değerlerinin grafik olarak gösterimi yer almaktadır.

Çizelge 4.4 Farklı küme sayıları sonucunda k-ortalamalar yöntemi için doğruluk değerleri

| **Doğruluk İndeksi** | **Küme Sayısı** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **PC** | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **CE** | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **SC** | 410 | 1041 | 1810 | 2585 | 3400 | 4654 | 5995 |
| **S** | 0.1361 | 0.0776 | 0.0294 | 0.0231 | 0.0214 | 0.0141 | 0.0123 |

Doğruluk indeksleri yöntemleri de kendi aralarında farklı sonuçlar üretebilir. Şekil 4.42'ye bakıldığında SC' ye göre küme sayısının 2 seçilmesi uygun gözükürken S'ye göre küme sayısının 8 olarak seçilmesi uygundur. Ancak küme sayısı 2 seçildiğinde veri seti içerisindeki farklı örüntüler tespit edilemeyeceğinden ve küme sayısı 8 seçildiğinde de istenilenden fazla örüntü ortaya çıkacağından küme sayısı 4 olarak belirlenmiştir. Çizelge 4.4'de ki değerlerin sonuçlarını daha iyi anlayabilmek için Şekil 4.42'ye bakmak faydalı olacaktır.




Şekil 4.42 K-ortalamalar yöntemi için geçerlilik indeksleri

K-ortalamalar kümeleme yönteminin verilere uygulanmasının ardından elde edilmiş kümeleme sonuçları Çizelge 4.10'da yer almaktadır. K-ortalamalar yönteminde çeşitli küme sayıları denenmiş ve küme sayısı 4 olarak seçilmiştir. Birinci küme de %47'lik oranla çevre mühendisliği ikinci öğretim, sınıf olarak 1.sınıf öğrencileri ağır basmaktadır.

1.kümede yer alan öğrencilerin çoğunluğu Marmara bölgesinden gelmiştir. Bölgelerin nüfus yoğunluğu düşünüldüğünde ise 1.küme için Karadeniz bölgesi demek gerekmektedir. Çünkü 1. kümede 108 öğrenci Marmara bölgesinden 68 öğrenci ise Karadeniz bölgesinden gelmiştir. Marmara bölgesinin nüfusu yaklaşık 22 milyon iken, Karadeniz bölgesinin nüfusu 7.5 milyon civarındadır. Öğrenci sayıları ile bölgelerin nüfus yoğunluklarını birlikte değerlendirdiğimizde 1.kümeye neden Karadeniz bölgesi dediğimiz ortaya çıkmaktadır. Bölgelerin nüfus yoğunluğuna göre hesabı aşağıdaki formüller yardımıyla yapılmaktadır;

$$BYK_j = \frac{TN}{BN_J} \tag{4.1}$$

$BYK_J$: Her bir bölgenin yoğunluk katsayısı

$BN_J$: j. bölgenin toplam nüfusu

TN: Tüm bölgelerin toplam nüfusu

$$KT_i = \sum_{j=1}^{7} KB_{ij} \tag{4.2}$$

$KT_i$: i. kümedeki toplam öğrenci sayısı

$KB_{ij}$: i. kümedeki j. bölge öğrenci sayısı

$$KBY_{ij} = \frac{KB_{ij}}{KT_i} \tag{4.3}$$

$$KBYK_{ij} = KBY_{ij} * BYK_J \tag{4.4}$$

$KBY_{ij}$: i. kümedeki j. bölge yoğunluğu

$KBYK_{ij}$: i. kümedeki j. bölgenin yoğunluk katsayısı

Yukarıda verilen formüller ışığında 1. küme için nüfus bölgelerinin yoğunluğu şu şekilde hesaplanır;

Çizelge 4.5'de her bölgenin nüfus miktarı yer almaktadır. Çizelge 4.6'da ise her bir bölgenin yoğunluk katsayıları yer almaktadır. Bu katsayılar (4.1) ile hesaplanan değerlerdir. Çizelge 4.7'de ise küme-1 için bölgelerden gelen öğrenci sayıları yer

almaktadır. Bu değerlerin (4.3)'e göre küme-1 de yer alan toplam öğrenci sayısına bölünmesiyle elde edilen katsayı oranları Çizelge 4.8'de görülmektedir. En son aşamada ise (4.4) yardımıyla Çizelge 4.6'da yer alan katsayılar ile Çizelge 4.8' de yer alan katsayılar çarpılarak her bölge için yoğunluk katsayısı bulunmuştur. Çizelge 4.9'da görülen bu yoğunluk katsayısı ne kadar büyük ise o bölgeden gelen öğrenci sayısı o kadar yoğun olmaktadır.

Çizelge 4.5 Bölgelere Göre Nüfus Dağılımları

| | |
|---|---|
| $BN_1$: | Marmara= 22.387.173 |
| $BN_2$: | İç Anadolu= 11.965.642 |
| $BN_3$: | Ege= 9.687.692 |
| $BN_4$: | Akdeniz= 9.620.240 |
| $BN_5$: | Karadeniz= 7.508.293 |
| $BN_6$: | Güney Doğu Anadolu= 6.923.256 |
| $BN_7$: | Doğu Anadolu= 6.631.973 |
| TN: | TOPLAM NÜFUS=74.724.269 |

Çizelge 4.6 Bölgelere Göre Nüfus Katsayıları

| |
|---|
| $BYK_1$= TN /$BN_1$:=74.724.269/22.387.173=3.3 |
| $BYK_2$= TN /$BN_2$=74.724.269/11.965.642=6.2 |
| $BYK_3$= TN /$BN_3$=74.724.269/9.687.692=7.7 |
| $BYK_4$= TN /$BN_4$=74.724.269/9.620.240=7.8 |
| $BYK_5$= TN /$BN_5$=74.724.269/7.508.293=9.9 |
| $BYK_6$= TN /$BN_6$=74.724.269/6.923.256=10.8 |
| $BYK_7$= TN /$BN_7$=74.724.269/6.631.973=11.3 |

Çizelge 4.7 Küme-1 için Bölgelerden Gelen Öğrenci Sayıları

| | |
|---|---|
| $KB_{11}$: | Marmara= 108 |
| $KB_{12}$: | İç Anadolu= 31 |
| $KB_{13}$: | Ege= 17 |
| $KB_{14}$: | Akdeniz= 8 |
| $KB_{15}$: | Karadeniz= 68 |

Çizelge 4.7 (devamı)

| | |
|---|---|
| $KB_{16}$: | Güney Doğu Anadolu= 5 |
| $KB_{17}$: | Doğu Anadolu= 31 |
| $KT_1$: | 1.Küme için toplam öğrenci sayısı=268 |

Çizelge 4.8 Küme-1 için bölgelere göre ağırlık katsayıları

| |
|---|
| $KBY_{11}= KB_{11}/ KT_1=108/268=0.40$ |
| $KBY_{12}= KB_{12}/ KT_1=31/268=0.12$ |
| $KBY_{13}= KB_{13}/ KT_1=17/268=0.06$ |
| $KBY_{14}= KB_{14}/ KT_1=8/268=0.03$ |
| $KBY_{15}= KB_{15}/ KT_1=68/268=0.25$ |
| $KBY_{16}= KB_{16}/ KT_1=5/268=0.02$ |
| $KBY_{17}= KB_{17}/ KT_1=31/268=0.12$ |

Çizelge 4.9 Küme-1 için bölgelere göre yoğunluk katsayıları

| |
|---|
| $KBYK_{11}= BYK_1 * KBY_{11}=0.40*3.3= 1.32$ |
| $KBYK_{12}= BYK_2 * KBY_{12}=0.12*6.2=0.74$ |
| $KBYK_{13}= BYK_3 * KBY_{13}=0.06*7.7=0.46$ |
| $KBYK_{14}= BYK_4 * KBY_{14}=0.03*7.8=0.234$ |
| $KBYK_{15}= BYK_5 * KBY_{15}=0.25*9.9=2.475$ |
| $KBYK_{16}= BYK_6 * KBY_{16}=0.02*10.8=0.216$ |
| $KBYK_{17}= BYK_7 * KBY_{17}=0.12*11.3=1.356$ |

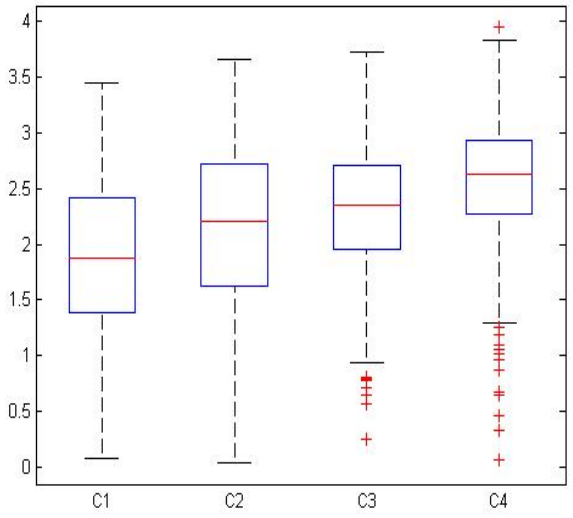
Küme-1 için yapılan değerlendirme sonucunda Çizelge 4.9'a bakıldığında 2.475 katsayı oranı ile Karadeniz Bölgesi birinci sırada, 1.356 katsayı oranı ile Doğu Anadolu Bölgesi ikinci sırada ve 1.32 katsayı oranı ile Marmara bölgesi üçüncü sırada yer almaktadır. Diğer kümeler içinde aynı değerlendirme yapılmıştır. Çizelge 4.10'da dört küme için elde edilmiş küme merkezi sonuçları yer almaktadır. Çalışmada kullanılan tüm yöntemlerde yukarıda bahsedilen nüfus yoğunluk hesabı yapılmış ve elde edilen yoğunluk katsayıları küme merkezlerini gösteren çizelgelerde en yoğun üç bölge için bölgenin satırında yer almaktadır.

Çizelge 4.10'a göre 1. küme dört küme içerisinde başarısız öğrencilerin en yoğun olduğu kümedir. Yerleştirme sırası 200 bin civarındadır. Genel ortalama notları da yerleştirme sıraları ile doğru orantılı bir şekilde çok düşüktür. Yapılan kümeleme sonucunda en başarılı bölümler biyomedikal ve bilgisayar mühendisliği olmuştur. 4. kümede ağırlıklı olan biyomedikal ve bilgisayar mühendisliği öğrencileri üniversiteye iyi sıralardan girdikleri gibi üniversitede de başarılarını devam ettirmişlerdir. Çizelge 4.10'da yer alan kümeler de dikkat çeken bir diğer nokta örgün öğretim öğrencilerinin ikinci öğretim öğrencilerinden daha başarılı bir grafik çizmeleridir. Kümelere genel olarak bakıldığında tüm kümelerin 1. Yerleştirme ile yerleşmiş, disiplin cezası almamış, bekâr, Türk asıllı ve erkek ağırlıklı, öğrenim kredisi ve burs almayan öğrencilerden oluştuğu görülmektedir.

Çizelge 4.10 k-ortalamalar yöntemi sonucunda elde edilmiş küme merkezleri

| **k-ortalamalar** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Bölümü(%)** | Çevre İ.Ö.:47<br>Çevre:30 | Makine İ.Ö.:38<br>Çevre:33 | Makine:23<br>İnşaat:22 | Biyomedikal:23<br>Bilgisayar:21 |
| **Yer. Kon. Tipi** | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme |
| **Sınıfı** | 1 | 1 | 1 | 4 |
| **Eğitim Yılı** | 2 | 1 | 2 | 4 |
| **Öğretim Türü** | Örgün | İkinci Ö. | Örgün | Örgün |
| **Disiplin Cezası** | Almamış | Almamış | Almamış | Almamış |
| **Cinsiyeti(%)** | Erkek:57<br>Kadın:43 | Erkek:71<br>Kadın:29 | Erkek:75<br>Kadın:25 | Erkek:67<br>Kadın:33 |
| **Yaşı** | 20.76 | 20.97 | 20.83 | 21.81 |
| **Medeni Hal** | Bekâr | Bekâr | Bekâr | Bekâr |
| **Türk Asıllı** | Evet | Evet | Evet | Evet |
| **Nüfus Bölgesi (Katsayı Oranı)** | Karadeniz:2.47<br>Doğu A.:1.356<br>Marmara:1.32 | Karadeniz:2.4<br>Doğu A.:2.08<br>Marmara:0.86 | Karadeniz:2.4<br>Doğu A.:1.35<br>Marmara:1.26 | Doğu A.:2.28<br>Karadeniz:2.2<br>G.Doğu A.:0.96 |
| **Okul Türü(%)** | Resmi Lise: 61<br>Anadolu L.: 20 | Resmi Lise:44<br>Anadolu L.:37 | Resmi Lise:40<br>Anadolu L.:45 | Resmi Lise:28<br>Anadolu L.:51 |
| **Yerleştirme Puanı** | 289.37 | 324.22 | 362.74 | 359.49 |
| **Yerleştirme Sırası** | 206978.68 | 121058.52 | 80719.53 | 47410.49 |
| **Tercih Sırası** | 11.29 | 9.97 | 10.29 | 10.43 |
| **Genel Ortalama** | **1.87** | **2.14** | **2.32** | **2.57** |
| **Katkı Kredisi** | Almıyor | Almıyor | Almıyor | Alıyor |
| **Öğrenim Kr.** | Almıyor | Almıyor | Almıyor | Almıyor |
| **Burs** | Almıyor | Almıyor | Almıyor | Almıyor |
| **Yüzde On(%)** | Evet:7<br>Hayır:93 | Evet:9<br>Hayır:91 | Evet:9<br>Hayır:91 | Evet:10<br>Hayır:90 |

K-ortalamalar kümeleme yönteminin veri setine uygulanmasının ardından genel başarı notları için aykırı değer tespiti yapılmıştır. Genel başarı notları için elde edilen aykırı değerler Şekil 4.43'de yer alan Box-plot analizinde görülmektedir. C1, C2, C3 ve C4 küme numaralarını göstermektedir. 1. ve 2. küme için aykırı değer oluşmadığı ve en yoğun aykırı değerlerin küme 4 de oluştuğu görülmektedir. Toplam 1588 kayıtın 21'i aykırı değer olarak işaretlenmiştir.




Şekil 4.43 K-ortalamalar yöntemi GNO özelliği için Box-plot analizi

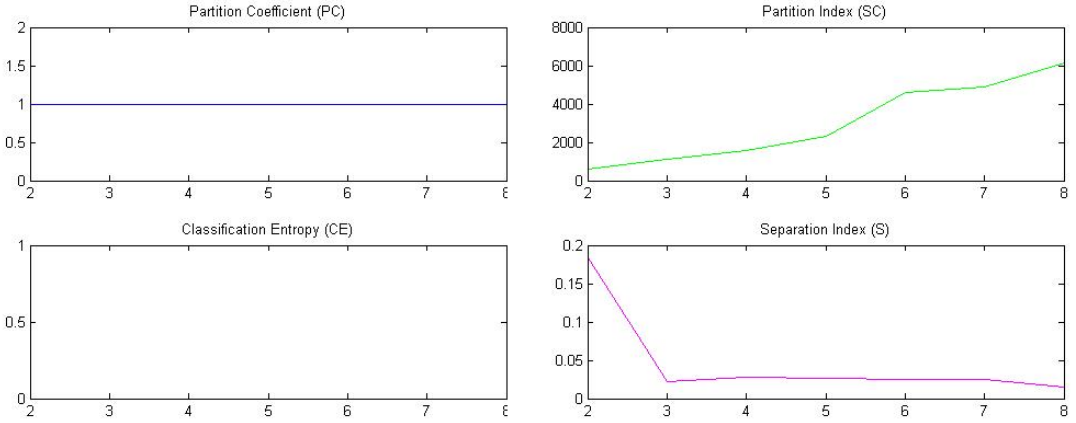
### 4.4.2 K-medoids Yönteminin Verilere Uygulanması

K-medoids yönteminde küme sayısının belirlenmesi için kullanılan doğruluk indekslerinden elde edilmiş sonuçlar Çizelge 4.11'de yer almaktadır. Çizelge incelendiğinde elde edilen sonuçların k-ortalamalar yöntemi ile benzer sonuçlar ürettiği görülebilmektedir. K-medoids yönteminde de PC ve CE yöntemlerinin kullanılamayacağı çizelgeden görülmektedir.

Çizelge 4.11 Farklı küme sayıları sonucunda k-medoids yöntemi için doğruluk değerleri

| **Doğruluk İndeksi** | **Küme Sayısı** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **PC** | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **CE** | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **SC** | 577 | 1088 | 1574 | 2308 | 4588 | 4886 | 6156 |
| **S** | 0.1846 | 0.0216 | 0.0285 | 0.0269 | 0.025 | 0.0256 | 0.0151 |

Şekil 4.44'de doğruluk indeksi değerlerinin grafik olarak gösterimi yer almaktadır.




Şekil 4.44 K-medoids yöntemi için geçerlilik indeksleri

Doğruluk indeksleri yöntemleri de kendi aralarında farklı sonuçlar üretebilmektedir. K-ortalamalar yönteminde de olduğu gibi Şekil 4.43'e bakıldığında SC' ye göre küme sayısının 2 seçilmesi uygun gözükürken S'ye göre küme sayısının 8 olarak seçilmesi uygundur. K- ortalamalar yönteminde olduğu gibi bu yöntemde aynı sebeplerden dolayı küme sayısı 4 olarak seçilmiştir.

Katı kümeleme yöntemlerinden olan k-medoids yönteminin veri setine uygulanması ile elde edilmiş küme merkezleri Çizelge 4.12'de yer almaktadır. K-medoids yöntemi k-ortalamalar yöntemine benzer sonuçlar üretmiştir.

En kötü yerleştirme sırasına ve en kötü genel not ortalamasına sahip olan bir numaralı küme elemanları ağırlıklı olarak çevre mühendisliği öğrencilerinden oluşmaktadır. Bu kümede yer alan öğrenciler 2.55 yoğunluk katsayısı ile en fazla Karadeniz bölgesinden gelmişlerdir. Dördüncü kümede yer alan öğrenciler ağırlıklı olarak Doğu Anadolu bölgesinden gelmişlerdir ve en başarılı küme olma özelliği taşımaktadırlar. Dördüncü kümedeki öğrenciler çoğunlukla son sınıf, biyomedikal ve elektronik-haberleşme mühendisliği öğrencilerinden oluşmaktadırlar. Genel not ortalamaları ve yerleştirme sıraları ile başarı puanları da en yüksek olan kümedir. Ve en çok Anadolu liselerinden gelen öğrenciler yer almaktadır.

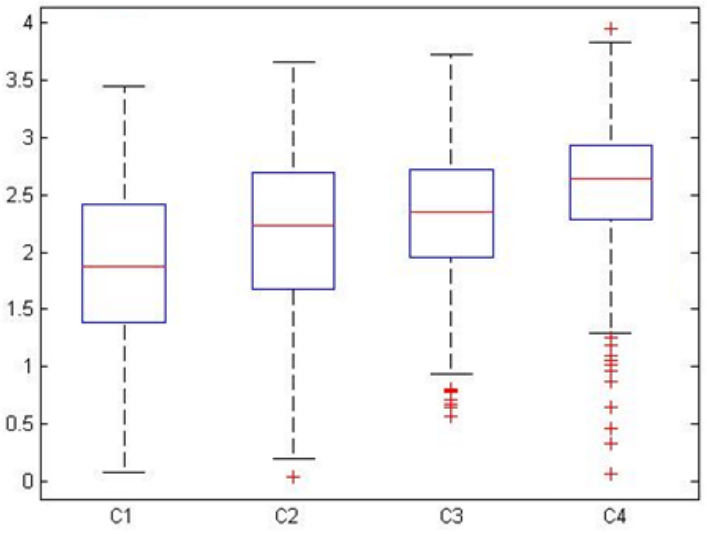
Üç numaralı küme yerleştirme sırası, genel not ortalaması ve yerleştirme puanı bakımından en başarılı ikinci kümedir. İki numaralı kümede yer alan öğrencilerin ise devlet liselerinden mezun ve makine mühendisliği ikinci öğretim ağırlıklı oldukları, yerleştirme sıraları ve puanları bakımından da en başarısız ikinci küme oldukları görülmektedir.

Çizelge genel anlamda incelendiğinde yüksek puanlarla yerleşmiş öğrencilerin üniversite sıralarında başarılı olma eğilimlerinin düşük puanlı öğrencilere göre daha yüksek olduğu görülebilmektedir. Çizelgeye okul türleri anlamında bakıldığında da Anadolu liselerinden gelen öğrencilerin, devlet liselerinden gelen öğrencilere göre hem daha iyi puanlarla yerleştikleri hem de yerleştikten sonrada daha başarılı oldukları sonucuna ulaşmak mümkündür.

Çizelge 4.12 k-medoids yöntemi sonucunda elde edilmiş küme merkezleri

| **k-medoids** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Bölümü(%)** | Çevre İ.Ö.:46<br>Çevre:33 | Mak. İ.Ö.:40<br>Çevre: 28 | İnşaat:23<br>Makine:22 | Biyomedikal:23<br>Elek.Hab.:21 |
| **Yer. Kon. Tipi** | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme |
| **Sınıfı** | 1 | 1 | 1 | 4 |
| **Eğitim Yılı** | 2 | 1 | 2 | 4 |
| **Öğretim Türü** | Örgün | İkinci Ö. | Örgün | Örgün |
| **Disiplin Cezası** | Almamış | Almamış | Almamış | Almamış |
| **Cinsiyeti(%)** | Erkek:58<br>Kadın:42 | Erkek:71<br>Kadın:29 | Erkek:75<br>Kadın:25 | Erkek:67<br>Kadın:33 |
| **Yaşı** | 21 | 21 | 21 | 22 |
| **Medeni Hal** | Bekâr | Bekâr | Bekâr | Bekâr |
| **Türk Asıllı** | Evet | Evet | Evet | Evet |
| **Nüfus Bölgesi** | Karadeniz:2.55<br>Marmara:1.32<br>Doğu A.:1.29 | Karadeniz:2.59<br>Doğu A.:1.59<br>Marmara:1.23 | Karadeniz:2.4<br>Doğu A.:2.15<br>Marmara:0.8 | Doğu A.:2.30<br>Karadeniz:2.14<br>G.D.Anadolu:0.99 |
| **Okul Türü(%)** | Resmi Lise:61<br>Anadolu L.:19 | Resmi Lise:41<br>Anadolu L.: 38 | Resmi Lise:39<br>Anadolu L.:46 | Resmi Lise:26<br>Anadolu L.:50 |
| **Yer. Puanı** | 289.18 | 327.98 | 365.4 | 357.98 |
| **Yer. Sırası** | 205019 | 116310 | 79158 | 46923 |
| **Tercih Sırası** | 1 | 3 | 1 | 4 |
| **Genel Ortalama** | **1.88** | **2.14** | **2.32** | **2.58** |
| **Katkı Kredisi** | Almıyor | Almıyor | Almıyor | Alıyor |
| **ÖğrenimKredisi** | Almıyor | Almıyor | Almıyor | Almıyor |
| **Burs** | Almıyor | Almıyor | Almıyor | Almıyor |
| **Yüzde On(%)** | Evet:7<br>Hayır:93 | Evet:9<br>Hayır:91 | Evet:9<br>Hayır:91 | Evet:10<br>Hayır:90 |

K-medoids kümeleme yönteminin veri setine uygulanmasının ardından genel başarı notları için yapılan aykırı değer tespiti Şekil 4.45'de görülmektedir. C1, C2, C3 ve C4 küme numaralarını göstermek üzere 1. küme için aykırı değer oluşmadığı ve en yoğun aykırı değerlerin küme 4'te oluştuğu görülmektedir. Toplam 1588 nesnenin 21 tanesi aykırı değer olarak işaretlenmiştir.




Şekil 4.45 K-medoids yöntemi GNO özelliği için Box plot analizi

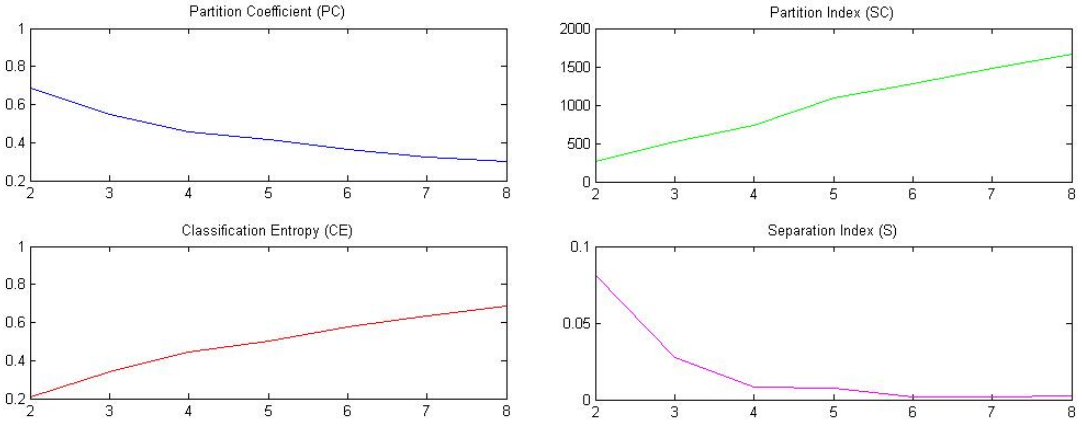
### 4.4.3 Bulanık c-ortalamalar Yönteminin Verilere Uygulanması

BCO yönteminde de küme sayısının belirlenmesi hususunda geçerlilik indeksleri kullanılmıştır. Küme sayısı olarak 2 ile 8 aralığında küme için Bölünme Katsayısı (Partititon Coefficient), Sınıflandırma Entropisi (Classification Entropy), Bölümlendirme İndeksi (Partition Index) ve Ayırma İndeksi ( Seperation Index) yöntemleri ayrı ayrı denenmiş ve aşağıdaki çizelgede yer alan sonuçlar elde edilmiştir. Çizelgeye bakıldığında küme sayısının PC, CE ve SC'ye göre 2 olarak seçilmesinin veri seti için en uygun değer olacağı görülmektedir. Küme sayısını 2 seçmenin veriler içerisinden aranan bilgilerin bulunmasını zorlaştıracağından küme sayısı 4 olarak seçilmiştir.

Çizelge 4.13 Bulanık c-ortalamalar yöntemi için doğruluk indeksi sonuçları

| **Doğruluk İndeksi** | **Küme Sayısı** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **PC** | 0.6887 | 0.5459 | 0.4587 | 0.4162 | 0.3614 | 0.3264 | 0.2992 |
| **CE** | 0.2093 | 0.3438 | 0.443 | 0.504 | 0.579 | 0.637 | 0.688 |
| **SC** | 258 | 515 | 739 | 1091 | 1276 | 1482 | 1663 |
| **S** | 0.0816 | 0.0272 | 0.0079 | 0.0072 | 0.0019 | 0.0021 | 0.0025 |

Şekil 4.44'de Çizelge 4.13'de yer alan çizelgenin grafiksel gösterimi yer almaktadır.




Şekil 4.46 Bulanık c-ortalamalar yöntemi için geçerlilik indeksleri

BCO yönteminin mühendislik fakültesi öğrenci verilerine uygulanmasının ardından elde edilen küme merkezleri Çizelge 4.14'de yer almaktadır. Çizelge incelendiğinde elde edilen sonuçların katı kümeleme yöntemlerinden elde edilen sonuçlarla benzerlikler gösterdiği görülebilmektedir.

1. Küme de bölüm olarak çevre mühendisliği ikinci öğretim, sınıf olarak 1.sınıf öğrencileri ağır basmaktadır. 1. kümede yer alan öğrenciler en yoğun Karadeniz bölgesinden gelmişler ve 200 bininci sıralardan üniversiteye yerleşmişlerdir. Üniversiteye yerleşme sıraları ile doğru orantılı bir şekilde, üniversitedeki başarı notları da en kötü durumdadır. Hatta 4 küme içerisindeki en kötü başarı notları ve en yüksek yani en kötü yerleştirme sıraları da bu kümeye aittir.

Küme 2'de bölüm olarak çevre mühendisliği, sınıf olarak 3.sınıf öğrencileri çoğunluktadır. Bu kümedeki öğrenciler yoğun olarak Karadeniz bölgesinden gelmişler ve 170 bininci sıralardan üniversiteye yerleşmişlerdir. Genel başarı ortalamaları bakımından ise en başarısız ikinci küme özelliği göstermektedirler.

3 ve 4. kümeler en başarılı öğrencilerin yer aldığı kümelerdir. 3.kümede yer alan öğrencilerin üniversiteye yerleşme sıraları 90 binlerde iken, 4. küme de yer alan öğrencilerin yerleşme sıraları 50 binlerdedir. 4. küme de bilgisayar ve biyomedikal mühendisliği öğrencileri ve 3. küme de makine mühendisliği öğrencileri yoğunluktadır. En başarılı öğrencilere sahip kümelerde yer alan öğrencilerin Karadeniz ve Doğu Anadolu bölgesi ağırlıklı olduğu gözlemlenmektedir.

Çizelge 4.14 Bulanık c-ortalamalar yöntemi sonucunda elde edilmiş küme merkezleri

| **BCO** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Bölümü(%)** | Çevre İ.Ö.:57<br>Tekstil:22 | Çevre: 54<br>Çevre İ.Ö.:27 | Makine İ.Ö.:22<br>Makine:19 | Bilgisayar:23<br>Biyomedikal:19 |
| **Yer. Kon. Tipi** | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme |
| **Sınıfı** | 1 | 3 | 1 | 4 |
| **Eğitim Yılı** | 1 | 3 | 1 | 2 |
| **Öğretim Türü** | İkinci | Örgün | Örgün | Örgün |
| **Disiplin Cezası** | Almamış | Almamış | Almamış | Almamış |
| **Cinsiyeti(%)** | Erkek:61<br>Kadın:39 | Erkek:55<br>Kadın:45 | Erkek:78<br>Kadın:22 | Erkek:65<br>Kadın:35 |
| **Yaşı** | 20.56 | 21.25 | 20.88 | 21.73 |
| **Medeni Hal** | Bekâr | Bekâr | Bekâr | Bekâr |
| **Türk Asıllı** | Evet | Evet | Evet | Evet |
| **Nüfus Bölgesi (Katsayı Oranı)** | Karadeniz:2.5<br>Marmara:1.38<br>Doğu A.:1.09 | Karadeniz:2.60<br>Doğu A.:1.55<br>Marmara:1.34 | Karadeniz:2.39<br>Doğu A.:1.96<br>Marmara:1.01 | Karadeniz:2.31<br>Doğu A.:2.30<br>Marmara:0.58 |
| **Okul Türü(%)** | Resmi Lise:65<br>Anadolu L.:20 | Resmi Lise:53<br>Anadolu L.:22 | Resmi Lise:41<br>Anadolu L.:43 | Resmi Lise:28<br>Anadolu L.:50 |
| **Yerleştirme Puanı** | 286.52 | 295.18 | 354.53 | 362.12 |
| **Yerleştirme Sırası** | 228079.76 | 170761.71 | 90975.70 | 50368.53 |
| **Tercih Sırası** | 11.58 | 10.00 | 10.45 | 10.46 |
| **Genel Ortalama** | **1.76** | **2.08** | **2.26** | **2.56** |
| **Katkı Kredisi** | Almıyor | Almıyor | Almıyor | Alıyor |
| **Öğrenim Kredisi** | Almıyor | Almıyor | Almıyor | Almıyor |
| **Burs** | Almıyor | Almıyor | Almıyor | Almıyor |
| **Yüzde On(%)** | Evet:6<br>Hayır:94 | Evet:7<br>Hayır:93 | Evet:8<br>Hayır:92 | Evet:10<br>Hayır:90 |

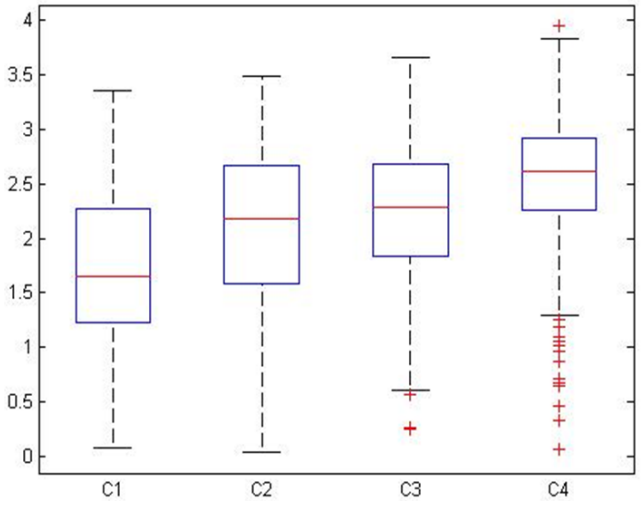
Bulanık c-ortalamalar yönteminin k-ortalamalar yöntemine göre en büyük avantajı bir nesnenin kesin olarak bir kümeye dâhil olması yerine belli bir üyelik değeri ile o kümeye bağlı olmasıdır. Çalışma sonucunda elde edilmiş üyelik matrislerinden bir kesit Çizelge 4.15'te yer almaktadır. Çizelge incelendiğinde her bir nesnenin tüm kümelere belli bir üyelik değeri ile bağlı olduğu görülmektedir. Örneğin 354 numaralı nesneye çizelgeden bakıldığında 1.kümeye 0.55229 üyelik değeri ile bağlıyken, 3. kümeye de 0.44675 üyelik değeri ile bağlıdır. Bu bize 354 numaralı nesnenin 1. kümenin özelliklerinin yanı sıra 3. kümenin özelliklerini de bünyesinde barındırdığını göstermektedir. Yani tek başına 354 numaralı nesne 1. kümeye aittir demek tamamıyla doğru olmayacaktır. Bunun yerine 354 numaralı nesne ağırlıklı olarak 1.kümenin

özeliklerini gösterirken 3. kümenin özelliklerine de sahiptir demek daha doğru olacaktır. BCO yöntemi işte bu noktada k-ortalamalar yöntemine göre daha anlamlı sonuçlar üretebilmektedir. K-ortalamalar yönteminde 354 numaralı nesne 1. kümeye aittir demekten başka şansımız bulunmamaktadır. Çünkü üyelik değeri gibi bir tanımlama k-ortalamalar yöntemi içerisinde bulunmamaktadır.

Çizelge 4.15 Bulanık c-ortalamalar yöntemi için üyelik matrisi

| **Nesneler** | **Bulanık Üyelik Matrisleri** | | | | **Ağırlıklı Bulunduğu Küme** |
|---|---|---|---|---|---|
| | **Küme-1** | **Küme-2** | **Küme-3** | **Küme-4** | |
| 346 | 0.71017 | 0.00068 | 0.28903 | 0.00012 | 1 |
| 347 | 0.70754 | 0.00069 | 0.29166 | 0.00012 | 1 |
| 348 | 0.66973 | 0.00073 | 0.32942 | 0.00012 | 1 |
| 349 | 0.66700 | 0.00073 | 0.33215 | 0.00012 | 1 |
| 350 | 0.64573 | 0.00075 | 0.35339 | 0.00013 | 1 |
| 351 | 0.60067 | 0.00079 | 0.39841 | 0.00013 | 1 |
| 352 | 0.57140 | 0.00081 | 0.42765 | 0.00013 | 1 |
| 353 | 0.56143 | 0.00082 | 0.43762 | 0.00013 | 1 |
| 354 | 0.55229 | 0.00082 | 0.44675 | 0.00014 | 1 |
| 355 | 0.45164 | 0.00086 | 0.54736 | 0.00014 | 3 |
| 356 | 0.41325 | 0.00086 | 0.58575 | 0.00014 | 3 |
| 357 | 0.33047 | 0.00084 | 0.66856 | 0.00013 | 3 |
| 358 | 0.29886 | 0.00083 | 0.70019 | 0.00013 | 3 |
| 359 | 0.25715 | 0.00080 | 0.74193 | 0.00013 | 3 |

Bulanık c-ortalamalar kümeleme yönteminin veri setine uygulanmasının ardından genel başarı notları için yapılan aykırı değer tespiti Şekil 4.47'de görülmektedir. C1, C2, C3 ve C4 küme numaralarını göstermek üzere 1. ve 2. küme için aykırı değer oluşmadığı ve en yoğun aykırı değerlerin küme 4 de oluştuğu görülmektedir. Toplam 1588 nesnenin 19 tanesi outlier olarak işaretlenmiştir.




Şekil 4.47 Bulanık c-ortalamalar yöntemi GNO özelliği için Box plot analizi

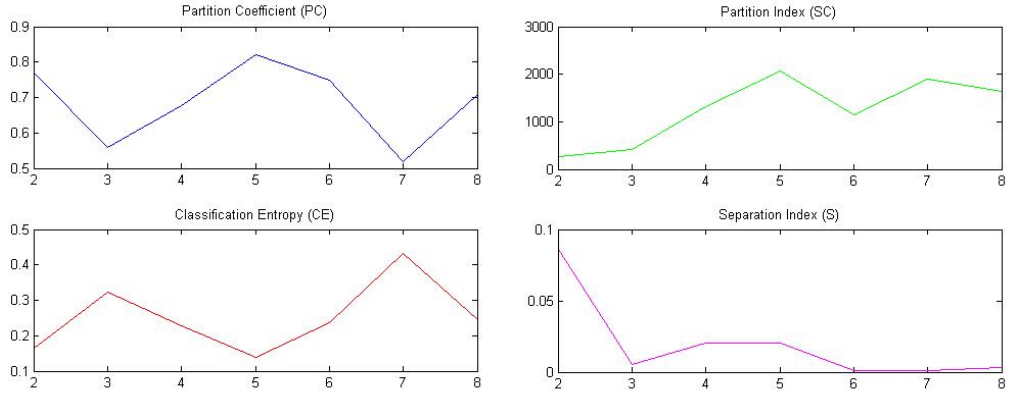
### 4.4.4 Gustafson Kessel Yönteminin Verilere Uygulanması

Küme sayısını belirlemek için yapılmış geçerlilik indeksleri sonuçları Çizelge 4.16'da yer almaktadır. GK yönteminde de küme sayısı 2 ile 8 aralığında seçilmiştir. GK yöntemi için de küme sayısı diğer yöntemlerde olduğu gibi 4 seçilmiştir.

Çizelge 4.16 Gustafson-Kessel yöntemi için doğruluk indeksi sonuçları

| **Doğruluk İndeksi** | **Küme Sayısı** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **PC** | 0.7701 | 0.5593 | 0.677 | 0.82 | 0.748 | 0.5184 | 0.71 |
| **CE** | 0.1628 | 0.3244 | 0.229 | 0.1398 | 0.238 | 0.4314 | 0.2454 |
| **SC** | 276 | 417 | 1319 | 2065 | 1153 | 1899 | 1633 |
| **S** | 0.0869 | 0.0054 | 0.0201 | 0.0206 | 0.001 | 0.0009 | 0.0033 |

Şekil 4.45'de, Çizelge 4.16'da yer alan doğruluk değerlerinin grafiksel gösterimi yer almaktadır.




Şekil 4.48 Gustafson-Kessel yöntemi için geçerlilik indeksleri

Bulanık yöntemlerden bir diğeri olan Gustafson-Kessel yöntemi de mühendislik fakültesi öğrencileri veri setine uygulanmıştır. Uygulamanın sonuçlarında elde edilen küme merkezi sonuçları aşağıdaki çizelgede yer almaktadır.

Çizelge 4.17'de yer alan sonuçlara bakıldığında 4.kümenin genel not ortalaması en iyi öğrencilerden oluştuğu görülmektedir. Bunun yanında 4. küme de yer alan öğrenciler çoğunlukla Makine, İnşaat ve Bilgisayar Mühendisliği ve 3.sınıf öğrencilerinden oluşmaktadır. Öğrenciler 1.kümede yoğun olarak Karadeniz ve Doğu Anadolu bölgesinden ve Anadolu liselerinden mezun olarak gelmişlerdir.

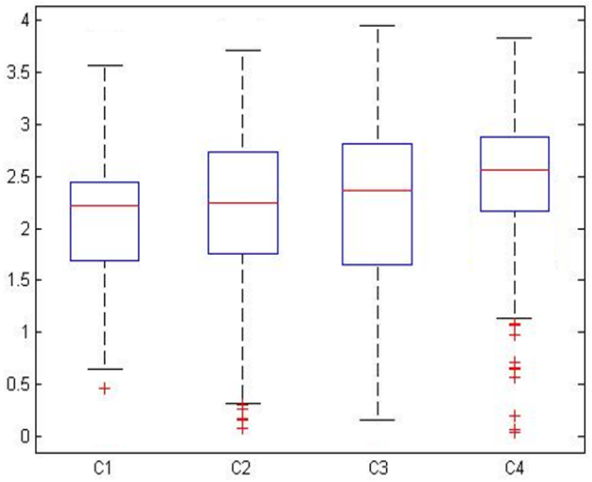
3. Kümede yer alan öğrencilerin ise yerleştirme sıralarının dört küme içerisinde en kötü olmasına rağmen genel başarı notları nispeten iyi durumdadır. Bölüm olarak tekstil ve elektronik-haberleşme mühendisliği öğrencileri çoğunluktadır. Bilgisayar mühendisliği

ikinci öğretim, inşaat ve çevre mühendisliği öğrencilerinin ağırlıkta olduğu üçüncü küme ise yerleştirme puanı anlamında ilk, yerleştirme sırası anlamında en kötü ikinci sırada yer almaktadır. Bu da yerleştirme puanları ile yerleştirme sıralarının paralellik göstermeyebileceği anlamına gelmektedir.

Çizelge 4.17 Gustafson-Kessel yöntemi sonucunda elde edilmiş küme merkezleri

| **Gustafson-Kessel** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Bölümü(%)** | İnşaat:35<br>Makine:26 | Bil. İ.Ö.:12<br>İnşaat:12<br>Çevre:12 | Tekstil:21<br>Elekt.Hab.:20 | Makine:16<br>İnşaat:15<br>Bilgisayar:14 |
| **Yer. Kon. Tipi** | 1.yerleştirme | 1.yerleştirme | 1.yerleştirme | 1.yerleştirme |
| **Sınıfı** | 3 | 2 | 1 | 3 |
| **Eğitim Yılı** | 5 | 2 | 2 | 3 |
| **Öğretim Turu** | 1 | 1 | 1 | 1 |
| **Disiplin Cezası** | HAYIR | HAYIR | HAYIR | HAYIR |
| **Cinsiyet Kodu(%)** | Erkek:88<br>Kadın:12 | Erkek:68<br>Kadın:32 | Erkek:65<br>Kadın:35 | Erkek:70<br>Kadın:30 |
| **Yaşı** | 23 | 20 | 21 | 22 |
| **Medeni D.** | BEKÂR | BEKÂR | BEKÂR | BEKÂR |
| **Türk Asıllı** | EVET | EVET | EVET | EVET |
| **Nüfus Bölgesi** | Karadeniz:2.4<br>Doğu A.:2.22<br>İç A.:1.22 | Karadeniz:2.6<br>Doğu A.:1.91<br>Marmara:0.87 | Karadeniz:1.78<br>Doğu A.:1.37<br>Marmara:1.24 | Karadeniz:2.30<br>Doğu A.:2.23<br>Marmara:0.87 |
| **Okul Turu(%)** | Resmi Lise:44<br>Anadolu L.:35 | Resmi Lise:42<br>Anadolu L.:45 | Resmi Lise:51<br>Anadolu L.:33 | Resmi Lise:31<br>Anadolu L.:42 |
| **Yer. Puanı** | 326 | 370 | 359 | 302 |
| **Yer. Sırası** | 68407 | 106222 | 126666 | 77026 |
| **Tercih Sırası** | 12 | 11 | 14 | 7 |
| **Genel Ort.** | **2.05** | **2.2** | **2.21** | **2.49** |
| **Katkı Kredisi** | Hayır | Hayır | Hayır | Hayır |
| **Öğrenim Kredi** | Hayır | Hayır | Evet | Evet |
| **Burs Durumu** | Hayır | Hayır | Hayır | Hayır |
| **Yüzde On(%)** | Evet:3<br>Hayır:97 | Evet:8<br>Hayır:92 | Evet:11<br>Hayır:89 | Evet:9<br>Hayır:91 |

Gustafson-Kessel kümeleme yönteminin veri setine uygulanmasının ardından genel başarı notları için yapılan aykırı değer tespiti Şekil 4.49'de görülmektedir. C1, C2, C3 ve C4 küme numaralarını göstermek üzere 3. küme için aykırı değer oluşmadığı ve en yoğun aykırı değerlerin küme 4'te oluştuğu görülmektedir. Toplam 1588 nesnenin 17 tanesi aykırı değer olarak işaretlenmiştir.




Şekil 4.49 Gustafson-Kessel yöntemi GNO özelliği için Box plot analizi

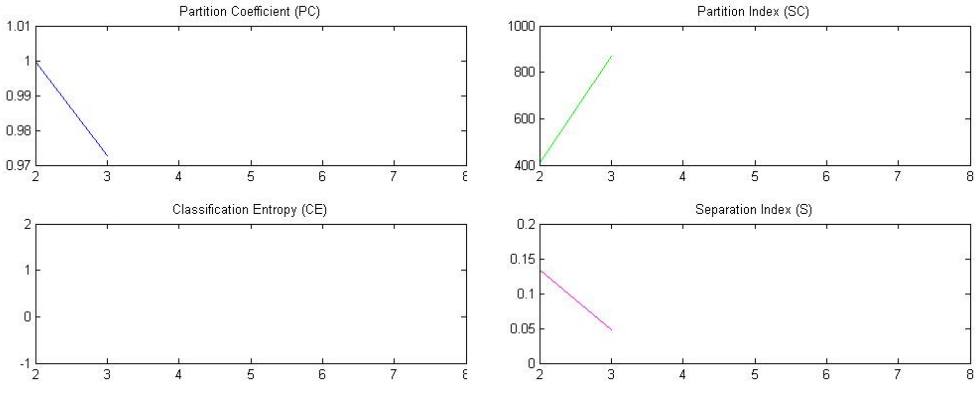
### 4.4.5 Gath-Geva Yönteminin Verilere Uygulanması

Gath-Geva yöntemi için doğruluk indekslerinden elde edilen sonuçlar Çizelge 4.18'de görülmektedir. En uygun küme sayısı iki olarak gözükmektedir. Ancak daha önceki yöntemlerde de söylendiği gibi küme sayısını 2 seçtiğimizde veri seti içerisinde mevcut olan değerli örüntüleri kaçırma ihtimali vardır. Bu sebepten küme sayısı dört olarak belirlenmiştir.

Çizelge 4.18 Gath-Geva yöntemi için doğruluk indeksi sonuçları

| **Doğruluk İndeksi** | **Küme Sayısı** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **PC** | 0.9997 | 0.9728 | 0.9913 | NaN | NaN | NaN | NaN |
| **CE** | 0.000234 | NaN | NaN | NaN | NaN | NaN | NaN |
| **SC** | 409 | 870 | NaN | NaN | NaN | NaN | NaN |
| **S** | 0.1336 | 0.0485 | NaN | NaN | NaN | NaN | NaN |

Çizelge 4.18'in grafiksel gösterimi Şekil 4.50'da yer almaktadır.




Şekil 4.50 Gath-Geva yöntemi için geçerlilik indeksleri

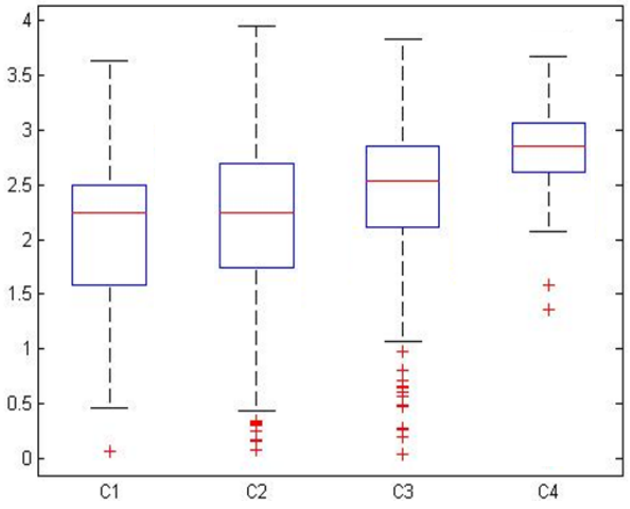
Bulanık yöntemlerden sonuncusu olan Gath-Geva yöntemi de mühendislik fakültesi öğrencileri veri setine uygulanmıştır. Elde edilen sonuçlar Gustafson-Kessel yöntemi ile benzerlik göstermektedir. Zaten Gath-Geva yöntemi Gustafson-Kessel yönteminin geliştirilmesi ile elde edilmiş bir yöntemdir. Çizelge 4.19'da yer alan sonuçlara bakıldığında 4.kümenin yerleştirme sırası ve genel not ortalaması en iyi öğrencilerden oluştuğu görülmektedir. Bunun yanında 4. küme de yer alan öğrenciler çoğunlukla Biyomedikal Mühendisliği ve 2.sınıf öğrencilerinden oluşmaktadır. Öğrenciler 4.kümede çoğunlukla Karadeniz bölgesinden gelmişlerdir.

En kötü genel not ortalamasına sahip 1. kümenin inşaat ve makine mühendisliği öğrencileri ağırlıklı olduğu görülmektedir. 3.küme ise GNO bakımından en iyi ikinci kümedir.

Çizelge 4.19 Gath-Geva yöntemi sonucunda elde edilmiş küme merkezleri

| **Gath-Geva** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Bölümü(%)** | İnşaat:40<br>Makine:26 | Bilg. İ.Ö.:12<br>Mak. İ.Ö.:11<br>İnşaat:11 | Makine:16<br>İnşaat:14 | Biyomedikal:82<br>Elekt.Hab.:10 |
| **Yer. Kon.Tipi** | 1.yerleştirme | 1.yerleştirme | 1.yerleştirme | 1.yerleştirme |
| **Sınıfı** | 3 | 1 | 3 | 2 |
| **Eğitim Yılı** | 5 | 1 | 3 | 2 |
| **Öğretim Türü** | 1 | 1 | 1 | 1 |
| **Disiplin Cezası** | HAYIR | HAYIR | HAYIR | HAYIR |
| **Cinsiyet Kodu(%)** | Erkek:91<br>Kadın:9 | Erkek:68<br>Kadın:32 | Erkek:69<br>Kadın:31 | Erkek:43<br>Kadın:57 |
| **Yaşı** | 25 | 20 | 22 | 21 |
| **Medeni Durumu** | BEKÂR | BEKÂR | BEKÂR | BEKÂR |
| **Türk Asıllı** | EVET | EVET | EVET | EVET |
| **Nüfus Bölgesi** | Doğu A.:3.77<br>Karadeniz:1.76<br>Marmara:0.84 | Karadeniz:2.4<br>Doğu A.:1.76<br>Marmara:0.94 | Karadeniz:2.3<br>Doğu A.:2.06<br>Marmara:0.94 | Karadeniz:2.82<br>G.D.Anadolu:1.54<br>Doğu A.:1.29 |
| **Okul Turu(%)** | Resmi Lise:46<br>Anadolu L.: 31 | Resmi Lise:46<br>Anadolu L.:41 | Resmi Lise:31<br>Anadolu L.:42 | Resmi Lise:23<br>Anadolu L.:60 |
| **Yerleştirme Puanı** | 307 | 365 | 300 | 459 |
| **Yerleştirme Sırası** | 64199 | 116139 | 81133 | 36804 |
| **Tercih Sırası** | 10 | 12 | 7 | 15 |
| **Genel Ortalama** | **2.05** | **2.18** | **2.44** | **2.82** |
| **Katkı Kredisi** | Hayır | Hayır | Hayır | Evet |
| **Öğrenim Kredi** | Hayır | Hayır | Hayır | Hayır |
| **Burs Durumu** | Hayır | Hayır | Hayır | Hayır |
| **Yüzde On(%)** | Evet:2<br>Hayır:98 | Evet:9<br>Hayır:91 | Evet:9<br>Hayır:91 | Evet:14<br>Hayır:86 |

Gath-Geva kümeleme yönteminin veri setine uygulanmasının ardından genel başarı notları için yapılan aykırı değer tespiti Şekil 4.51'de görülmektedir. C1, C2, C3 ve C4 küme numaralarını göstermek üzere en yoğun aykırı değerlerin küme 2 ve küme 3'te oluştuğu görülmektedir. Toplam 1588 nesnenin 28 tanesi aykırı değer olarak işaretlenmiştir.




Şekil 4.51 Gath-Geva yöntemi GNO özelliği için Box plot analizi

### 4.4.6 Geçerlilik Değerleri ile Yöntemlerin Karşılaştırılması

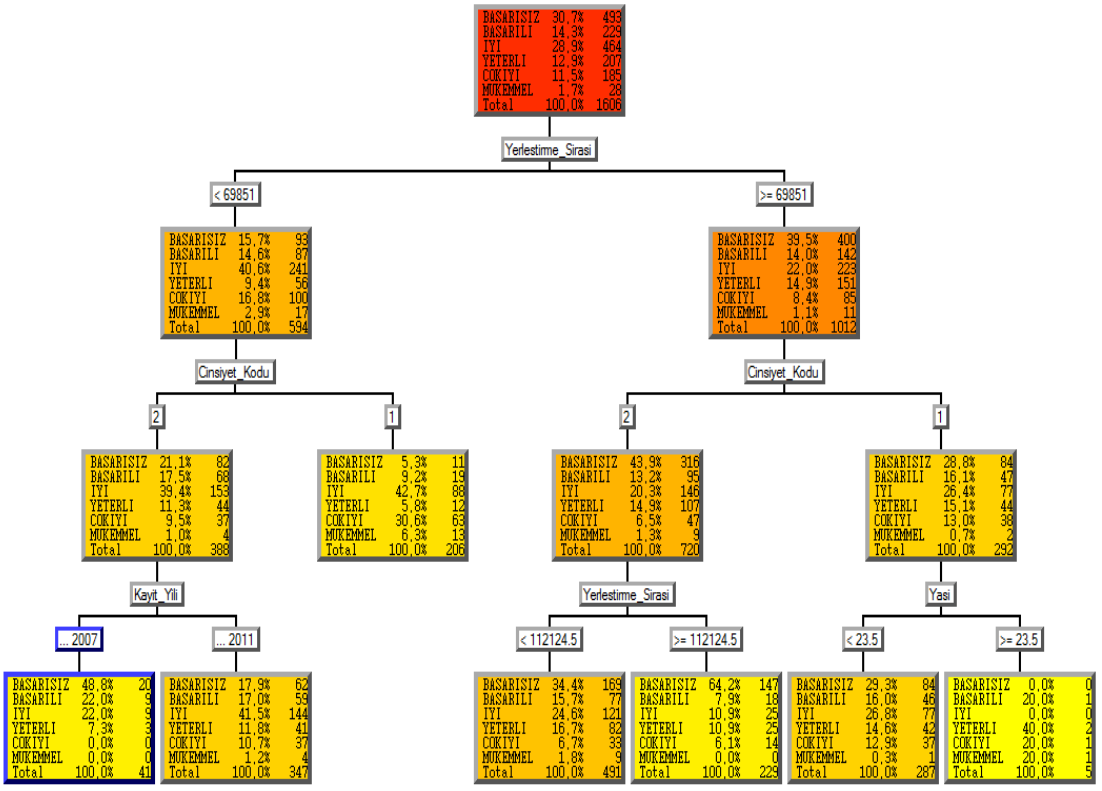
Mevcut veri setimize 5 farklı kümeleme yöntemi uygulanmıştır. Bu yöntemler sonucunda elde edilen sonuçlar yukarıda anlatılmıştır. Çizelge 4.20'de ise doğruluk indeksleri ile elde edilmiş, kullanılan bu yöntemlerin birbirlerine göre üstünlüklerini inceleyebileceğimiz değerler yer almaktadır. Bu değerler her bir yöntem için küme sayısının dört olarak seçilmesi ile elde edilmiştir. Çizelge incelendiğinde PC yöntemine göre en başarılı kümeleme algoritmasının Gath-Geva olduğu görülmektedir. Ayrıca bulanık kümeleme yöntemlerinin katı kümeleme yöntemlerine göre daha başarılı oldukları da Çizelge 4.20'de açıkça görülmektedir.

Çizelge 4.20 Doğruluk indeksi yöntemlerinin kümeleme algoritmaları için başarı sonuçları

| Yöntem | PC | CE | SC | S |
|---|---|---|---|---|
| **K-ortalamalar** | 1 | NaN | 1810 | 0.0294 |
| **K-medoids** | 1 | NaN | 1574 | 0.0285 |
| **BCO** | 0.4587 | 0.443 | 739 | 0.0079 |
| **GK** | 0.677 | 0.229 | 1319 | 0.0201 |
| **GG** | 0.9913 | NaN | NaN | NaN |

### 4.5 Karar Ağacı Yönteminin Verilere Uygulanması

Verilerle ilgili çeşitli ön işlem aşamaları gerçekleştirildikten sonra son olarak verilerin sınıflandırılması işlemi yapılmıştır. Verileri sınıflamak için SAS Enterprise Miner veri madenciliği aracı kullanılmıştır. Sınıflandırma için Karar Ağacı(Decision Tree) yöntemi kullanılmıştır. Verilerin %67'si öğrenme, %33'ü test için kullanılmıştır. Elde edilen karar ağacı aşağıdaki şekilde görülmektedir.




Şekil 4.52 Karar Ağacı

Karar ağacı incelendiğinde şu sonuçlar elde edilmektedir; Üniversiteye yerleştirme sırası kök düğüm özelliği kazanmıştır. Öğrenciler yerleştirme sırası 69851'den büyük ve küçük olanlar şeklinde ikiye ayrılmıştır. Ardından öğrencilerin cinsiyeti bir sonraki düğüm olmuştur. Kayıt yılı, yerleştirme sırası ve yaşı özellikleri de ağaçta yer alan diğer düğümlerdir. Karar ağacından elde edilen karar kurallarından bazıları ise şöyledir;

Eğer Yerlestirme_Sirasi > 69455 ve
Okul_Turu_Kodu = 10 ve
Ogretim_Turu_Kodu = 2:
sonuç BASARISIZ

Eğer Yerlestirme_Sirasi <= 44020 ve
Cinsiyet_Kodu = 1 ve Nufus_Bolgesi = 6:
sonuc IYI

Eğer Yerlestirme_Sirasi > 69455 ve
Okul_Turu_Kodu = 12 ve
Bolum_Kodu = 6:
sonuç YETERLI

Eğer Yerlestirme_Sirasi > 69455 ve
Okul_Turu_Kodu = 13 ve
Cinsiyet_Kodu = 2 ve
Sinifi_Kodu = 1:
sonuç BASARISIZ

Karar ağacının uygulanmasının ardından veri setinin % 68'inin doğru sınıflandırıldığı Çizelge 4.16 da görülmektedir. Yine aynı çizelgeden her bir sınıf için doğru-pozitif, yanlış pozitif ve tahmin değerlerini görmek mümkündür. Çizelge 4.17 de ise sınıf karışıklık matrisi yer almaktadır.

Çizelge 4.21 Karar ağacı sınıflandırma istatistikleri

| | | | | |
|---|---|---|---|---|
| Doğru Sınıflandırılmış Örnekler | | | 1079 | 68.3064 % |
| Yanlış Sınıflandırılmış Örnekler | | | 509 | 31.6936 % |
| **TP** | **FP** | **Hassasiyet** | **Sınıf** | |
| 0.899 | 0.201 | 0.664 | BASARISIZ | |
| 0.485 | 0.035 | 0.698 | BASARILI | |
| 0.759 | 0.13 | 0.703 | IYI | |
| 0.473 | 0.039 | 0.641 | YETERLI | |
| 0.481 | 0.021 | 0.748 | COKIYI | |
| 0.143 | 0.002 | 0.571 | MUKEMMEL | |
| W. Avg. 0.683 | 0.112 | 0.685 | | |

Çizelge 4.22 Karar ağacı sınıf karışıklık matrisi

| **Sınıf Karışıklık Matrisi** | | | | | | |
|---|---|---|---|---|---|---|
| **a** | **B** | **c** | **d** | **e** | **f** | |
| 443 | 7 | 26 | 13 | 4 | 0 | a = BASARISIZ |
| 52 | 111 | 44 | 13 | 8 | 1 | b = BASARILI |
| 64 | 24 | 352 | 14 | 9 | 1 | c = IYI |
| 68 | 9 | 28 | 98 | 3 | 1 | d = YETERLI |
| 39 | 6 | 42 | 9 | 89 | 0 | e = COKIYI |

### 4.6 Öğrenci Anketlerinin Değerlendirilmesi

Tezin bundan önceki bölümlerinde N.K.Ü. Çorlu Mühendislik Fakültesi öğrenci bilgi sistemi üzerinde kayıtlı olan lisans öğrencilerinin üniversitedeki başarı durumları ile üniversite sınavında almış oldukları başarı puanları ve yerleştirme sıraları baz alınarak profillerinin incelenmesi işlemleri gerçekleştirilmiştir. Tezin bu kısmında ise mevcut veri setine ek olarak N.K.Ü. Çorlu Mühendislik Fakültesi Bilgisayar Mühendisliği bölümü öğrencilerine bir anket yapılmıştır. Ankette öğrencilerin anne-baba eğitim durumları ve meslekleri gibi sorular yöneltilmiştir. Burdaki amaç öğrencinin ailesinin başarısına etkisi nedir sorusuna yanıt alabilmektir.

Yapılan anket sonucunda alınan cevaplar üzerinden bir veri seti oluşturulmuştur. Elde edilen bu veri setine tezin önceki bölümlerinde bahsedilen kümeleme yöntemleri uygulanmıştır. Küme sayısının yapılan analizler neticesinde dört seçilmesi uygun bulunmuştur.

Oluşturulan veri setinin özellikleri, açıklamaları ve değerleri Çizelge 4.23'de yer almaktadır.

Çizelge 4.23 Anket veri seti

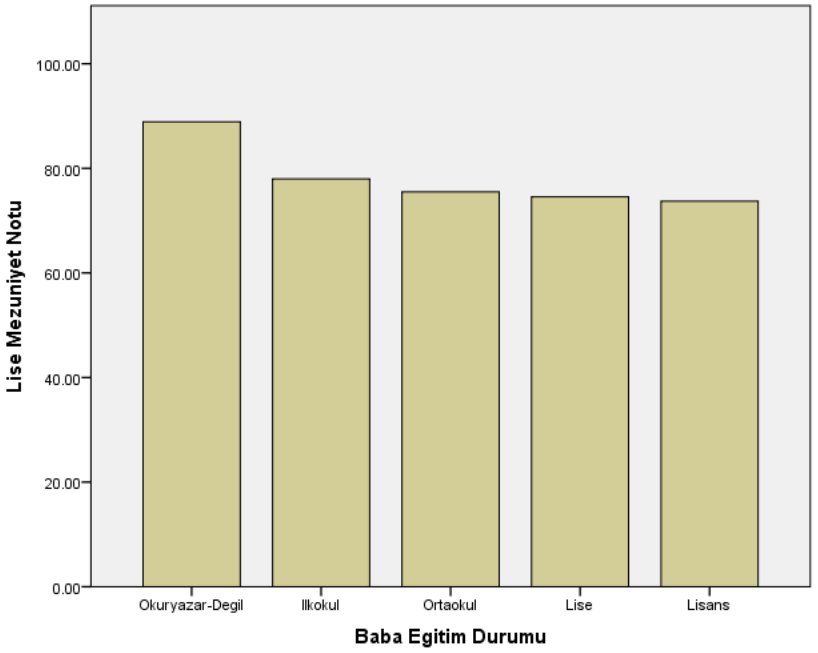
| **Özellik** | **Açıklaması** | **Değeri** |
|---|---|---|
| Yerleştiği Kontenjan Tipi | Üniversiteyi kazanma şekli | Kategorik(1.Yerleştirme, DGS, Yatay Geçiş vb.) |
| Sınıfı | Öğrencinin öğrenimine devam ettiği sınıf bilgisi | Nümerik |
| Öğretim Türü | Örgün öğretim-İkinci öğretim bilgisi | Kategorik(Normal, Örgün) |
| Cinsiyet | Öğrencinin cinsiyeti | Kategorik(Kadın, Erkek) |
| Yaşı | Öğrencinin yaşı | Nümerik |
| Medeni Durum | Öğrencinin medeni durumu | Kategorik(Evli, Bekar) |
| Nüfus Bölgesi | Öğrencinin nüfusa kayıtlı olduğu bölge | Kategorik(Marmara, Ege, Akdeniz vb.) |
| Okul Türü | Öğrencinin ortaöğretim kurum tipi | Kategorik(Resmi lise, teknik lise, Anadolu lisesi vb.) |
| Yerleştirme Puanı | Üniversiteye yerleştirme puanı | Nümerik |
| Yerleştirme Sırası | Üniversiteye yerleştirme sırası | Nümerik |
| Tercih Sırası | Öğrencinin kazandığı bölümü tercih sırası | Nümerik |
| Genel Ortalama | Üniversitedeki genel not ortalaması | Nümerik |

Çizelge 4.23 (devamı)

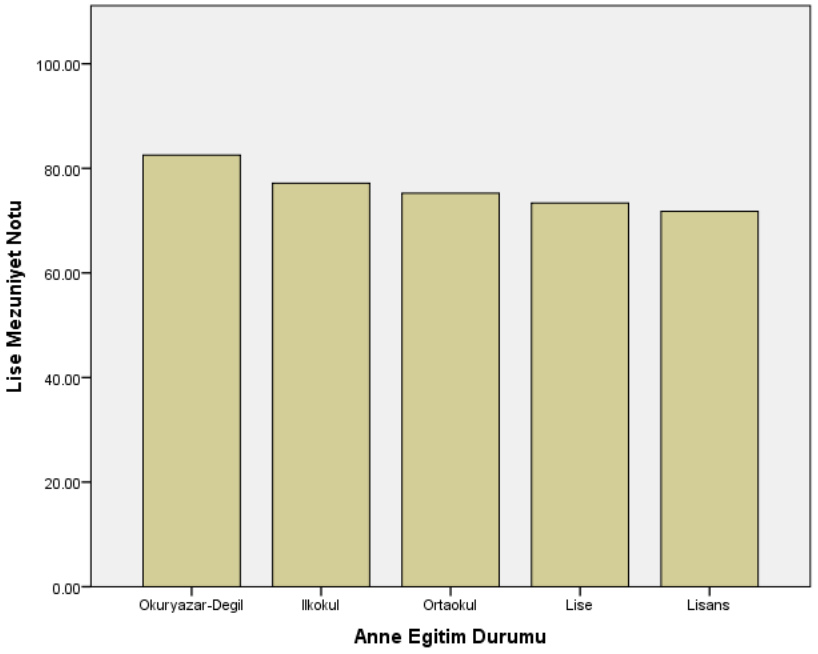
| Özellik | Açıklaması | Değeri |
|---|---|---|
| Katkı Kredisi | Katkı kredisi bilgisi | Kategorik(Evet, Hayır) |
| Öğrenim Kredisi | Öğrenim kredisi bilgisi | Kategorik(Evet, Hayır) |
| Kyk Burs | Burs durum bilgisi | Kategorik(Evet, Hayır) |
| Yüzde On | Öğrencinin yüzde onluk dilime girip girmediği | Kategorik(Evet, Hayır) |
| Baba Eğitim Durumu | Baba eğitim düzeyi | Kategorik(İlkokul, Ortaokul, lise vb.) |
| Anne Eğitim Durumu | Anne eğitim düzeyi | Kategorik(İlkokul, Ortaokul, lise vb.) |
| Baba Meslek | Baba mesleği | Kategorik(Memur, Serbest Meslek, İşçi vb.) |
| Anne Meslek | Anne mesleği | Kategorik(Memur, Ev Hanımı, İşçi vb.) |
| Kardeş Sayısı | Kardeş sayısı | Nümerik |
| Okuyan Kardeş Sayısı | Okumakta olan kardeş sayısı | Nümerik |
| Lise Mezuniyet Notu | Lise mezuniyet ort. | Nümerik |
| GSM | Cep tel. operatörü | Nümerik |

### 4.6.1 Anket Veri Setinin Analiz Edilmesi

Anket veri setinde yer alan özellikler Çizelge 4.23'de görülmektedir. Tezin ilk kısmında kullanılan mühendislik fakültesi öğrenci veri setinden farklı olarak anket veri setinde anne-baba eğitim durumu, anne-baba meslek bilgisi, kardeş-okuyan kardeş sayıları, lise mezuniyet notları ve GSM operatörü bilgileri yer almaktadır. Anket veri seti üzerinde kümeleme işlemine başlamadan önce bu özelliklerin nasıl dağılım gösterdiklerinin analiz edilmesi önem taşımaktadır. Bu doğrultuda Şekil 4.53 ve Şekil 4.54'de anne-baba eğitim durumu ile öğrencilerin lise mezuniyetleri arasındaki ilişki görülmektedir. Şekillere bakıldığında eğitim durumu aşağı seviyelerde olan ailelerin çocuklarının lisedeki başarılarının daha yüksek olduğu görülmektedir. Bu şekillerden eğitim düzeyi düşük ailelerin çocuklarının başarılı olma eğilimlerinin eğitim düzeyi yüksek olan ailelerin çocuklarının başarılı olma eğilimlerinden daha fazla olduğu sonucuna ulaşılabilir.

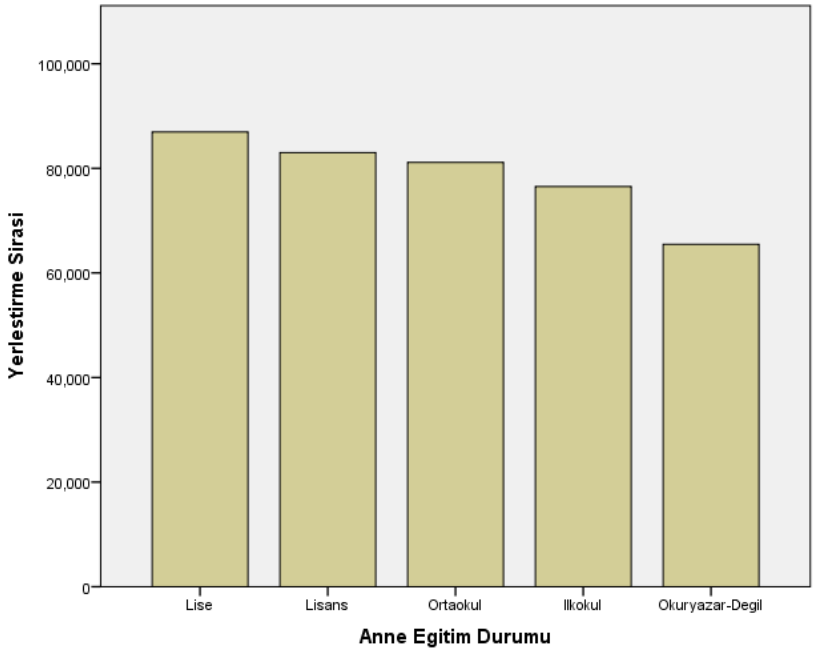



Şekil 4.53 Baba Eğitim Durumu-Öğrencinin Lise Mezuniyet Notu İlişkisi

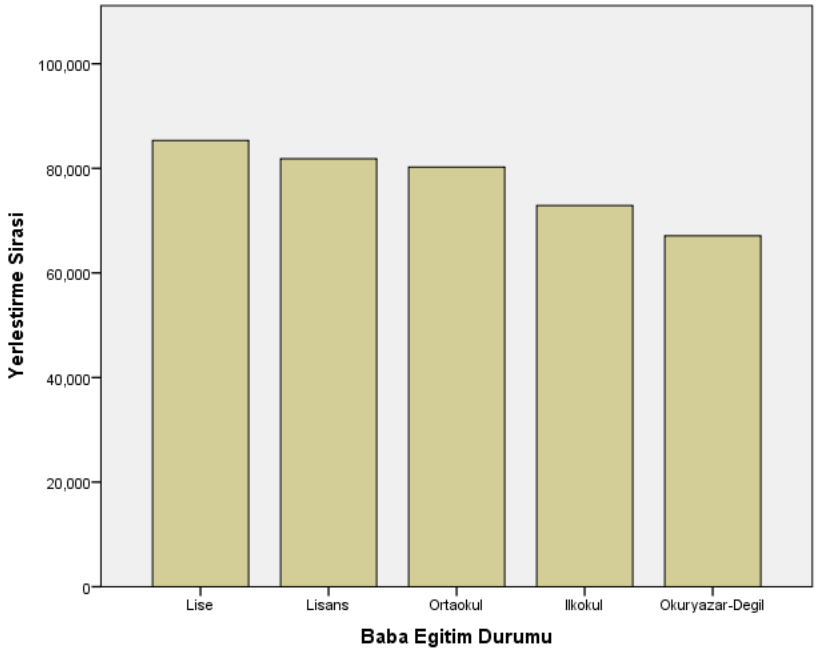



Şekil 4.54 Anne Eğitim Durumu- Öğrencinin Lise Mezuniyet Notu İlişkisi

Şekil 4.55 ve 4.56'da ise anne-baba eğitim durumları ile öğrencilerin üniversite sınavında başarılı olma durumları görülmektedir. Bir önceki şekilde olduğu gibi burada da eğitim seviyesi düşük olan ailelerin çocuklarının diğer ailelerin çocuklarına göre üniversite sınavlarında daha başarılı olduklarını görmek mümkündür.



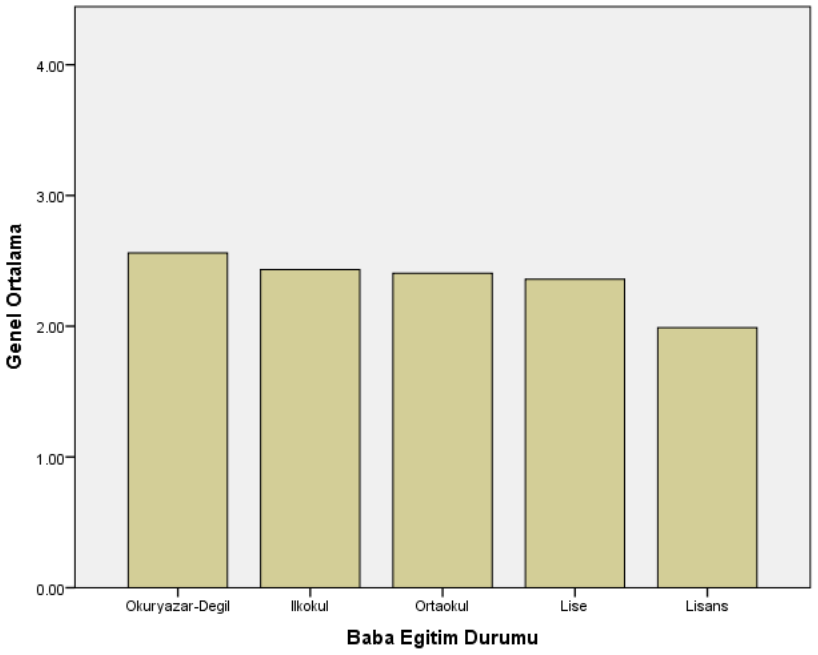

Şekil 4.55 Anne Eğitim Durumu- Öğrencinin Yerleştirme Sırası İlişkisi



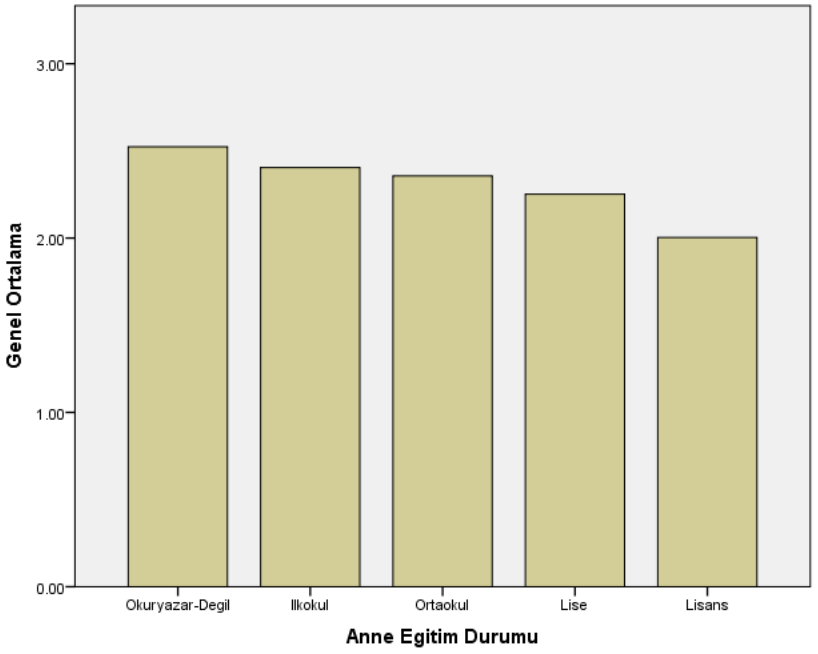

Şekil 4.56 Baba Eğitim Durumu- Öğrencinin Yerleştirme Sırası İlişkisi

Şekil 4.57 ve 4.58'de ise anne-baba eğitim durumları ile öğrencilerin üniversite öğrenimlerindeki başarıları arasında nasıl bir ilişki olduğu görülmektedir. Bu grafik ışığında eğitim düzeyi düşük ailelerin çocuklarının, üniversite öğrenimlerinde, diğer ailelerin çocuklarından daha başarılı olduğunu söylemek doğru olacaktır. Yine bu grafiğe bakarak kendileri yeteri kadar eğitim alamamış aileler, çocuklarına bu konuda

gereken desteği ve azmi vermek konusunda, diğer ailelerden daha isteklidirler, sonucuna ulaşmak yanlış olmayacaktır.

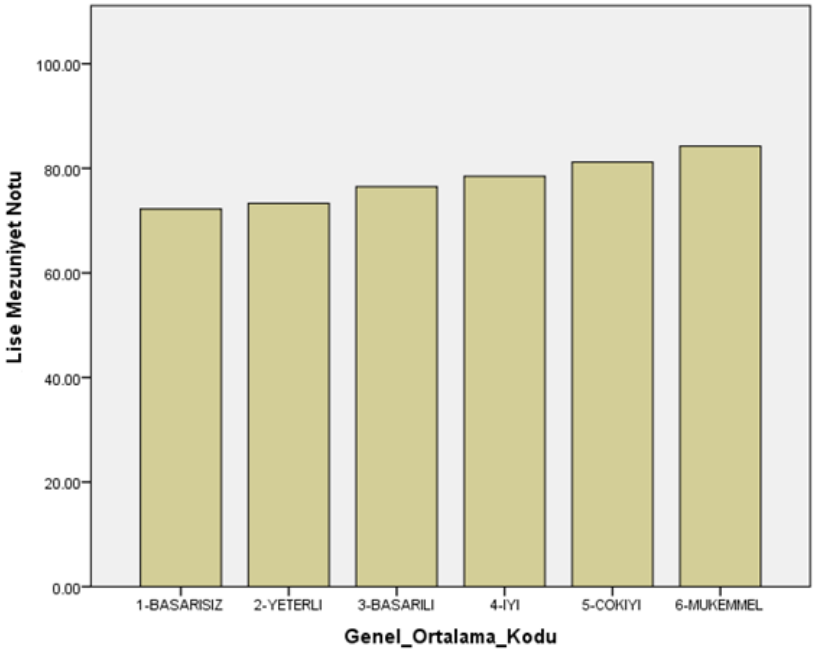



Şekil 4.57 Baba Eğitim Durumu- Öğrencinin Genel Not Ortalaması İlişkisi



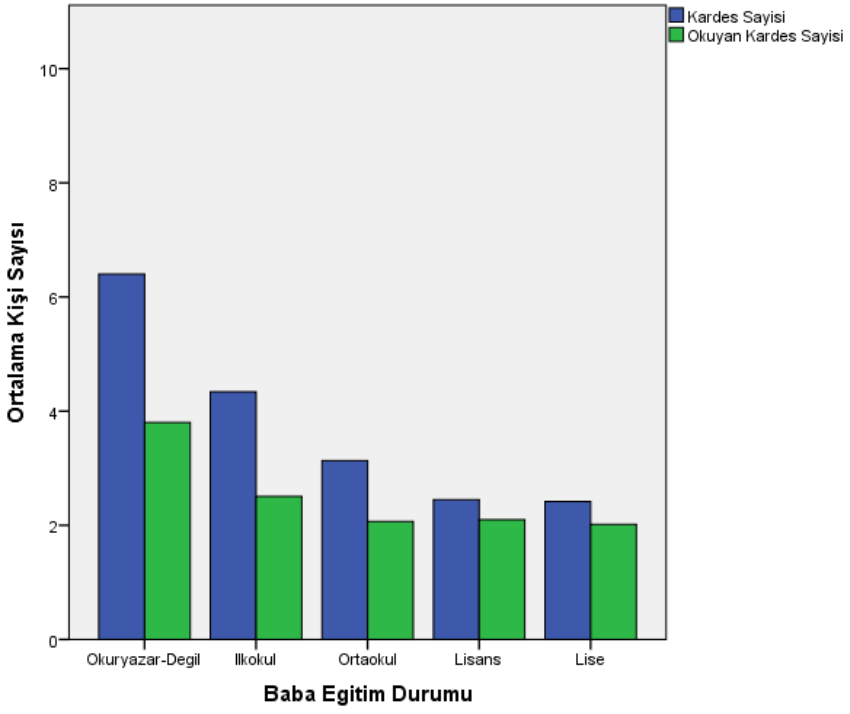

Şekil 4.58 Anne Eğitim Durumu- Öğrencinin Genel Not Ortalaması İlişkisi

Şekil 4.59'da ise öğrencilerin lise mezuniyet notları ile üniversite genel not ortalamaları arasındaki ilişki görülmektedir. Şekil incelendiğinde lise öğrenimlerinde başarılı olan öğrenciler bu başarılarını üniversite öğrenimlerinde de devam ettirme eğilimindedirler.

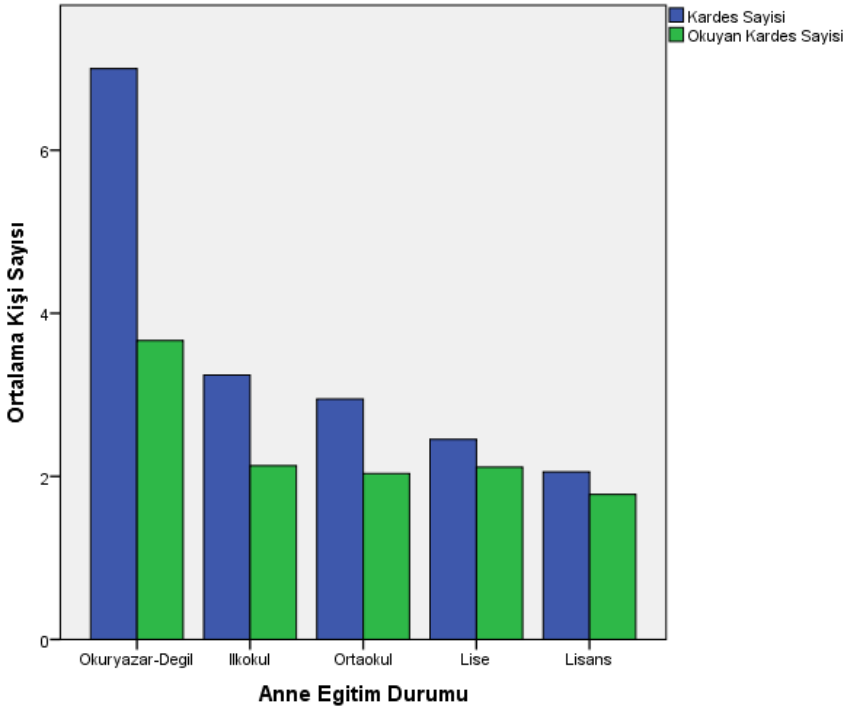



Şekil 4.59 Lise Mezuniyet Notu- Genel Not Ortalaması İlişkisi

Şekil 4.60 ve 4.61'de anne-babanın eğitim durumları, kardeş ve okuyan kardeş sayılarının yer aldığı grafikler görülmektedir. Grafikler incelendiğinde eğitim durumu düşük ailelerin çocuk sayılarının, eğitim durumu nispeten daha iyi olan ailelere göre daha fazla olduğu görülmektedir. Bu da eğitim seviyesi düşük ailelerin aile planlaması konusunda bilgisiz olduklarını göstermektedir. Bunun yanında kardeş sayısı fazla olan ailelerde okuyan kardeş sayısının da fazla olduğu görülmektedir.

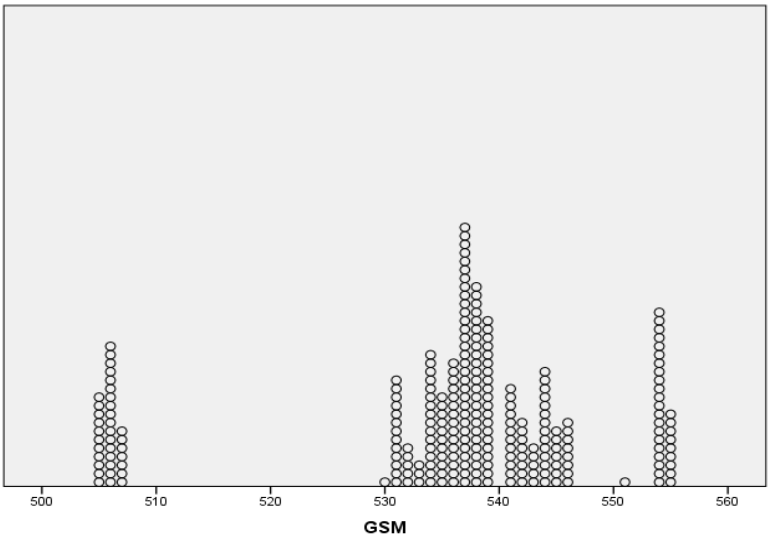



Şekil 4.60 Baba Eğitim Durumu-Kardeş-Okuyan Kardeş Sayısı İlişkisi




Şekil 4.61 Anne Eğitim Durumu-Kardeş-Okuyan Kardeş Sayısı İlişkisi

Şekil 4.62'de öğrencilerin kullandıkları cep telefonu operatörlerinin dağılımları görülmektedir. Grafiğe bakıldığında öğrencilerin ağırlıklı olarak 530 ve 540'lı numaraları kullandıklarını görmek mümkündür.




Şekil 4.62 GSM Operatörleri Dağılımı

### 4.6.2 Anket Verilerinin K-Ortalamalar Yöntemi İle Kümelenmesi

Verilere k-ortalamalar yöntemi uygulanmış ve Çizelge 4.24'de yer alan sonuçlar elde edilmiştir. Yöntemin uygulanmasının ardından dört farklı küme elde edilmiştir. Bu kümelere bakıldığında üniversiteye giriş sınavında en başarılı olan öğrencilerin yer

aldığı küme dördüncü kümedir. Üniversite sınavında başarılı olan bu öğrencilerin lise mezuniyet notlarına bakıldığında aynı başarıyı lisede de gösterdikleri görülmektedir. Aynı zamanda bu öğrenciler üniversite sıralarında da başarılı olma eğilimindedirler. Başarılı olan bu öğrencilerin ağırlıklı olarak Anadolu liselerinden mezun oldukları görülmektedir. En başarısız öğrencilerin yer aldığı bir numaralı kümeye bakıldığında öğrencilerin hem lise sıralarında, hem üniversite sınavında hem de üniversite öğrenciliklerinde bu başarısızlıklarını devam ettirdikleri görülmektedir. Başarısız olan öğrencilerin yer aldığı bu kümede anne-baba eğitim durumlarına bakıldığında, eğitim durumlarının başarılı olan kümelerdeki öğrencilerin ebeveynlerinin eğitim durumlarından daha yüksek olduğu görülmektedir. Buradan eğitim durumları iyi olan ailelerin çocuklarının diğer ailelerin çocuklarına göre daha başarılı olmadıkları sonucuna ulaşmak mümkündür. Bir numaralı kümede yer alan öğrencilerin babalarının %29, annelerinin ise %14'ünün lisans mezunu olduğuna da dikkat çekmekte fayda vardır. Yine çizelge incelendiğinde annesi ev hanımı olan öğrencilerin daha başarılı oldukları görülmektedir.

Çizelge 4.24 K-ortalamalar yöntemi ile elde edilmiş küme merkezleri

| **k-ort.** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Yer. Kon. Tipi** | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme |
| **Sınıfı** | 1 | 2 | 2 | 1 |
| **Öğretim Türü** | İkinci Ö. | Örgün | Örgün | Örgün |
| **Cinsiyeti(%)** | Erkek:75<br>Kadın:25 | Erkek:56<br>Kadın:44 | Erkek:76<br>Kadın:24 | Erkek:61<br>Kadın:39 |
| **Yaşı** | 20.2 | 21.1 | 22.8 | 19.7 |
| **Medeni Hal** | Bekâr | Bekâr | Bekâr | Bekâr |
| **Nüfus Bölgesi (Katsayı oranı)** | Karadeniz:2.88<br>İç A.:1.02<br>Doğu A.:1.00 | Karadeniz:2.25<br>Doğu A.:1.85<br>G.DoğuA.:1.08 | Doğu A.:4.84<br>G. Doğu A.:4.11 | Karadeniz:2.1<br>Doğu A.:1.66<br>İç A.:1.11 |
| **Okul Türü(%)** | Resmi Lise: 42<br>Anadolu L.: 41 | Resmi Lise : 33<br>Anadolu L.: 53 | Resmi Lise:86<br>Anadolu L.:5 | Resmi L.:41<br>Anadolu L.:44 |
| **Yer. Puanı** | 366.7 | 367.6 | 389.3 | 392.3 |
| **Yer. Sırası** | 104281 | 63506 | 69259 | 77176 |
| **Tercih Sırası** | 9 | 10 | 11 | 9 |
| **Genel Ort.** | **1.86** | **2.51** | **2.51** | **2.56** |
| **Katkı Kredisi(%)** | Evet:11<br>Hayır:89 | Evet:35<br>Hayır:65 | Evet:43<br>Hayır:57 | Evet:36<br>Hayır:64 |

Çizelge 4.24 (devamı)

| k-ort. | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Öğrenim Kredisi(%)** | Evet:46<br>Hayır:54 | Evet:44<br>Hayır:56 | Evet:33<br>Hayır:67 | Evet:34<br>Hayır:66 |
| **Burs(%)** | Evet:15<br>Hayır:85 | Evet:27<br>Hayır:73 | Evet:43<br>Hayır:57 | Evet:30<br>Hayır:70 |
| **Yüzde On(%)** | Evet:3<br>Hayır:97 | Evet:12<br>Hayır:88 | Evet:10<br>Hayır:90 | Evet:15<br>Hayır:85 |
| **Baba Eğitim Durumu(%)** | Lisans: 29<br>Lise: 27<br>İlköğretim: 22<br>İlkokul: 15 | İlkokul: 4<br>İlköğretim: 24<br>Lise: 23 | İlkokul: 57<br>İlköğretim: 14<br>Okuryazar Değil: 14 | Lise: 33<br>İlköğretim: 23<br>Lisans: 18<br>İlkokul: 16 |
| **Anne Eğitim Durumu(%)** | İlkokul: 33<br>Lise: 27<br>İlköğretim: 19<br>Lisans: 14 | İlkokul: 51<br>İlköğretim: 19<br>Lise: 17<br>Okuryazar Değil: 9 | OkuryazarDeğil: 52<br>İlkokul 29<br>İlköğretim: 14 | İlkokul: 31<br>İlköğretim: 31<br>Lise: 20<br>Lisans: 11 |
| **Baba Meslek(%)** | Emekli: 25<br>Serbest M.: 16<br>Memur: 11<br>İşçi: 10 | Emekli: 22<br>Serbest M.: 16<br>İşçi: 13<br>Memur: 7 | Çiftçi: 33<br>İşçi: 14<br>Emekli: 10<br>Serbest M.: 10 | Emekli: 25<br>Serbest M.: 21<br>Memur: 10<br>İşçi: 8 |
| **Anne Meslek(%)** | Ev Hanımı: 67<br>Emekli: 14<br>Öğretmen: 5 | Ev Hanımı: 84<br>Emekli: 7<br>İşçi: 5 | Ev Hanımı: 95 | Ev Hanımı: 72<br>Emekli: 11<br>Öğretmen: 5 |
| **Kardeş Sayısı** | 2.7 | 3.12 | 7.7 | 2.63 |
| **Okuyan Kardeş Say.** | 2.075 | 2.1 | 4.85 | 1.65 |
| **Lise Mez. Notu** | 67.9 | 79.568 | 82.48 | 77.9 |
| **GSM(%)** | 530'lu: 36<br>500'lı: 35<br>540'lü: 29 | 500'lu: 48<br>530'lı: 29<br>540'lü: 23 | 530'lu: 37<br>500'lı: 33<br>540'lü: 30 | 500'lu: 41<br>530'lı: 36<br>500'lü: 23 |

### 4.6.3 Anket Verilerinin K-Medoids Yöntemi ile Kümelenmesi

Verilere k-medoids yönteminin uygulanmasının ardından elde edilen küme merkezleri Çizelge 4.25'de yer almaktadır. Çizelge incelendiğinde k-ortalamalar yöntemi için söylenenlerin k-medoids içinde geçerli olduğu görülmektedir. Bu kümelere bakıldığında üniversiteye giriş sınavında en başarılı olan öğrencilerin yer aldığı küme dördüncü kümedir. Aynı şekilde üç numaralı kümede bir numaralı kümeye benzer şekilde üniversite sınavında başarılı olan öğrencileri kapsamaktadır. Üniversite sınavında başarılı olan bu öğrencilerin, lise mezuniyet notlarının ise dört numaralı küme için 81,

üç numaralı küme için 77 olduğu görülmektedir. Ve diğer kümelere bakıldığında bu değerlerin diğerlerinden daha yüksek olduğu görülmektedir. Buradan yola çıkarak üniversite sınavındaki başarıyı lisede de gösterme eğiliminde oldukları anlaşılmaktadır. Aynı zamanda bu öğrencilerin üniversite genel not ortalamaları da diğer kümelerde yer alan öğrencilerinkinden daha yüksektir. En başarısız öğrencilerin yer aldığı bir numaralı kümeye bakıldığında öğrencilerin hem lise sıralarında, hem üniversite sınavında hem de üniversite öğrenciliklerinde bu başarısızlıklarını devam ettirdikleri görülmektedir. Başarısız olan öğrencilerin yer aldığı bu kümede anne-baba eğitim durumlarına bakıldığında, eğitim durumlarının başarılı olan kümelerdeki öğrencilerin ebeveynlerinin eğitim durumlarından daha yüksek olduğu görülmektedir. Buradan eğitim durumları iyi olan ailelerin çocuklarının diğer ailelerin çocuklarına göre daha başarılı olmadıkları sonucuna ulaşmak mümkündür. Bir numaralı kümede yer alan öğrencilerin babalarının %28'inin, annelerinin ise %11'inin lisans mezunu olduğu ancak genel başarı notu anlamında en başarısız öğrencilerin yer aldığı küme olduğuna da dikkat çekmekte fayda vardır. Okuyan kardeş sayısı en yüksek olan kümenin başarı oranının da en yüksek küme olduğu görülmektedir. Ayrıca dört kümenin dördünde de öğrenciler cep telefonu operatörü olarak 530'lu numaraları tercih etmektedirler.

Çizelge 4.25 k-medoids yöntemi ile elde edilmiş küme merkezleri

| **k-medoids** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Yer. Kon. Tipi** | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme |
| **Sınıfı** | 1 | 1 | 2 | 3 |
| **Öğretim Türü** | İkinci Ö. | İkinci Ö. | Örgün | Örgün |
| **Cinsiyeti(%)** | Erkek:70<br>Kadın:30 | Erkek:69<br>Kadın:31 | Erkek:55<br>Kadın:45 | Erkek:60<br>Kadın:40 |
| **Yaşı** | 20.3 | 20.3 | 20.8 | 21.7 |
| **Medeni Hal** | Bekâr | Bekâr | Bekâr | Bekâr |
| **Nüfus Bölgesi (Katsayı oranı)** | Karadeniz:2.93<br>Akdeniz:1.01<br>Marmara:0.86 | Karadeniz:1.94<br>Doğu A.:1.63<br>G.Doğu A.: 1.1 | Doğu A.:2.36<br>Karadeniz:2.22<br>İç Anadolu: 1.2 | Doğu A.:2.56<br>Karadeniz:2.05<br>G.Doğu A.: 0.8 |
| **Okul Türü(%)** | Resmi Lise: 46<br>Anadolu L.: 30 | Resmi Lise : 45<br>Anadolu L.: 43 | Resmi Lise : 45<br>Anadolu L.: 43 | Resmi Lise : 26<br>Anadolu L.: 58 |
| **Yer. Puanı** | 349.14 | 387.07 | 411.59 | 331.05 |
| **Yer. Sırası** | 119453 | 83283 | 65753 | 46258 |
| **Genel Ortalama** | **1.93** | **2.32** | **2.42** | **2.64** |
| **Katkı Kredisi(%)** | Evet:2<br>Hayır:98 | Evet:15<br>Hayır:85 | Evet:57<br>Hayır:43 | Evet:45<br>Hayır:55 |
| **Öğrenim Kredisi(%)** | Evet:30<br>Hayır:70 | Evet:46<br>Hayır:54 | Evet:39<br>Hayır:61 | Evet:47<br>Hayır:53 |

Çizelge 4.25 (devamı)

| k-medoids | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Burs(%)** | Evet:19<br>Hayır:81 | Evet:22<br>Hayır:78 | Evet:36<br>Hayır:64 | Evet:26<br>Hayır:74 |
| **Tercih Sırası** | 8 | 10 | 12 | 7 |
| **Yüzde On(%)** | Evet:7<br>Hayır:93 | Evet:9<br>Hayır:91 | Evet:7<br>Hayır:93 | Evet:15<br>Hayır:85 |
| **Baba Eğitim Durumu(%)** | Lise: 30<br>Lisans: 28<br>İlköğretim: 26<br>İlkokul: 5 | Lise: 29<br>İlkokul: 27<br>İlköğretim: 19<br>Lisans: 13 | İlkokul: 33<br>İlköğretim: 27<br>Lise: 24<br>Lisans: 9 | İlkokul: 38<br>İlköğretim: 19<br>Lise: 13<br>Lisans: 11 |
| **Anne Eğitim Durumu(%)** | Lise: 31<br>İlkokul: 30<br>İlköğretim: 22<br>Lisans: 11 | İlkokul: 37<br>Lise: 22<br>İlköğretim: 21<br>Okuryazar Değil: 9 | İlkokul: 48<br>İlköğretim: 24<br>Lise: 12<br>Okuryazar Değil: 7 | İlkokul 43<br>İlköğretim: 19<br>Okuryazar Değil: 17<br>Lise: 13 |
| **Baba Meslek(%)** | Emekli: 22<br>Serbest M..: 19<br>Memur: 15<br>İşçi: 15 | Emekli: 28<br>Serbest M..: 16<br>İşçi: 10<br>Memur: 7 | Serbest M.: 21<br>Emekli: 16<br>İşçi: 7<br>Çiftçi: 7 | Emekli: 20<br>İşçi : 13<br>Çiftçi: 11<br>Serbest M.: 11 |
| **Anne Meslek(%)** | Ev Hanımı: 67<br>Emekli: 17<br>Memur: 6 | Ev Hanımı: 77<br>Emekli: 9<br>İşçi: 3 | Ev Hanımı: 79<br>İşçi: 6<br>Emekli: 4 | Ev Hanımı: 87<br>Emekli: 9<br>Öğretmen: 2 |
| **Kardeş Sayısı** | 2.9 | 3.13 | 3.21 | 3.8 |
| **Okuyan Kardeş Sayısı** | 2.14 | 2.11 | 2.24 | 2.4 |
| **Lise Mez. Notu** | 71.4 | 75.6 | 77.14 | 81.67 |
| **GSM(%)** | 500'lu: 42<br>530'lı: 31<br>540'lü: 27 | 500'lu: 40<br>530'lı: 33<br>540'lü: 27 | 500'lu: 44<br>530'lı: 38<br>540'lü: 18 | 500'lu: 47<br>530'lı: 31<br>540'lü: 22 |

### 4.6.4 Anket Verilerinin BCO Yöntemi İle Kümelenmesi

Verilere bulanık c-ortalamalar yönteminin uygulanmasının ardından elde edilen küme merkezleri Çizelge 4.26'da yer almaktadır. Çizelge incelendiğinde dört numaralı kümede yer alan öğrencilerin üniversite sınavında en başarılı olan, lise sıralarında en başarılı olan, üniversite genel not ortalaması en yüksek olan öğrencilerden oluştuğu görülmektedir. Buradan ortaöğretim başarısı, üniversite sınavı başarısı ve üniversite sıralarındaki başarı arasında bir doğru orantı vardır yargısına ulaşmak mümkündür. En başarısız öğrencilerin yer aldığı bir ve iki numaralı kümelere bakıldığında öğrencilerin hem lise sıralarında, hem üniversite sınavında hem de üniversite öğrenciliklerinde bu başarısızlıklarını devam ettirdikleri görülmektedir. Bulanık c-ortalamalar yönteminde

elde edilen sonuçlar ile k-ortalamalar ve k-medoids yöntemlerinde elde edilen sonuçlar benzerlik göstermektedir.

Çizelge 4.26 Bulanık c-ortalamalar yöntemi ile elde edilmiş küme merkezleri

| **BCO** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Yer. Kon. Tipi** | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme |
| **Sınıfı** | 2 | 1 | 1 | 3 |
| **Öğretim Türü** | Örgün | İkinci Ö. | Örgün | Örgün |
| **Cinsiyeti(%)** | Erkek:71<br>Kadın:29 | Erkek:67<br>Kadın:33 | Erkek:64<br>Kadın:36 | Erkek:60<br>Kadın:40 |
| **Yaşı** | 22.8 | 20.2 | 20.5 | 21.5 |
| **Medeni Durumu** | Bekâr | Bekâr | Bekâr | Bekâr |
| **Nüfus Bölgesi(Katsayı Oranı)** | Marmara:1.41<br>Karadeniz:1.1<br>Doğu A.: 1.1 | Karadeniz:2.34<br>G. Doğu A.:1.28<br>Doğu A.: 1.04 | Karadeniz:2.45<br>Doğu A:1.68<br>İç Anadolu:0.92 | Doğu A.:2.87<br>Karadeniz:1.77<br>Marmara: 0.89 |
| **Okul Türü(%)** | Resmi Lise : 25<br>Teknik Lise: 65 | Resmi Lise : 54<br>Anadolu L.: 33 | Resmi Lise : 43<br>Anadolu L.: 45 | Resmi Lise : 25<br>Anadolu L.: 51 |
| **Yerleştirme Puanı** | 281.92 | 367.77 | 397.94 | 349.95 |
| **Yer. Sırası** | 237350 | 97497 | 74631 | 49022 |
| **Tercih Sırası** | 6 | 9 | 10 | 9 |
| **Genel Ort.** | **1.52** | **2.09** | **2.38** | **2.6** |
| **Katkı Kredisi(%)** | Evet:0<br>Hayır:100 | Evet:1<br>Hayır:99 | Evet:34<br>Hayır:66 | Evet:54<br>Hayır:46 |
| **Öğrenim Kredisi(%)** | Evet:0<br>Hayır:100 | Evet:42<br>Hayır:58 | Evet:42<br>Hayır:58 | Evet:43<br>Hayır:57 |
| **Burs(%)** | Evet:0<br>Hayır:100 | Evet:21<br>Hayır:79 | Evet:26<br>Hayır:74 | Evet:33<br>Hayır:67 |
| **Yüzde On(%)** | Evet:0<br>Hayır:100 | Evet:7<br>Hayır:93 | Evet:9<br>Hayır:91 | Evet:15<br>Hayır:85 |
| **Baba Eğitim Durumu(%)** | Lise: 43<br>İlkokul: 28<br>İlköğretim: 28 | Lise: 29<br>İlköğretim: 25<br>Lisans: 22<br>İlköğretim: 10 | İlkokul: 30<br>Lise: 27<br>İlköğretim: 20<br>Lisans: 13 | İlkokul: 37<br>İlköğretim: 22<br>Lise: 13<br>Lisans: 10 |
| **Anne Eğitim Durumu(%)** | İlkokul: 43<br>İlköğretim: 28<br>Okuryazar Değil: 28 | İlkokul: 33<br>İlköğretim: 25<br>Lise: 24<br>Lisans: 11 | İlkokul: 41<br>Lise: 21<br>İlköğretim: 20<br>Okuryazar Değil: 7 | İlkokul:43<br>İlköğretim: 19<br>Okuryazar Değil: 18<br>Lise: 12 |
| **Baba Meslek(%)** | Emekli: 43<br>Memur: 14<br>Mobilyacı: 14 | Serbest M.: 22<br>Emekli: 21<br>İşçi: 13 | Emekli: 22<br>Serbest Ö.: 16<br>İşçi: 10 | Emekli: 22<br>Serbest M.: 13<br>İşçi: 12 |
| **Anne Meslek(%)** | Ev Hanımı: 71<br>Emekli: 14 | Ev Hanımı: 70<br>Emekli: 13 | Ev Hanımı: 78<br>Emekli: 7 | Ev Hanımı: 87<br>Emekli: 9 |

Çizelge 4.26 (devamı)

| BCO | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Kardeş Sayısı** | 3 | 3.13 | 3 | 3.82 |
| **Okuyan Kardeş Sayısı** | 2.14 | 2.25 | 2.07 | 2.402 |
| **Lise Mez.Notu** | 78.4 | 72.05 | 75.68 | 80.95 |
| **GSM(%)** | 540'lu: 43<br>500'lı: 41<br>530'lü: 16 | 500'lu: 44<br>530'lı: 34<br>540'lü: 22 | 500'lu: 42<br>530'lı: 34<br>540'lü: 24 | 500'lu: 41<br>530'lı: 34<br>540'lü: 25 |

### 4.6.5 Anket Verilerinin Gustafson-Kessel Yöntemi İle Kümelenmesi

Verilere Gustafson-Kessel yönteminin uygulanmasının ardından elde edilen küme merkezleri Çizelge 4.27'de yer almaktadır. Çizelge incelendiğinde bulanık c-ortalamalar yönteminde de olduğu gibi dört numaralı kümede yer alan öğrencilerin üniversite sınavında en başarılı olan, lise sıralarında en başarılı olan, üniversite genel not ortalaması en yüksek olan öğrencilerden oluştuğu görülmektedir. Buradan ortaöğretim başarısı, üniversite sınavı başarısı ve üniversite sıralarındaki başarı arasında bir doğru orantı vardır yargısı tekrardan doğrulanmış olmaktadır. Bulanık c-ortalamalar yönteminde elde edilen sonuçlar ile Gustafson-Kessel yönteminde elde edilen sonuçlar büyük benzerlik göstermektedir.

Çizelge 4.27 Gustafson-Kessel yöntemi ile elde edilmiş küme merkezleri

| GK | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Yer. Kon. Tipi** | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme |
| **Sınıfı** | 1 | 1 | 2 | 3 |
| **Öğretim Türü** | Örgün | Örgün | İkinci Ö. | Örgün |
| **Cinsiyeti(%)** | Erkek:60<br>Kadın:40 | Erkek:69<br>Kadın:31 | Erkek:64<br>Kadın:36 | Erkek:58<br>Kadın:42 |
| **Yaşı** | 20.28 | 19.97 | 21.29 | 21.54 |
| **Medeni Durumu** | Bekâr | Bekâr | Bekâr | Bekâr |
| **Nüfus Bölgesi(%)** | Karadeniz:2.5<br>Doğu A.:1.9<br>İç A.: 1.24 | Karadeniz:2.02<br>G. Doğu A.:1.3<br>Doğu A.: 1.32 | Karadeniz:2.45<br>Doğu A.:1.9<br>G. Doğu A.:0.9 | Doğu A.:2.35<br>Karadeniz:2.0<br>İç A.: 0.78 |
| **Okul Türü(%)** | Resmi Lise: 31<br>Teknik L.: 37 | Resmi Lise: 52<br>Anadolu L.: 36 | Resmi Lise: 42<br>Anadolu L.: 47 | Resmi Lise: 19<br>Anadolu L.: 58 |
| **Yer. Puanı** | 367.6 | 383.2 | 399.926 | 316.501 |
| **Yer. Sırası** | 110763 | 85184 | 76058 | 47300 |
| **Genel Ortalama** | **2.175** | **2.245** | **2.31** | **2.663** |
| **Katkı Kredisi(%)** | Evet:26<br>Hayır:74 | Evet:26<br>Hayır:74 | Evet:22<br>Hayır:78 | Evet:48<br>Hayır:52 |

Çizelge 4.27 (devamı)

| GK | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Öğrenim Kredisi(%)** | Evet:49<br>Hayır:51 | Evet:39<br>Hayır:61 | Evet:39<br>Hayır:61 | Evet:46<br>Hayır:54 |
| **Burs(%)** | Evet:11<br>Hayır:89 | Evet:30<br>Hayır:70 | Evet:22<br>Hayır:78 | Evet:31<br>Hayır:69 |
| **Yüzde On(%)** | Evet:2<br>Hayır:98 | Evet:10<br>Hayır:90 | Evet:9<br>Hayır:91 | Evet:15<br>Hayır:85 |
| **Baba Eğitim Durumu(%)** | Lise: 37<br>İlkokul: 22<br>Lisans: 22 | İlköğretim: 25<br>Lise: 22<br>İlkokul: 21 | Lise: 28<br>İlkokul: 26<br>İlköğretim: 25 | İlkokul: 40<br>İlköğretim: 17<br>Lise: 15 |
| **Anne Eğitim Durumu(%)** | İlkokul: 43<br>İlköğretim: 26<br>Lise: 23<br>Lisans: 8 | İlkokul: 30<br>İlköğretim: 24<br>Lise: 18<br>Lisans: 12 | İlkokul: 45<br>Lise: 24<br>İlköğretim: 18<br>Okuryazar Değil: 7 | İlkokul: 48<br>İlköğretim: 18<br>Okuryazar Değil: 17<br>Lise: 12 |
| **Baba Meslek(%)** | Emekli: 26<br>Serbest M.: 17<br>Memur: 12<br>İşçi: 9 | Emekli: 22<br>Serbest M.: 19<br>Memur: 12<br>İşçi: 10 | Emekli: 22<br>Serbest M.: 18<br>İşçi: 12<br>Öğretmen: 8 | Emekli: 21<br>Çiftçi: 15<br>İşçi: 15<br>Serbest M.: 10 |
| **Anne Meslek(%)** | Ev Hanımı: 77<br>Emekli: 6 | Ev Hanımı: 72<br>Emekli: 13 | Ev Hanımı: 78<br>Emekli: 8 | Ev Hanımı: 90<br>Emekli: 8 |
| **Tercih Sırası** | 9 | 8 | 12 | 7 |
| **Kardeş Sayısı** | 2.49 | 3.56 | 2.94 | 3.67 |
| **Okuyan Kardeş Sayısı** | 2.143 | 2.235 | 2.083 | 2.417 |
| **Lise Mez. Notu** | 73.475 | 74.194 | 75.904 | 82.1 |
| **GSM(%)** | 500'lu: 45<br>530'lı: 29<br>540'lü: 26 | 530'lu: 39<br>500'lı: 37<br>540'lü: 24 | 530'lu: 49<br>540'lı: 21<br>500'lü: 30 | 500'lu: 42<br>530'lı: 33<br>540'lü: 25 |

### 4.6.6 Anket Verilerinin Gath-Geva Yöntemi İle Kümelenmesi

Verilere Gath-Geva yönteminin uygulanmasının ardından elde edilen küme merkezleri Çizelge 4.28'de yer almaktadır. Çizelge incelendiğinde diğer yöntemlerden farklı olarak kümelerden bir tanesi yerleştiği kontenjan tipi bakımından DGS(Dikey Geçiş Sistemi) ile gelen öğrencilerden oluşmuştur. Bu öğrencilerin başarılarının diğer kümelerde yer alan öğrencilerin başarılarından çok düşük olduğu da gözden kaçmamaktadır. Buradan DGS ile yerleşen öğrencilerin normal yerleştirme ile yerleşen öğrencilerden daha başarısız oldukları sonucuna ulaşılabilmektedir.

En başarılı öğrencilerin yer aldığı kümenin dört numaralı küme olduğu görülmekte ve bu kümede yer alan öğrencilerin ebeveynlerin eğitim seviyelerinin diğer kümelerde yer

alan ebeveynlere göre daha düşük olduğu görülmektedir. Ayrıca en başarılı öğrencilerin doğu Anadolu bölgesinden geldiği de dikkat çekici bir başka noktadır.

Çizelge 4.28 Gath-Geva yöntemi ile elde edilmiş küme merkezleri

| **GG** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Yer. Kon. Tipi** | DGS ile Gelen | 1.Yerleştirme | 1.Yerleştirme | 1.Yerleştirme |
| **Sınıfı** | 2 | 1 | 1 | 3 |
| **Öğretim Türü** | Örgün | İkinci Ö. | Örgün | Örgün |
| **Cinsiyeti(%)** | Erkek:67<br>Kadın:33 | Erkek:65<br>Kadın:35 | Erkek:67<br>Kadın:33 | Erkek:60<br>Kadın:40 |
| **Yaşı** | 23.3 | 19.9 | 22.5 | 21.5 |
| **Nüfus Bölgesi (Katsayı Oranı)** | Doğu A.:1.88<br>Karadeniz:1.65<br>Ege: 1.28 | Karadeniz:2.49<br>Doğu A.:1.49<br>İç A.:0.92 | Doğu A.:2.42<br>G. Doğu A.:1.8<br>Karadeniz: 1.4 | Doğu A.:2.35<br>Karadeniz:2.06<br>Marmara: 0.83 |
| **Medeni Hal** | Bekâr | Bekâr | Bekâr | Bekâr |
| **Okul Türü(%)** | Teknik Lise:100 | Resmi Lise: 47<br>Anadolu L.: 44 | Resmi Lise:48<br>Anadolu L.:31 | Resmi Lise:21<br>Anadolu L.:58 |
| **Yerleştirme P.** | 284.158 | 387.218 | 401.956 | 316.06 |
| **Yerleştirme S.** | 233394 | 83237 | 71179 | 50620 |
| **Tercih Sırası** | 6 | 10 | 11 | 7 |
| **Genel Ort.** | **1.33** | **2.279** | **2.32** | **2.654** |
| **Katkı Kredisi(%)** | Evet:0<br>Hayır:100 | Evet:23<br>Hayır:77 | Evet:36<br>Hayır:64 | Evet:46<br>Hayır:54 |
| **Öğrenim Kredisi(%)** | Evet:0<br>Hayır:100 | Evet:41<br>Hayır:59 | Evet:43<br>Hayır:57 | Evet:46<br>Hayır:54 |
| **Burs(%)** | Evet:0<br>Hayır:100 | Evet:26<br>Hayır:74 | Evet:21<br>Hayır:79 | Evet:29<br>Hayır:71 |
| **Yüzde On(%)** | Evet:0<br>Hayır:100 | Evet:9<br>Hayır:91 | Evet:7<br>Hayır:93 | Evet:15<br>Hayır:85 |
| **Baba Eğitim Durumu(%)** | Lise: 50<br>İlköğretim: 33<br>İlkokul: 16 | Lise: 29<br>İlköğretim: 24<br>İlkokul: 21<br>Lisans: 15 | İlkokul: 31<br>Lisans: 21<br>İlköğretim: 19<br>Okuryazar Değil: 7 | İlkokul: 42<br>İlköğretim: 17<br>Lise: 15<br>Lisans: 8 |
| **Anne Eğitim Durumu(%)** | İlkokul 50<br>İlköğretim: 33<br>Lise: 16 | İlkokul: 38<br>İlköğretim: 24<br>Lise: 23<br>Lisans: 8 | İlkokul: 36<br>Okuryazar Değil: 29<br>Lise: 12<br>İlköğretim: 12<br>Lisans: 10 | İlkokul: 48<br>İlköğretim: 19<br>Lise: 15<br>Okuryazar Değil: 15 |
| **Baba Meslek(%)** | Emekli: 33<br>Öğretmen 16<br>Mobilyacı: 16<br>Serbest M.: 16 | Emekli: 21<br>Serbest M.: 21<br>İşçi: 13<br>Memur: %10 | Emekli: 29<br>Çiftçi: 17<br>Öğretmen: 14<br>İşsiz: 10 | Emekli: 23<br>İşçi: 15<br>Çiftçi: 13<br>Serbest M.: 10 |

Çizelge 4.28 (devamı)

| GG | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Anne Meslek(%)** | Ev Hanımı: 67<br>Emekli: 16<br>Memur: 16 | Ev Hanımı: 73<br>Emekli: 10<br>İşçi: 4 | Ev Hanımı: 83<br>Emekli: 7<br>Öğretmen: 5 | Ev Hanımı: 90<br>Emekli: 8<br>İşçi: 2 |
| **Kardeş Sayısı** | 3.16 | 2.74 | 4.88 | 3.64 |
| **Okuyan Kardeş Sayısı** | 2.16 | 2.15 | 2.19 | 2.43 |
| **Lise Mezuniyet Notu** | 77.9 | 73.507 | 79.595 | 81.98 |
| **GSM(%)** | 540'lu: 54<br>500'lı: 26<br>530'lü: 20 | 500'lu: 41<br>530'lı: 36<br>540'lü: 23 | 500'lu: 47<br>530'lı: 28<br>540'lü: 25 | 500'lu: 45<br>530'lı: 31<br>540'lü: 24 |

# BÖLÜM 5

## SONUÇ VE ÖNERİLER

Bu çalışmada, veri madenciliği standart sürecinin tüm aşamaları Namık Kemal üniversitesi Çorlu Mühendislik Fakültesinde 2011 yılsonu itibariyle kayıtlı lisans öğrencilerinin verileri üzerinde uygulanmış ve çok geniş bir çalışma alanı olan veri madenciliğinin, ağırlıklı olarak kümeleme fonksiyonları üzerinde durulmuştur.

Veri seti, başlangıçta 25 değişken ve 1934 kayıt içerirken, veri hazırlama aşamasında yürütülen faaliyetlerle 20 değişken ve 1588 kayıt içerecek şekilde biçimlendirilmiştir. Bu veri setinin yanı sıra Çorlu Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümünde öğrenimine devam eden 271 öğrenciye anket yapılmıştır. Anketler sayesinde öğrencilerin aileleri ile ilişkili bilgilerinin de öğrencilerin başarısını nasıl etkilediği araştırılmıştır.

Uygulama, veri setleri üzerinde birden çok kümeleme tekniğini kullanarak bu tekniklerin karşılaştırılması üzerine kurgulanmıştır. Kümeleme yöntemleri katı ve bulanık kümeleme algoritmaları olmak üzere iki ana başlık altında incelenmiştir. Katı kümeleme yöntemlerinden k-ortalamalar ve k-medoids algoritmaları ile bulanık kümeleme yöntemlerinden bulanık c-ortalamalar, Gustafson-Kessel ve Gath-Geva algoritmaları veri setlerine uygulanmıştır. Uygulanan bu yöntemler neticesinde bulanık kümeleme algoritmalarının katı kümeleme algoritmalarından daha başarılı sonuçlar ürettiği ortaya çıkmıştır. Bu belirlemenin gerçekleştirilmesi için Doğruluk İndekslerinden ve aykırı değer analizinden faydalanılmıştır. Literatürde en sık kullanılan Bölünme Katsayısı (PC), Sınıflandırma Entropisi (CE), Bölümlendirme İndeksi (SC) ve Ayırma İndeksi (S) yöntemleri kümeleme algoritmalarının başarılarını tespit etmek için kullanılan geçerlilik indeksleridir. Kümeleme tekniklerinin yanı sıra karar ağaçları kullanılarak veri seti üzerinde bir sınıflandırma işlemi de gerçekleştirilmiştir. Bu sınıflandırma işleminde verilerin %67'si öğrenme ve %33'ü test

için kullanılmıştır. Ve sınıflandırma işlemi sonucunda verilerin %68'i doğru sınıflandırılmıştır.

Kümeleme yöntemleri ile mevcut öğrenci verileri içerisinden anlamlı bilgiler ortaya çıkarılmaya çalışılmıştır. Tüm fakülte öğrencilerini kapsayan veri seti için yapılan kümeleme işlemi sonuçları şu şekildedir;

- Öğrencilerin üniversite sınavında elde ettikleri başarılarla üniversitedeki başarıları arasında doğru orantı olduğu ortaya çıkmıştır.
- Veri setinde yer alan öğrencilerden üniversite başarı ortalamaları bakımından en başarılı öğrenciler Karadeniz ve Doğu Anadolu Bölgesinden gelen öğrenciler olmuştur.
- Fakülteyi tercih eden öğrencilerin büyük kısmının resmi ve gündüz öğretim yapan devlet liseleri ile yabancı dille öğretim yapan Anadolu liselerinden geldiği tespit edilmiştir.
- Anadolu liselerini okuyarak üniversite sınavına giren öğrencilerin, düz lise öğrencilerine göre hem üniversite giriş sınavında daha başarılı oldukları hem de üniversite öğrenimlerinde daha başarılı oldukları tespit edilmiştir.

Anketlerle elde edilen veri setinden oluşturulan küme merkezleri şunları göstermektedir;

- Öğrencilerin lise mezuniyet notları ile üniversite öğrenimlerinde gösterdikleri başarı arasında pozitif ilişki olduğu ortaya çıkmıştır.
- Öğrencilerin başarıları ile ebeveynlerinin eğitim durumları arasında negatif ilişki bulunmaktadır. Yani eğitim seviyesi düşük ailelerin çocuklarının hem lise öğrenimlerinde, hem üniversite sınavında hem de üniversite öğrenimleri esnasında başarı durumlarının, diğer ailelerin çocuklarına göre daha iyi olduğu görülmektedir. Kümeleme işleminde kullanılan tüm yöntemlerde elde edilen sonuçlar bunu göstermektedir. Buradan eğitim seviyesi düşük olan ailelerin, eğitim seviyesi yüksek olan ailelere oranla çocuklarını eğitim anlamında daha iyi motive ettikleri sonucuna ulaşmak mümkündür.
- Anne mesleğinde ev hanımı oranı en yüksek olan kümelerin tüm yöntemler de en başarılı öğrencilerden oluşan kümeler olduğu dikkat çekici diğer bir noktadır. Buradan da çocukları ile daha fazla vakit geçiren annelerin, onların eğitimine pozitif katkı sağladıkları ortaya çıkmaktadır.

- Eğitim seviyesi düşük ailelerin çocuk sayılarının diğer ailelere nazaran oldukça yüksek olduğu görülmektedir. Buradan hareketle eğitim seviyesi düşük ailelerin aile planlaması hakkında daha bilgisiz oldukları sonucuna da ulaşmak mümkündür.
- Küme merkezlerini gösteren çizelgeler incelendiğinde, en başarılı öğrencilerin yer aldığı kümelerde okuyan kardeş sayısının hep daha yüksek olduğu görülmektedir. Bu da okuyan kardeşlerin eğitim anlamında birbirini motive ettiklerini ve desteklediklerini göstermektedir.

Ayrıca SAS Enterprise Miner yazılımının Multiplot ve Distrubituon Explorer özellikleri kullanılarak elde edilen analiz sonuçları ise şöyledir;

- Üniversite yaşamları boyunca öğrencilerin genel başarı notlarının 1.sınıftan 4.sınıfa doğru artış gösterdiği belirlenmiştir(Bkz. Çizelge 4.3).
- Mühendislik fakültesini tercih eden öğrencilerin ağırlıklı olarak erkek öğrencilerden oluştuğu ortaya çıkmıştır.
- Mühendislik fakültesindeki erkek öğrencilerin başarı yüzdelerinin bayan öğrencilere göre daha düşük olduğu tespit edilmiştir. Erkek öğrencilerin %50'si başarılı iken bayanlar da bu oran %67'lerdedir.
- Örgün öğretim öğrencilerinin genelde ikinci öğretim öğrencilerinden daha başarılı oldukları ortaya çıkmıştır.
- Öğrencilerin burs alma durumları ile başarıları arasında pozitif ilişki olduğu tespit edilmiştir.
- Fakülteyi tercih eden öğrencilerin yoğun olarak geldikleri bölgeler sırasıyla Karadeniz, Doğu Anadolu ve Marmara bölgeleridir.

Sonuç olarak, gerçekleştirilen veri madenciliği süreci ve uygulanan kümeleme teknikleri ışığında, sürecin en zorlu kısmının veri hazırlama aşaması olduğu, sürecin her aşamasında ise tekniklere ilişkin güçlü bilgi birikimine ihtiyaç duyulduğu, veri sayısının ve veri kalitesinin uygulamaların başarısında önemli birer faktör olduğu, bulanık kümeleme yöntemlerinin çalışmada kullanılan diğer yöntemlerden daha başarılı sonuçlar ürettiği ortaya çıkmıştır.

Bu çalışmanın devamı olarak mevcut öğrenci verilerine öğrencilerin üniversiteye yerleşmeden önceki sosyal ve ekonomik durumları ile üniversiteye yerleştikten sonraki sosyal ve ekonomik durumları ile ilgili bilgilerinde katılmasıyla daha başarılı sonuçlar

elde edilebilir. Aynı zamanda bu tez çalışmasının sonucunda elde edilen veriler göz önüne alınarak aynı çalışma farklı üniversitelerde ki öğrencilere uygulanabilir. Bu sayede farklı üniversitelerde öğrenim gören öğrencilerin başarılarını etkileyen faktörler arasındaki benzerlikler ve farklılıklar ortaya çıkarılabilir.

# KAYNAKLAR


[1] Guruler, H., İstanbullu, A. ve Karahasan M., (2010). "A new student performance analysing system using knowledge discovery in higher educational databases", Elsevier, 55:247-254.

[2] Kabakchieva, D., Stefanova, K. ve Kisimov, V., (2011). "Analyzing University Data for Determining Student Profiles and Predicting Performance", Educational Data Mining Symposium, 6-8 July 2011, Eindhoven.

[3] Bindiya, V., Jose, T., Unnikrishnan, A. ve Poulose, J., (2011). "Clustering Student Data to Characterize Performance Patterns", International Journal of Advanced Computer Science and Applications(IJACSA), 2(9):138-140.

[4] Tair, M. ve Halees, A., (2012). "Mining Educational Data to Improve Students' Performance: A Case Study", International Journal of Information and Communication Technology Research, 2(2):140-146.

[5] Yadav, S. ve Pal, S., (2012). "Data Mining A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal (WCSIT), 2(2):51-56.

[6] Halees, A., (2008). "Mining Students Data To Analyze Learning Behavior:A Case Study", The International Arab Conference on Information Technology (ACIT), 2008

[7] Radaideh, Q., AI-Shawakfa, E., ve AI-Najjar, M., (2006). "Mining student data using decision trees", In Proceedings of the 7th International Arab Conference on Information Technology (ACIT2006), Jordan.

[8] Sembiring, S., Zarlis, M., Hartama, D., Ramliana S. ve Wani E., (2011). "Prediction Of Student Academic Performance By An Application Of Data Mining Techniques", International Conference on Management and Artificial Intelligence, 6:110-114.

[9] Boumedyen S., Yusupov R. ve Alexandro, V., "Student Relationship in Higher Education Using Data Mining Techniques", Global Journal of Computer Science and Technology, 10(11):54-59.

[10] Erdoğan, Ş. ve Timor, M., (2005). "A Data Mining Application In A Student Database", Journal Of Aeronautics And Space Technologies, 2(2):53-57

[11] Han, J. ve Kamber,M., (2000). Data Mining Concepts And Tecniques, Morgan Kaufmann Publishers, San Francisco.

[12] Babadağ, K., (2006). "Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri", Industrial Application Software, 85-87.

[13] Alpaydın, E., (2000), "Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri", Bilişim 2000 Eğitim Semineri, 1-10.

[14] Hung, S., Yen, D. ve Wang, H., (2005). “Applying Data Mining To Telecom Churn Management”, Expert Systems With Applications, 1-10.

[15] Jacobs, P., (1999). ”Data Mining: What General Managers Need To Know”, Harvard Management Update, 4 (10): 8.

[16] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth,R. ve Uthurusamy, R., (1996). Advances In Knowledge Discovery And Data Mining, AAAI/MIT Press, London.

[17] Çakır, Ö., (2008). Veri Madenciliğinde Sınıflandırma Yöntemlerinin Karşılaştırılması-Bankacılık Müşteri Veri Tabanı Üzerinde Bir Uygulama, Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

[18] Sumathi, S. ve Sivanandam, S.N., (2006). Introduction to Data Mining and its Applications, Springer, New York.

[19] Piramuthu, S., (2004). “Evaluating Feature Selection Methods for Learning in Data Mining Applications”, European Journal of Operational Research, 156(2):483-494.

[20] Pyle, D., (1999). Data Preparation For Data Mining, Morgan Kaufmann Publishers, San Francisco.

[21] Zhu, J., (2003). “Sampling- An Efficient Solution For Data Mining Of Association Rules”, Department of Computer Science, Dalhousie University.

[22] Gülçe, G., (2010). Veri Ambarı ve Veri Madenciliği Teknikleri Kullanılarak Öğrenci Karar Destek Sistemi Oluşturma, Yüksek Lisans Tezi, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü, Denizli.

[23] Esen, M.F., (2009). Veri Tabanlarından Bilgi Keşfi: Veri Madenciliği ve Bir Sağlık Uygulaması, Yüksek Lisans Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

[24] Özkan, Y., (2008). Veri Madenciliği Yöntemleri, 1.Basım, Papatya Yayıncılık, İstanbul.

[25] Akpınar, H., (2000), “Veri tabanlarında Bilgi Keşfi ve Veri Madenciliği”, İ.Ü. İşletme Fakültesi Dergisi, 29:1-22.

[26] Akyüz, R., (2009). Sosyal Ağlarda Emniyet Verilerinin İncelenmesi, Yüksek Lisans Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya.

[27] Döşlü, A., (2008). Veri Madenciliğinde Market Sepet Analizi ve Birliktelik Kurallarının Belirlenmesi, Yüksek Lisans Tezi, YTÜ Fen Bilimleri Enstitüsü, İstanbul.

[28] Kalikov, A., (2006), Veri Madenciliği ve Bir e-Ticaret Uygulaması, Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

[29] Altıntaş, T., (2006). Veri Madenciliği Metotlarından Olan Kümeleme Algoritmalarının Uygulamalı Etkinlik Analizi, Yüksek Lisans Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, , Sakarya.

[30] Akbulut, S., (2006). Veri Madenciligi Teknikleri ile Bir Kozmetik Markasının Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu, Yüksek Lisans Tezi, G.Ü. Fen Bilimleri Enstitüsü, Ankara.

[31] Kona, H. V., (2003). Association Rule Mining Over Multiple Databases: Partioned and Incremental Approaches, The University of Texas, Arlington.

[32] Özçakır, F.C., (2006). Müşteri işlemlerindeki birlikteliklerin belirlenmesinde veri madenciliği uygulaması, Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[33] Özdamar, E. Ö., (2002). Veri Madenciliğinde Kullanılan Teknikler ve Bir Uygulama, Yüksek Lisans Tezi, Mimar Sinan Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[34] Bozkır, A., S., Sezer, E. ve Gök, B., (2009). "Öğrenci Seçme Sınavında(ÖSS) Öğrencilerin Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti", Uluslararası İleri Teknolojiler Sempozyumu, 13-15 Mayıs 2009, Karabük.

[35] Anand, A., (2003). "A study and comparison of data clustering techniques", Master Thesis, Faculty of The Graduate School of The University of Texas, El Paso.

[36] Karabatak, M. ve Önce, M., C., (2004). "Apriori Algoritması ile Öğrenci Başarısı Analizi", Elektrik Elektronik Bilgisayar Mühendisliği Sempozyumu, ELECO, 2004, 348-352.

[37] Sever, H. ve Oğuz, B., "Veri Tabanlarında Bilgi Keşfine Formal Bir Yaklaşım, Kısım 1: Eşleştirme Sorguları ve Algoritmalar", Bilgi Dünyası, 3(2), Ekim, 173-204.

[38] Mucha, H. J. ve Sofyan, H., "Cluster Analysis", http://edoc.hu-berlin.de/series/sfb-373-papers/2000-49/PDF/49.pdf, 07 Ağustos 2012.

[39] Davidson, Y., (2002). "Understanding K-means Non-hierarchical Clustering", Teknik Rapor, Computer Science Department of State University of New York, Albany.

[40] Işık M., (2006). Bölünmeli Kümeleme Yöntemleri İle Veri Madenciliği Uygulamaları, Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[41] Pang-Ning T., Steinbach, M. ve Kumar, V., (2005). Introduction to Data Mining., Addison Wesley.

[42] Ng, R. T. ve Han, J., (2002). "Clarans: A method for clustering objects for spatial data mining", IEEE Trans. on KDE, 14(5).

[43] Höppner, F., Klawonn, F., Kruse, R. ve Runkler, T., (2000). Fuzzy Cluster Analysis, John Wiley&Sons, Chichester.

[44] Kruse, R., Borgelt, C. ve Nauck, D., (1999). "Fuzzy Data Analysis: Challenges and Perspectives", IEEE Int. Conf. on Fuzzy Systems (FUZZIEEE99), Seoul, 1211-1216.

[45] Azem, Z., (2003). "A Comprehensive Cluster Validity Framework For Clustering Algorithms", MSc Thesis, The University of Guelph, Canada, 15-19.

[46] Moertini, V.S., (2002). "Introduction to Five Clustering Algorithms", Integral, 7(2).

[47] Özmen Ş., (2003). "Veri Madenciliği Süreci", Veri Madenciliği ve Uygulama Alanları Konferansı, İstanbul Ticaret Üniversitesi, İstanbul.

[48] Aydogan, F., (2003). E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirimi, Yayınlanmamış Yüksek Lisans Tezi, H.Ü. Fen Bilimleri Enstitüsü, Şanlıurfa.

[49] Sugumaran V., Muralidharan V. ve Ramachandran K. I., (2007). "Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing", Mechanical Systems and Signal Processing, 21(2), 930-942.

[50] Ayık, Y., Z., Özdemir, A. ve Yavuz, U., (2007). Lise Türü ve Lise Mezuniyet Başarısının Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10(2).

[51] Bozkır, A., S., Sezer, E. ve Gök, B., (2009). "Öğrenci Seçme Sınavında(ÖSS) Öğrencilerin Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti, Uluslararası İleri Teknolojiler Sempozyumu (IATS'09).

[52] Vandamme, J., P., Meskens, N. ve Superby, J., F., (2007). "Predicting Academic Performance by Data Mining Methods", Education Economics , 15:405-419.

[53] Şeker, Ş. E., Karar Ağacı Öğrenmesi (Decision Tree Learning), http://www.bilgisayarkavramlari.com/2012/04/11/karar-agaci-ogrenmesi-decision-tree-learning/, 28 Kasım 2012.

[54] Mitchell T. M., (1997). Machine learning, MIT press and the McGraw-Hill Companies Inc., Singapore.

[55] Yıldırım S., (2007). Bulanık Kümeleme Analizi Tekniği Kullanılarak Türkiye'deki Bir GSM Operatörünün Kullanıcılarının Profil Analizi, Yüksek Lisans Tezi, Marmara Üniversitesi, Sayısal Teknikler Bilim Dalı, İstanbul.

[56] Güler, N., (2006). Bulanık Kümeleme Analizi ve Bulanık Modellemeye Uygulamaları, Yüksek Lisans Tezi, Muğla Üniversitesi Fen Bilimleri Enstitüsü, Muğla.

[57] Koçyiğit, Y. ve Korükek, M., (2005), "EMG İşaretlerini Dalgacık Dönüşümü ve Bulanık Mantık Sınıflayıcı Kullanarak Sınıflama", İTÜ Dergisi/d Mühendislik, 4(3).

[58] Abonyi J. ve Feil B., (2007). Cluster Analysis for Data Mining and Systems Identification, Birkhäuser Verlag AG, Berlin.

[59] Büyükköz, D., (2010). Fuzzy Kümeleme Teknikleri ve Avrupa Birliği Üye Ülkeleri ile Türkiye'nin Fuzzy Kümelenmesi, Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

[60] Gath, J. ve Geva, A., (1989). "Unsupervised Optimal Fuzzy Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence", 11(7): 773-781.

[61] Kaufman, L., ve Rousseeuw, P. J., (1987), "Clustering by Means of Medoids Statistical Data Analysis Based on The L1–Norm and Related Methods", North-Holland, 405–416.

# ÖZGEÇMİŞ

## KİŞİSEL BİLGİLER

| | |
|---|---|
| **Adı Soyadı** | :Ahmet SAYGILI |
| **Doğum Tarihi ve Yeri** | :20.10.1985/Şereflikoçhisar |
| **Yabancı Dili** | :İngilizce |
| **E-posta** | :ahmetsygl@gmail.com |

## ÖĞRENİM DURUMU

| **Derece** | **Alan** | **Okul/Üniversite** | **Mezuniyet Yılı** |
|---|---|---|---|
| Lisans | Bilgisayar Mühendisliği | Çukurova Üniversitesi | 2009 |
| Lise | Fen Bilimleri | Hazım Kulak Anadolu Lisesi | 2004 |

## İŞ TECRÜBESİ

| **Yıl** | **Firma/Kurum** | **Görevi** |
|---|---|---|
| 2010 | Namık Kemal Üni. Çorlu Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü | Araştırma Görevlisi |