



T.C.
KONYA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ



MAKİNE ÖĞRENMESİ ALGORİTMALARI
KULLANARAK ÖĞRENCİLERİN AKADEMİK
BAŞARISINI ETKİLEYEN FAKTÖRLERİN
TESPİT EDİLMESİ

Fatih Hüseyin KAYA

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Ocak 2022
KONYA
Her Hakkı Saklıdır

TEZ KABUL VE ONAYI

Fatih Huseyin KAYA tarafından hazırlanan “Makine Öğrenmesi Algoritmaları Kullanarak Öğrencilerin Akademik Başarısını Etkileyen Faktörlerin Tespit Edilmesi” adlı tez çalışması 20/01/2022 tarihinde aşağıdaki jüri tarafından oy birliği ile Konya Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı Anabilim Dalı’nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri

İmza

Başkan

Dr. Öğr. Üyesi Hazım İŞCAN

.....

Danışman

Prof. Dr. Halife KODAZ

.....

Üye

Dr. Öğr. Üyesi Ahmet Cevahir ÇINAR

.....

Yukarıdaki sonucu onaylarım.

Prof. Dr. Saadettin Erhan KESEN
Enstitü Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Fatih Hüseyin KAYA
Tarih: 15/02/2022

ÖZET

YÜKSEK LİSANS

MAKİNE ÖĞRENMESİ ALGORİTMALARI KULLANARAK ÖĞRENCİLERİN AKADEMİK BAŞARISINI ETKİLEYEN FAKTÖRLERİN TESPİT EDİLMESİ

Fatih Hüseyin KAYA

**Konya Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı**

Danışman: Prof. Dr. Halife KODAZ

2022, 102 Sayfa

Jüri

**Prof. Dr. Halife KODAZ
Dr. Öğr. Üyesi Hazım İŞCAN
Dr. Öğr. Üyesi Ahmet Cevahir ÇINAR**

Bu tez çalışmasında makine öğrenmesi algoritmaları kullanılarak öğrencilerin akademik başarısını etkileyen faktörlerin tespit edilmesi amaçlanmaktadır. Araştırmada ilk veri seti olarak Hindistan'ın Assam eyaletinde gerçekleştirilen bir araştırmadan elde edilen veriler kullanılmıştır. İkinci veri seti olarak Portekiz'in Alentejo bölgesinden 2005 yılında iki devlet okulundan anket ile toplanmış veriler kullanılmıştır. Çalışmada uygulama çerçevesi çapraz endüstri standart işlem modeli kapsamında geliştirilmiştir. Elde edilen sonuçlar, Doğruluk, F-1 Score sonuçları ile doğrulanmış ve karşılaştırılmıştır. Rastgele Orman, Aşırı Gradyan Güçlendirmesi ve Destek Vektör Makinaları ile sınıflandırma modelleri oluşturarak öğrencilerin akademik başarısına etki eden önemli faktörler incelenmiştir. Buna göre öğrencilerin akademik başarısını etkileyen faktörlerin tespit edilmesinde Aşırı Gradyan Güçlendirmesi'nin en iyi sonuçlar verdiği görülmektedir. Ayrıca hesaplanan başarı ölçüleri ve tespit edilen faktörler literatürdeki benzer çalışmalar ile karşılaştırılmış ve önemli ölçüde benzerlik gösterdiği görülmüştür.

Anahtar Kelimeler: CRIDSP-DM, Rastgele Orman, Makine Öğrenmesi, Özellik Çıkarma, Öğrenci Başarısı, XGBoost.

ABSTRACT

MS THESIS

IDENTIFYING THE FACTORS AFFECTING STUDENTS' ACADEMIC ACHIEVEMENT USING MACHINE LEARNING ALGORITHMS

Fatih Hüseyin KAYA

**Konya Technical University
Institute of Graduate Studies
Department of Computer Engineering**

Advisor: Prof. Dr. Halife KODAZ

2022, 102 Pages

**Jury
Prof. Dr. Halife KODAZ
Asist. Prof. Dr. Hazım İŞCAN
Asist. Prof. Dr. Ahmet Cevahir ÇINAR**

The aim of this study was to identify the factors affecting students' academic achievement using machine learning algorithms. The first data covered in the study, the data obtained from a study conducted in the Assam state of India were used. The second data covers, data collected from two public schools in 2005 from the Alentejo region of Portugal were used. The application framework in the study was developed under the cross industry standard process for data mining. Using classification models with Random Forest, Extreme Gradient Boosting and Support Vector Machines, important factors affecting students' academic success were examined. The results obtained were verified and compared utilizing classification Accuracy, F-1 Score results. By creating classification models with Random Forest, Extreme Gradient Reinforcement and Support Vector Machines, important factors affecting students' academic success were examined. According to these results, it can be said that Extreme Gradient Boosting provides the best results in identifying the factors affecting students' academic achievement. Besides these results, it was predicted success scores and the identified factors were compared with similar studies in the literature and it was seen that they showed significant similarity.

Keywords: CRIDSP-DM, Feature Extraction, Random Forest, Machine Learning, Student Success, XGBoost,

ÖNSÖZ

Bu tez çalışmasında öğrencilerin başarısına etki eden faktörlerin tespitinde makine öğrenmesi yöntemleri ile tespiti üzerine çalışılmıştır. Çalışmalarında yardımlarını ve desteklerini esirgemeyen danışmanım Konya Teknik Üniversitesi Bilgisayar Mühendisliği öğretim üyesi Prof. Dr. Halife KODAZ'a ve bölümümüzün tüm değerli öğretim elemanlarına teşekkür ederim.

Fatih Hüseyin KAYA
KONYA-2022



İÇİNDEKİLER

ÖZET	iv
ABSTRACT.....	v
ÖNSÖZ	vi
İÇİNDEKİLER.....	vii
SİMGELER VE KISALTMALAR.....	ix
ŞEKİLLER.....	x
ÇİZELGELER.....	xii
1. GİRİŞ.....	1
1.1. Çalışmanın Amacı.....	1
1.2. Tezin Organizasyonu	4
2. KAYNAK ARAŞTIRMASI	5
3. MATERYAL VE YÖNTEM.....	8
3.1. Tez İçin Kullanılan Donanım ve Yazılımlar.....	8
3.1.1. Kişisel Bilgisayar.....	8
3.1.1.1. PyCharm 2019.3.4	8
3.1.1.2. Google Colaboratory.....	8
3.2. Teorik Esaslar	9
3.2.1. Çapraz Endüstri Standart İşlem Modeli (CRISP-DM)	9
3.2.1.1. İşin Anlaşılması	10
3.2.1.2. Verinin Anlaşılması	10
3.2.1.3. Verinin Hazırlanması	10
3.2.1.3.1. Korelasyon Analizi	11
3.2.1.4. Modelleme	12
3.2.1.5. Değerlendirme	14
3.2.1.6. Konuşlandırma.....	16
3.3 Makine Öğrenmesi Yöntemleri	18
3.3.1. Rastgele Orman Algoritması	19
3.3.2. Aşırı Gradyan Güçlendirme Algoritması.....	21
3.3.3. Destek Vektör Makineleri Algoritması.....	24
4. ARAŞTIRILAN YÖNTEM VE DENEYSEL SONUÇLAR.....	27
4.1. Araştırmanın CRISP-DM Yöntemine Göre Gerçekleştirilmesi	27
4.1.1. İşin Anlaşılması	27
4.1.2. Verinin Anlaşılması	28
4.1.2.1. Veri Seti I.....	28
4.1.2.2. Veri Seti II	37

4.1.3. Verinin Hazırlanması	50
4.1.3.1 Veri Seti I Hazırlanması	50
4.1.3.2 Veri Seti II Hazırlanması	58
4.1.4. Modelleme	65
4.1.1.1. Rastgele Orman.....	66
4.1.1.2. Aşırı Gradyan Güçlendirmesi	67
4.1.1.3. Destek Vektör Makineleri.....	68
4.1.5. Değerlendirme	69
4.2.5.1. Rastgele Orman Sınıflandırılması Modeli Değerlendirmesi	69
4.2.5.2. Aşırı Gradyan Güçlendirmesi Sınıflandırılması Modeli Değerlendirmesi	74
4.2.5.3. Destek Vektör Makineleri Modeli Değerlendirmesi	79
4.2.5.4. Genel Değerlendirme	84
4.2.5.4.1. Veri Seti I Değerlendirmesi	84
4.2.5.4.2. Veri Seti II Değerlendirilmesi.....	87
4.2.5.4.3. Elde Edilen Sonuçların Literatürdeki Sonuçlar İle Değerlendirilmesi	90
4.1.6. Konuşlandırma.....	95
SONUÇLAR VE ÖNERİLER	96
KAYNAKÇA.....	98

SİMGELER VE KISALTMALAR

Simgeler

d^+	: Destek Vektör Makineleri Algoritmasında Ayırıcı Yüzey
d^-	: Destek Vektör Makineleri Algoritmasında Ayırıcı Yüzey
ω	: Destek Vektör Makineleri Algoritmasında Hiper Düzlemin Normali
x	: Destek Vektör Makineleri Algoritmasında Vektör
r	: Pearson Moment Çarpım Korelasyon Katsayısı
$\bar{x}$	: X_i İçin Örneklem Aritmetik Ortalaması
$\bar{y}$	: Y_i İçin Örneklem Aritmetik Ortalaması
S_x	: X_i İçin Örneklem Standart Sapma
S_y	: Y_i İçin Örneklem Standart Sapma

Kısaltmalar

CRISP-DM	: Çapraz Endüstri Standart İşlem Modeli
TP	: Gerçek Pozitifler
TN	: Gerçek Negatifler
FP	: Yanlış Pozitifler
FN	: Yanlış Negatifler
GBM	: Gradyan Güçlendirme Makineleri
DVM	: Destek Vektör Makinesi
ARFF	: Özellik İlişki Dosya Formatı
CSV	: Virgülle Ayrılmış Olarak Veriler
XGBOOST	: Aşırı Gradyan Güçlendirme

ŞEKİLLER

Şekil 3.1. CRISP-DM Yaşam Döngüsü.....	9
Şekil 3.2. Karar Ağaçlarının Gelişimi	13
Şekil 3.3. Karmaşıklık Matrisi.....	14
Şekil 3.4. Makine Öğrenmesi Algoritması Türleri	18
Şekil 3.5. Rastgele Orman Algoritmasının Çalışma Mantığı	20
Şekil 3.6. Güçlendirme Algoritmalarının Gelişimi	22
Şekil 3.7. XGBoost'un GBM'yi Geliştirdiği Yönler.....	23
Şekil 3.8. DVM İle İki Sınıfa Ayrılmış Düzlem.....	24
Şekil 3.9. Doğrusal Olarak İki Hiper Düzleme Ayrılabilen Destek Vektörleri.....	25
Şekil 4.1. ARFF Formatında Veri Setinin Görünümü.....	50
Şekil 4.2. ESP Kolonunun Hiyerarşisi.....	54
Şekil 4.3. ESP Kolonunun Düzenlenmiş Hiyerarşisi	55
Şekil 4.4. Veri Seti I Korelasyon Matrisi	58
Şekil 4.5. CSV Formatında Veri Setinin Görünümü	59
Şekil 4.6. Veri Seti II Korelasyon Matrisi	65
Şekil 4.7. Rastgele Orman Sınıflandırması İçin İlk Alt Ağaç	67
Şekil 4.8. Veri Seti I İçin Rastgele Orman Sınıflandırması Modelinin Karmaşıklık Matrisi.....	70
Şekil 4.9. Veri Seti I İçin Rastgele Orman Sınıflandırması İle Önemli Faktörler.....	71
Şekil 4.10. Veri Seti II İçin Rastgele Orman Sınıflandırması Modelinin Karmaşıklık Matrisi.....	72
Şekil 4.11. Veri Seti II İçin Rastgele Orman Sınıflandırması İle Önemli Faktörler	73
Şekil 4.12. Veri Seti I İçin Aşırı Gradyan Güçlendirmesi Modelinin Karmaşıklık Matrisi	75
Şekil 4.13. Veri Seti I İçin Aşırı Gradyan Güçlendirmesi Sınıflandırması İle Önemli Faktörler.....	76
Şekil 4.14. Veri Seti II İçin Aşırı Gradyan Güçlendirmesi Modelinin Karmaşıklık Matrisi.....	77
Şekil 4.15. Veri Seti II İçin Aşırı Gradyan Güçlendirmesi Sınıflandırması İle Önemli Faktörler.....	78
Şekil 4.16. Veri Seti I İçin Destek Vektör Makineleri Modelinin Karmaşıklık Matrisi	80
Şekil 4.17. Veri Seti II İçin Destek Vektör Makineleri Sınıflandırması İle Önemli Faktörler.....	81
Şekil 4.18. Veri Seti II İçin Destek Vektör Makineleri Modelinin Karmaşıklık Matrisi.....	82
Şekil 4.19. Veri Seti II İçin Destek Vektör Makineleri Sınıflandırması İle Önemli Faktörler.....	83
Şekil 4.20. Veri Seti I İçin Sınıflandırma Algoritmalarının Karşılaştırılması.....	85
Şekil 4.21. Veri Seti I İçin Sınıflandırma Algoritmalarının F1-Score Karşılaştırmaları	86
Şekil 4.22. Veri Seti I İçin Sınıflandırma Algoritmalarının MSE,MAE ve RMSE Karşılaştırmaları.....	86
Şekil 4.23. Veri Seti I İçin Sınıflandırma Algoritmalarının RAE, RSE Karşılaştırmaları	87
Şekil 4.24. Veri Seti II İçin Sınıflandırma Algoritmalarının Karşılaştırılması	88
Şekil 4.25. Veri Seti II İçin Sınıflandırma Algoritmalarının F1-Score Karşılaştırmaları	88
Şekil 4.26. Veri Seti II İçin Sınıflandırma Algoritmalarının MSE, MAE ve RMSE Karşılaştırmaları.....	89
Şekil 4.27. Veri Seti II İçin Sınıflandırma Algoritmalarının RAE, RSE Karşılaştırmaları	89

Şekil 4.28. Veri Seti I İçin Tez Çalışması ile Orijinal Çalışmanın Doğruluk Oranların Karşılaştırması	90
Şekil 4.29. Veri Seti II İçin Tez Çalışması ile Orijinal Çalışmanın Doğruluk Oranların Karşılaştırması	91



ÇİZELGELER

Çizelge 3.1. Verinin Hazırlanması Adımında Yapılan İşlemler.....	10
Çizelge 3.2. Kaynak Araştırmasına Göre Algoritmaların Kullanım Miktarı	13
Çizelge 3.3. CRISP-DM İş Akışına Bir Genel Bakış	17
Çizelge 4.1. Veri Seti I Özellikleri	29
Çizelge 4.2. Veri Seti I Cinsiyet Kolonunun Özellikleri.....	29
Çizelge 4.3. Veri Seti I Kast Sınıfı Kolonunun Özellikleri	30
Çizelge 4.4. Veri Seti I 10.Sınıf Ortalaması Yüzdesi Kolonunun Özellikleri.....	30
Çizelge 4.5. Veri Seti I 12.Sınıf Ortalaması Yüzdesi Kolonunun Özellikleri.....	30
Çizelge 4.6. Veri Seti I İç Değerlendirme Ortalaması Yüzdesi Kolonunun Özellikleri	31
Çizelge 4.7. Veri Seti I Dönem Sonu Ortalaması Yüzdesi Kolonunun Özellikleri.....	31
Çizelge 4.8. Veri Seti I Önceki Dönemden Kalan Dersi Kolonunun Özellikleri.....	31
Çizelge 4.9. Veri Seti I Medeni Hal Durumu Kolonu Özellikleri.....	31
Çizelge 4.10. Veri Seti I Şehir veya Köyde Yaşanması Durumu Kolonunun Özellikleri	32
Çizelge 4.11. Veri Seti I Öğrenci Okula Giriş Tipi Kolonunun Özellikleri	32
Çizelge 4.12. Veri Seti I Aile Aylık Geliri Kolonunun Özellikleri	32
Çizelge 4.13. Veri Seti I Aile Genişliği Kolonunun Özellikleri.....	32
Çizelge 4.14. Veri Seti I Baba Eğitim Seviyesi Kolonunun Özellikleri.....	33
Çizelge 4.15. Veri Seti I Anne Eğitim Seviyesi Kolonunun Özellikleri	33
Çizelge 4.16. Veri Seti I Baba Mesleği Kolonunun Özellikleri	34
Çizelge 4.17. Veri Seti I Anne Mesleği Kolonunun Özellikleri.....	34
Çizelge 4.18. Veri Seti I Arkadaş Sayısı Kolonunun Özellikleri	34
Çizelge 4.19. Veri Seti I Ders Çalışma Süresi Kolonunun Özellikleri.....	35
Çizelge 4.20. Veri Seti I Geldiği Okul Tipi Kolonunun Özellikleri.....	35
Çizelge 4.21. Veri Seti I Sınav Dili Kolonunun Özellikleri	35
Çizelge 4.22. Veri Seti I Evden Üniversiteye Seyahat Süresi Kolonunun Özellikleri...	35
Çizelge 4.23. Veri Seti I Sınıfa Devam Yüzdesi Kolonunun Özellikleri	36
Çizelge 4.24. Veri Seti I Dönem Sonu Ortalama Kolonunda Sınıf Dağılımı.....	36
Çizelge 4.25. Veri Seti I Kolonların Kısaltmaları	37
Çizelge 4.26. Veri Seti II Öğrenci Veri Setinin Özellikleri.....	38
Çizelge 4.27. Veri Seti II Cinsiyet Kolonunun Özellikleri.....	39
Çizelge 4.28. Veri Seti II Yaş Kolonunun Özellikleri.....	39
Çizelge 4.29. Veri Seti II Okul Kolonunun Özellikleri	39
Çizelge 4.30. Veri Seti II Adres Kolonunun Özellikleri	39
Çizelge 4.31. Veri Seti II Baba Eğitim Seviyesi Kolonunun Özellikleri	40
Çizelge 4.32. Veri Seti II Anne Eğitim Seviyesi Kolonunun Özellikleri.....	40
Çizelge 4.33. Veri Seti II Aile Genişliği Kolonunun Özellikleri	40
Çizelge 4.34. Veri Seti II Baba Mesleği Kolonunun Özellikleri.....	41
Çizelge 4.35. Veri Seti II Anne Mesleği Kolonunun Özellikleri	41
Çizelge 4.36. Veri Seti II Veli Kolonunun Özellikleri	41
Çizelge 4.37. Veri Seti II Aile İlişkisi Kalitesi Kolonunun Özellikleri.....	42
Çizelge 4.38. Veri Seti II Okulu Seçme Sebebi Kolonunun Özellikleri	42
Çizelge 4.39. Veri Seti II Evden Üniversiteye Seyahat Süresi Kolonunun Özellikleri..	42
Çizelge 4.40. Veri Seti II Ders Çalışma Süresi Kolonunun Özellikleri	43
Çizelge 4.41. Veri Seti II Önceki Dönemden Kalan Ders Sayısı Kolonunun Özellikleri	43
Çizelge 4.42. Veri Seti II Okul Desteği Kolonunun Özellikleri.....	43
Çizelge 4.43. Veri Seti II Ailenin Eğitim Desteği Kolonunun Özellikleri.....	43
Çizelge 4.44. Veri Seti II Etkinliklere Katılma Kolonunun Özellikleri	44

Çizelge 4.45. Veri Seti II Ekstra Ücretli Ders Kolonunun Özellikleri	44
Çizelge 4.46. Veri Seti II Evde İnternet Erişimi Kolonunun Özellikleri	44
Çizelge 4.47. Veri Seti II Anaokuluna Gidip Gitmediği Kolonunun Özellikleri	44
Çizelge 4.48. Veri Seti II Yüksek Eğitim Almak İstemesi Kolonunun Özellikleri.....	45
Çizelge 4.49. Veri Seti II Romantik İlişki Kolonunun Özellikleri	45
Çizelge 4.50. Veri Seti II Boş Zaman Kolonunun Özellikleri.....	45
Çizelge 4.51. Veri Seti II Dışarı Çıkma Kolonunun Özellikleri	45
Çizelge 4.52. Veri Seti II Hafta Sonu Alkol Tüketimi Kolonunun Özellikleri	46
Çizelge 4.53. Veri Seti II Hafta İçi Alkol Tüketimi Kolonunun Özellikleri	46
Çizelge 4.54. Veri Seti II Sağlık Durumu Kolonunun Özellikleri	46
Çizelge 4.55. Veri Seti II Ebeveyn Durumu Kolonunun Özellikleri.....	46
Çizelge 4.56. Veri Seti II Devamsızlık Sayısı Kolonunun Özellikleri	47
Çizelge 4.57. Veri Seti II Birinci Dönem Notu Kolonunun Özellikleri	47
Çizelge 4.58. Veri Seti II İkinci Dönem Notu Kolonunun Özellikleri.....	47
Çizelge 4.59. Veri Seti II Final Notu Kolonunun Özellikleri.....	48
Çizelge 4.60. Veri Seti II Kolonların Kısaltmaları	49
Çizelge 4.61. Veri Setinin İlk 3 Satırı	51
Çizelge 4.62. Karakter Seti Değiştirilmiş Veri Setinin İlk 3 Satırı	52
Çizelge 4.63. Veri Seti I Temel Analizi	53
Çizelge 4.64. Veri Seti I Tüm Kolonların Veri Tipi.....	54
Çizelge 4.65. One Hot Encoder Örneği.....	56
Çizelge 4.66. Kategorik Verilerin Dönüşümü	56
Çizelge 4.67. Kolonlar Arasındaki Standart Sapma Değerleri	57
Çizelge 4.68. Veri Seti II'nin İlk 3 Satırı.....	60
Çizelge 4.69. Veri Seti II Temel Analizi	61
Çizelge 4.70. Veri Seti II Tüm Kolonların Veri Tipi	62
Çizelge 4.71. Beş Sınıflı Dönüşüm Sistemi	63
Çizelge 4.72. Kategorik Verilerin Dönüşümü	63
Çizelge 4.73. Kolonlar Arasındaki Standart Sapma Değerleri	64
Çizelge 4.74. Veri Seti I İçin Rastgele Orman Sınıflandırmasının Başarımları	69
Çizelge 4.75. Veri Seti II İçin Rastgele Orman Sınıflandırmasının Başarımları	70
Çizelge 4.76. Veri Seti I İçin Rastgele Orman Sınıflandırması Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri.....	70
Çizelge 4.77. Veri Seti II İçin Rastgele Orman Sınıflandırması Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri.....	72
Çizelge 4.78. Veri Seti I İçin Aşırı Gradyan Güçlendirmesi Sınıflandırmasının Başarımları.....	74
Çizelge 4.79. Veri Seti II İçin Aşırı Gradyan Güçlendirmesi Sınıflandırmasının Başarımları.....	75
Çizelge 4.80. Veri Seti I İçin Aşırı Gradyan Güçlendirmesi Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri.....	75
Çizelge 4.81. Veri Seti II İçin Aşırı Gradyan Güçlendirmesi Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri.....	78
Çizelge 4.82. Veri Seti I İçin Destek Vektör Makineleri Sınıflandırmasının Başarımları	80
Çizelge 4.83. Veri Seti II İçin Destek Vektör Makineleri Sınıflandırmasının Başarımları	80
Çizelge 4.84. Veri Seti I İçin Destek Vektör Makineleri Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri.....	81

Çizelge 4.85. Veri Seti II İçin Destek Vektör Makineleri Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri.....	83
Çizelge 4.86. Veri Seti I İçin Orijinal Çalışma ile Bu Çalışmadaki Faktörlerin Korelasyon Önem Sıralarının Karşılaştırılması.....	91
Çizelge 4.87. Veri Seti II İçin Orijinal Makaledeki En İyi Sonuç Veren Algoritmaların En Önemli Faktörlerinin Karşılaştırılması.....	92
Çizelge 4.88. Veri Seti II'yi Kullanan Diğer Çalışmalar İle Tez Çalışmasının Karşılaştırılması.....	93



1. GİRİŞ

1.1. Çalışmanın Amacı

Günümüz teknolojisinde bilgiye ulaşmak kolaylaşmasına rağmen öğrencilerdeki akademik başarı oranı farklılık göstermektedir. Aynı şekilde bazı öğrenciler bilinçsiz bir şekilde sadece dersleri ezberleyerek notlarını yükseltmeye çalışabilmektedir ve ezberleyerek çalışma gerçek öğrenme değildir. Oysa öğrenme; bilgileri anlama, düşünme, ön bilgilerle bütünleştirme, yeni bilgiler üretme ve uygulamayı içermektedir. Öğrenme ile birey kendini sürekli geliştirmekte ve yenilemektedir (Güneş, 2020). Bundan dolayı ezberden kaçınılmalıdır.

Öğrencilik hayatında edinilen başarı iş dünyasına adım atarken bir birey için önemlidir. Üniversitede en üst düzeyde eğitim verilmektedir. Üniversite aşaması profesyonel iş dünyasına giriş öncesi son aşamadır. Öğrenci not ortalaması birçok işyerine başvururken veya akademik kariyer yapmak istenildiğinde bile bir sorun olarak ortaya çıkabilmektedir. Eğitim sürecinde başarıyı etkileyen çeşitli faktörler ortaya çıkmaktadır. Bunlar tespit edilerek akademik alanda başarı oranı yükseltilebilmektedir. Tespit edildiği takdirde öğrenciler ve akademisyenler bu faktörleri kontrol altında tutabilmektedir.

Üniversite öğrenimi hayatın en önemli dönüm noktalarından biridir ve önemi gittikçe artmaktadır. Ülkemizde 2019-2021 yılı istatistiklerine göre 7 milyon 940 bin 133 üniversite öğrencisi ve 2020-2021 yılı istatistiklerine göre ise 8 milyon 240 bin 997 üniversite öğrencisi bulunmaktadır (Yök, 2020). Buradaki veriler göz önünde bulunduğunda 300 bin civarında öğrenci sayısında artış olmuştur. Bu sayılar ülkemiz açısından oldukça yüksektir ve bu durumda öğrencinin aldığı eğitim kalitesi önemli hale gelmektedir. Üniversitelerin bütçeleri, Türkiye'nin eğitime yaptığı kamu harcamalarının yaklaşık dörtte birini oluşturmaktadır. Türkiye'de üniversitelerin geleceğine ilişkin tahminler incelendiğinde 2010-2025 yılları arasında üniversite öğreniminde büyüme, Ekonomik Kalkınma ve İşbirliği Örgütü ülkelerinden daha hızlı bir biçimde devam edeceği öngörülmüştür (Kavak, 2011). Bu nedenle üniversitede öğrenci performansını etkileyen faktörlerin belirlenmesi oldukça önemli hale gelmiştir.

Türkiye’de ve uluslararası yapılan çalışmalarda bu faktörler farklılık göstermektedir. Çünkü ülkelerin ekonomisi, kültürü, aile yapısı ve eğitim sistemi gibi konular işin içine girince oldukça farklı faktörler ortaya çıkabilmektedir. Özellikle bu faktörlere ebeveynlerin dikkat etmesi büyük önem taşımaktadır. Dam (2008) yapmış olduğu çalışmada, öğrenci başarısızlığında ailenin çocuğa karşı ilgisizliği ve eğitimin önemini kavrayamamış olmanın etkisinin büyük olduğunu belirtmiştir. Son yıllardaki nispi bilinçlenme ve ilerlemeye rağmen henüz ailelerin okullardaki eğitim sürecinin etkili bir unsuru haline gelemediği tespit edilmiştir.

Eğitim amaçları yönündeki davranış değişikliklerinin öğrencide ne ölçüde gerçekleşebildiğinin ölçülmesi ve değerlendirilmesi, bir eğitim sistemi çerçevesinde yapılan değerlendirme işlemlerinin en önemlileri arasında sayılabilir (Koç, 1981).

Eğitim kararları üç grupta toplanmıştır; öğrenciler, eğitim içeriği, eğitim ve değerlendirme ve değerlendirme süreçleri. Kararlara öğrenciler yönünden bakıldığında, değerlendirmenin görevlerinden biri öğrencilerin bulundukları okulları bitirdikten sonra ilgi ve yeteneklerine uygun alanlara yönltilmesidir. Diğer bir görev ise öğrenci başarısını değerlendirmektir (Turgut ve Baykul, 2012).

Eğitim seviyesi yüksek olan anne ve babalar çocuklarının akademik başarıları için en uygun ortamı sağlamak istemektedirler. Marzano (2003) yaptığı çalışmada, yönetimindekilerin öğrencilerin başarı kalitesinin artması için uygun ev ortamı yaratmak konusunda anne ve babalara danışmanlık ve rehberlik sağlayabileceğini belirtmiştir. Farklı okullarda öğrenim gören öğrencilerin başarı kaçınılmaz olarak farklı olabilmektedir. Bu farklılıkların yalnızca %10’u öğrencilerin gittiği okulların kalitesinden kaynaklanmaktadır (Marzano ve ark., 2008).

Aile evi, öğrenci evi, yurt ve apart gibi ortamlar öğrencinin akademik performansını etkilemektedir. Çemrek ve Filiz (2007), barınma yerinden memnun olmayan öğrencilerin memnun olmama nedenlerinin sırasıyla ders çalışmama (% 33.5), kendini rahat hissetmeme (% 28), arkadaşlarıyla anlaşmama (% 18), ulaşım zorluğu (% 14.3) ve diğer nedenler (% 6.2) olduğunu tespit etmişlerdir.

Akademik başarı olarak nitelendirilebilecek öğrenme ürünü, genel anlamıyla bireyin içinde tekrarlanmaktadır ve enerji sarfı bireyin içinde tekrarlanmaktadır (Özoğlu ve Koç, 1995).

Akademik başarı genellikle, öğrencinin psikomotor ve duyuşsal gelişiminin dışında kalan, bütün program alanlarındaki bilişsel davranış değişmelerini ifade etmektedir (Stanley ve Glock, 1967).

Cinsiyetin akademik başarı üzerinde önemli olup olmadığı günümüzde hala tartışılmaktadır. Chambers ve Schreiber (2004) öğrencilerin cinsiyetleri ile okul başarı tutumları arasında anlamlı ve pozitif yönlü bir ilişki olduğu ve kız öğrencilerin okul başarı tutumlarının erkek öğrencilere göre daha yüksek olduğunu öne sürmüşlerdir.

Peker (2003), öğrencilerin genel akademik performansının cinsiyete, lise diplomasına, lise bitirme derecesine, lisedeki okul türüne ve lisedeki alanına göre değişip değişmediğini incelemiştir. Araştırma sonucunda kız öğrencilerin genel akademik başarı not ortalamalarının erkek öğrencilerin genel akademik başarı not ortalamalarından daha yüksek olduğunu tespit etmiştir.

2020 yılındaki Covid-19 pandemisi göz önüne alındığında bu faktörlerin ne kadar önemli olduğu ortaya çıkmıştır. Bu faktörler değişebilmektedir. Bundan dolayı kontrol edilmesi gerekmektedir. Örneğin, Türkiye’de 2020 yılında dünyada görülen Covid-19 pandemisi nedeniyle örgün eğitimin yerini uzaktan eğitimin almasıyla birlikte başarıyı etkileyen başka faktörler ortaya çıkmıştır. Yolcu (2020), yapmış olduğu çalışmada, bilgisayara sahip olma, internet erişimi, derslere devamlılık gibi konular üzerinden bir çalışma gerçekleştirmiştir.

Thomas (1998) anketi, insanların yaşam koşullarını, davranışlarını, dini görüşlerini ve eğilimlerini betimlemeye yönelik bir dizi sorudan oluşan bir araştırma materyali olarak tanımlanmaktadır. Dini konular, cinsel eğilim ve çocuk sahibi olup olmama gibi konular hakkında anket ile bilgi almanın güç olduğu düşünülmektedir. Birey bu soruları kendi özel hayatına müdahale olarak düşünebilmektedir ve yanıltıcı cevaplar verebilmektedir. Eğer bu yanıltıcı cevaplar yapılacak olan çalışmada sınıflandırma işlemine uygulanırsa yanlış bulgular bulunmasına sebebiyet verebilir.

Bir araştırma için hedef kitlenin nüfussal özellikleri ile birlikte davranışları, değerleri, performansları ve görüşleri de tespit edilmeye çalışılabilir. Bu durumda hazırlanacak anket, ölçülmek istenilen özelliklere göre bölümlere ayrılır ya da her amaç için ayrı bir anket hazırlanabilir. Buna göre anketlerde, ölçülen özelliklere göre dört farklı soru grubu kullanılabilir;

i) Cevaplayan kişinin nüfussal özelliklerini (Cinsiyeti, mesleği, eğitim düzeyi, ailesinin sosyoekonomik durumu, eğitim seviyesi) betimlemeye yönelik gibi olgusal sorular,

ii) Cevaplayan kişinin bir soruda ne bildiği ve bilgiye ulaşma kaynağını belirlemeye yönelik bilgi sorular,

iii) Cevaplayan kişinin bir konu ya da nesneye göre davranışlarını (Bilgisayar kullanımı, öğrenci kulüplerine katılma, sınıfta öğretmenin anlattığı konuya katılma gibi) belirlemeye yönelik davranış soruları,

iv) Cevaplayan kişinin bir konu ya da nesneye göre duyguları (İş yerinde yüksek konuma gelme, Bölümünü sevip sevmemesi gibi) belirlemeye yönelik inanç ve kanı soruları,
şeklinde kullanılabilir.

1.2. Tezin Organizasyonu

Tezin 1. bölümünde giriş yapılmış ve tez konusu hakkında bilgi verilmiştir. 2. bölümde makine öğrenmesi algoritmaları kullanarak öğrencilerin akademik başarısını etkileyen faktörlerin tespit edilmesi konusunda daha önce yapılmış olan çalışmalar hakkında bilgiler verilerek kaynak araştırması yapılmıştır. 3. bölümde tez çalışmasında kullanılan materyal ve yöntemler hakkında bilgi verilmiştir. Bu bağlamda iki adet veri seti kullanılmıştır. İlk veri seti olarak Hussain ve ark. (2018) yaptıkları çalışmaya ait veriler kullanılmıştır. İkinci veri seti olarak ise tarafından Cortez ve Silva (2008) yaptıkları çalışmaya ait veriler kullanılmıştır. Yöntem olarak ise CRISP-DM metodolojisi kullanılmıştır. Tez çalışmasında kullanılan sınıflandırma algoritmaları olan Rastgele Orman algoritması, Aşırı Gradyan Güçlendirme algoritması ve Destek Vektör Makineleri algoritması hakkında bilgi verilmiştir. 4. bölümde tez çalışmasında iki veri seti için ayrı ayrı elde edilen sonuçlar verilmiştir. Son bölümde ise tez çalışması sonuçlandırılmış ve öneriler sunulmuştur.

2. KAYNAK ARAŞTIRMASI

Üniversitelerde gerek daha iyi eğitim vermek isteyen akademisyenler, gerekse de iyi bir not ortalamasına sahip olmaya çalışan öğrenciler başarı düzeyinin nasıl artması gerektiği hakkında bilgi sahibi olmak istemektedirler. Bilim ve teknolojideki gelişmeler sayesinde makine öğrenmesi kavramının önemi artmıştır. Bu durumda birçok kişi makine öğrenmesi alanında, bilinen materyal ve yöntemler ile birlikte yeni materyal ve yöntemlerle deneysel çalışmalar gerçekleştirmekte ve bu alanda çalışmalar devam etmektedir.

Ibrahim ve Rusli (2007) öğrencilerin okullarındaki başarısını tahmin etmede Doğrusal Regresyon, Karar Ağacı, Yapay Sinir Ağları yöntemlerini karşılaştırmıştır. Bu çalışmada, okuldaki genel başarı ortalaması önemli faktör olarak, aile geliri, okul programı, önceki okulundaki yatılı veya yatılı kalmama ve teknoloji kullanımı az önemli faktörler olarak belirlenmiştir. Yaptığı çalışmada en iyi sonuç Yapay Sinir Ağları yöntemi ile alınmıştır.

Wook ve ark. (2009) öğrencilerin akademik performanslarını tespit etmek için Kümeleme ve Karar Ağacı birleşimi hibrit bir algoritma ile Yapay Sinir Ağları algoritmasının sınıflandırma performanslarını karşılaştırmıştır.

El-Refae ve Al-Shayea (2010) üniversite öğrencilerinin gelecekteki başarısını tahmin etmek için Yapay Sinir Ağları kullanarak, İktisat Fakültesi'ne kayıtlı 208 öğrenciden alınan verilerle bir analiz gerçekleştirmiştir. Ortaöğretim başarı notu, bir üniversite öğrencisinin mezun olması için en önemli faktör olarak belirlenmiştir. Cinsiyet, ortaöğretim alanı, yatılı kalma durumu faktörleri daha az etkili olarak tespit edilmiştir.

Kumar ve Vijayalakshmi (2011) öğrenme sürecini en iyi bir şekilde gerçekleştirerek, kaliteli öğrenciler yetiştirmek için lise öğrencileri verisi üzerinde yaptığı çalışmada Karar Ağacı kullanarak final sınavında başarılı veya başarısız olacak öğrencilerin performansını tahmin etmek için Rapidminer uygulamasını kullanmıştır. Yapmış olduğu çalışmada sırası ile mezuniyet oranı ile mezun olma ve olmama durumunda en etkili faktörün üçüncü yarıyıl olduğunu tespit etmiştir.

Osmanbegović ve Suljić (2012) Çok Katmanlı Algılayıcı, Karar Ağacı ve Naive Bayes algoritmalarını kullanarak Tuzla Üniversitesi İktisat Fakültesi'nde 2010-2011 eğitim-öğretim yılı birinci sınıf öğrencileri ve kayıt sırasında alınan veriler ile öğrencilerin akademik başarısını tahmin etmeye çalışmıştır. Yaptıkları çalışmada %76,65 doğrulukla Naive Bayes algoritmasının en iyi performansı gösterdiğini tespit etmişlerdir.

Ramesh ve Ramar (2013) yaptıkları çalışmada, Hindistan'ın Kancheepuram bölgesinden rastgele dokuz okuldan seçilen, rastgele 900 öğrenciye çeşitli kişisel, sosyoekonomik, psikolojik sorulardan oluşan anket yapılmıştır. Yapılan anket verileri üzerinde Naive Bayes, Çok Katmanlı Algılayıcı, Sıralı Minimal Optimizasyon, Karar Ağacı, Azaltılmış Hata Budaması algoritmaları kullanılarak öğrencilerin son değerlendirmede performansını etkileyen faktörler belirlenmiş ve öğrencilerin notları tahmin edilmeye çalışılmıştır. Çok Katmanlı Algılayıcı algoritması %72.38 doğruluk sağlayarak en iyi performansı göstermiştir.

Cheewaparakobkit (2013) üniversite öğrencilerinin uluslararası bir programdaki akademik performansını tahmin etmek için 2001 ve 2011 yılları arasında Tayland'da bir üniversiteye kayıtlı 1.600 öğrenciye ait 22 özellik içeren bir veri seti kullanmıştır. Karar Ağacı ve Yapay Sinir Ağları sınıflandırma algoritmaları üzerinde veri setini uygulamıştır. Karar Ağacı sınıflandırma algoritması %85,188 oranında doğruluğa ulaşmıştır.

Ruby ve David (2014) üniversite öğrencilerinin performansını ölçmek için 165 kayıtlık bir veri seti kullanmıştır. Çok Katmanlı Algılayıcı, Azaltılmış Hata Budaması, Karar Tabloları ve 3 farklı Karar Ağacı algoritmaları kullanarak karşılaştırma yapılmıştır. Çok Katmanlı Algılayıcı algoritması %74.8 doğruluğa ulaşmıştır.

Christian ve Ayub (2014) öğrenciler ve mezun olmuş öğrencilerin performansını etkileyen faktörleri tespit etmek için bir çalışma gerçekleştirmişlerdir. Naive Bayes ve Karar Ağaçları algoritmaları kullanılarak öğrencilerin eğitim, kişisel, üniversiteye kabul türü ve akademik verileri üzerinde çalışmışlardır. Çalışmada öğrencilerin eğitimlerini bitirip, bitirememesi konusunda cinsiyetin önemli olduğunu tespit etmişlerdir.

Ogwoka ve ark. (2015) öğrencilerin akademik performansını tahmin etmek için yaptığı çalışmada K-Means ve Karar Ağacı Algoritması ile bir araya getirilmiş sistem kullanarak boyut indirgemiş ve öğrencilerin performansını artırmak için bir model oluşturmuşlardır.

Kaur ve ark. (2015) yavaş öğrenen lise öğrencilerini tespit etmek için, sınıflandırma algoritmaları kullanarak tahmine dayalı modeller geliştirmiştir. Naive Bayes, Karar Ağacı, Sıralı Minimal Optimizasyon, Azaltılmış Hata Budaması ve Çok Katmanlı Algılayıcı algoritmaları kullanarak WEKA üzerinde modelleri test etmişlerdir. Tüm sınıflandırma algoritmaları arasında en iyi performansı Çok Katmanlı Algılayıcılar %75 doğrulukla göstermiştir.

Khan ve ark. (2015) Karar Ağacı algoritmasını kullanarak öğrencilerin final notunu tahmin etmiştir. Final notunun tahmin edilmesinde ebeveynlerin mesleğinin önemli bir rol oynadığı ifade edilmiştir.

Guo ve ark. (2015) öğrencilerinin başarılarını derin öğrenme algoritması kullanarak tespit etmeye çalışmıştır. Naive Bayes, Çok Katmanlı Algılayıcılar ve Destek Vektör Makineleri algoritmalarını kullanmışlardır. En iyi başarıyı Destek Vektör Makineleri'nde elde etmişlerdir.

Mesarić ve Šebalj (2016) üniversite öğrencilerinin akademik başarılarını tahmin etmek için bir çalışma geliştirmiştir. Azaltılmış Hata Budaması ve Karar Ağacı algoritmalarını kullanmışlardır. Azaltılmış Hata Budaması algoritmasının %79 sınıflandırma doğruluğuna ulaştığı tespit edilmiştir. En önemli faktörlerin devletin yaptığı sınav puanları, lisede alınan puanlar ve Hırvatça dil sınavındaki puanlar olduğu ifade edilmiştir.

Barrak ve ark. (2016) King Saud Üniversitesi'ndeki öğrencilerin önceki dönem aldıkları derslerdeki notlarına göre öğrencilerin performansını artırmak ve nihai not ortalamalarının erken tahminini tespit etmek için WEKA üzerinde Karar Ağacı algoritmasını kullanmışlardır.

Hussain ve ark. (2018) yapmış oldukları çalışmada, Assam eyaletinden üç farklı okul seçilmiş ve 24 adet faktör içeren bir anket düzenlenmiştir. 300 adet anket sonucunda öznitelik seçme kullanarak Karar Ağacı, Naive Bayes, Part ve Rastgele Orman Sınıflandırması algoritmaları üzerinde testler yapmışlardır. Ve en iyi sonucu Rastgele Orman algoritmasının verdiği ifade edilmiştir.

Nurafifah ve ark. (2019) Naive Bayes, Karar Ağacı, Yapay Sinir Ağları ve Destek Vektör Makineleri algoritmaları kullanarak öğrencilerin mezuniyet tahminini yapmaya çalışmıştır. Yaptıkları çalışmada en yüksek tahmin doğruluğuna Yapay Sinir Ağları algoritmasının ulaştığını ifade etmişlerdir.

Vijayalakshi ve Venkatachalapathy (2019) Altı adet algoritma üzerinden doğrulukları ile karşılaştırma yaptıklarında %84 başarı ile Derin Sinir Ağı en iyi algoritma olarak ortaya çıkmıştır.

Harvey ve Kumar (2019) ilk on iki yıllık eğitimden oluşan bir veri setini analiz etmek için makine öğrenmesi modelleri geliştirmişlerdir. Doğrusal Regresyon, Karar Ağacı ve Naive Bayes algoritmalarını kullanmışlardır. Kumar kullandığı algoritmalar arasından Naive Bayes algoritmasının lise öğrencileri için matematik puanlarını tahmin ederken en yüksek doğruluğa ulaştığını ifade etmişlerdir.

3. MATERYAL VE YÖNTEM

3.1. Tez İçin Kullanılan Donanım ve Yazılımlar

3.1.1. Kişisel Bilgisayar

Makine öğrenmesi algoritmaları kullanarak öğrencilerin akademik başarısını etkileyen faktörlerin tespit edilmesi için PyCharm 2019.3.4 programı kullanılmıştır. PyCharm programının ve yapılan uygulamanın çalıştırıldığı bilgisayarın özellikleri aşağıdaki gibidir;

- BIOS: 0401 (type: UEFI)
- İşlemci: Intel(R) Core(TM) i3-4170 CPU @ 3.70GHz (4 CPUs), ~3.7GHz
- Bellek: 8192MB
- Sabit Disk: 670 GB
- Ekran Kartı: 1024 MB

Yüksek Bellek ve uzun süre gerektiren veri ön işleme çalışmalarında sıkıntı yaşamamak için Google Colaboratory üzerinde oluşturulan çevrimiçi python not defteri üzerinde çalışılmıştır.

3.1.1.1. PyCharm 2019.3.4

Uygulamanın gerçekleştirilmesinde Python programlama dili kullanılmıştır. PyCharm geliştiriciye hem Python 2.7 hem de Python 3.5 üzeri sürümler için geliştirme sağlamaktadır. PyCharm Python dışında JavaScript, TypeScript, CoffeeScript, CSS, HTML, SQL dilleri içinde destek sağlar. Python'un açık kaynak kodlu olması ve entegrasyonu kolay olması nedeni ile tercih edilmiştir. Dünyanın en popüler uygulamalarından Instagram, Google, Spotify, Dropbox, Netflix, Facebook kullanıcılarına hangi ürünleri sunacağını belirlemek için Python ile yazılımlar geliştirmektedir.

3.1.1.2. Google Colaboratory

Derin öğrenme, Makine Öğrenmesi, Yapay Zekâ gibi alanlarda büyük veri ile çalışılmak istendiği zaman yüksek bir ram ve yüksek bir işlemci gerekir. Google Colaboratory bu noktada sağlam bir işlemciye ücretsiz erişim için tamamen yapılandırılmış bir çalışma zamanı sağlar. Sınıflandırma için uygun parametre analizi yaparken bilgisayarın özellikleri yetersiz geldiğinden Google Colaboratory kullanılmıştır.

3.2. Teorik Esaslar

Gün geçtikçe teknoloji sayesinde veriler veri madenciliği teknikleri kullanılarak çıkarılabilmekte, bundan ötürü modellerin eğitim süreçlerinin kalitesi iyileştirilebilmektedir.

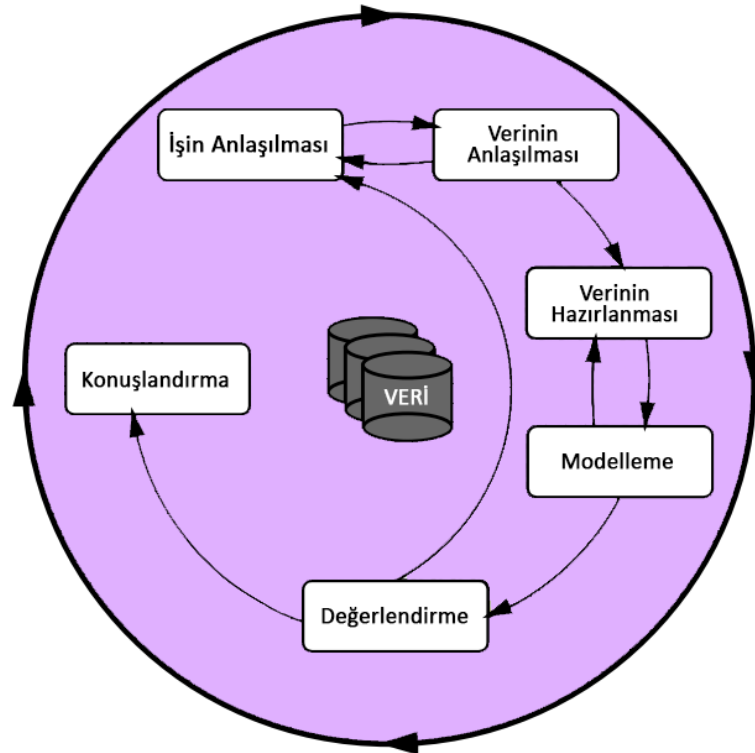
3.2.1. Çapraz Endüstri Standart İşlem Modeli (CRISP-DM)

Makine öğrenmesi sırasında probleme en uygun modeli bulmak önemlidir. Bu yüzden farklı algoritmalar kullanarak en iyisinin hangisi olduğunun bulunması gerekir. Makine öğrenmesi kullanarak veri bilimi projelerinin inceliklerini sağlamak için en iyi bilinen ve en yaygın çerçeveler;

- Sistem Geliştirme Yaşam Döngüsü (SDLC)
- Veri Tabanlarında Bilgi Keşfi (KDD)
- Çapraz Endüstri Standart İşlem Modeli (CRISP-DM)
- Örneklem, Keşfetme, Değiştirme, Modelleme, Değerlendirme (SEMMA)

yöntemleri olmuştur.

Bu çalışmada kullanılan metodoloji CRISP-DM olacaktır. CRISP-DM modeli, Şekil 3.1'de gösterildiği gibi altı adımdan oluşmaktadır.



Şekil 3.1. CRISP-DM Yaşam Döngüsü

3.2.1.1. İşin Anlaşılması

İlk adım projenin hedefi, ihtiyaçlarını tespit etmek ve değerlendirmektir. İlk adım, işi her yönüyle tamamen anlamaktan oluşur: uygulama alanı, proje çıktısı ve gereksinimler yer almaktadır.

3.2.1.2. Verinin Anlaşılması

İlk adımda sorunu tanımladıktan sonra ikinci adımda, soruna uygun verilerin elde edilmesi gerçekleştirilir. Veriler elde edildikten sonra veri kalitesini kontrol etmek, verileri anlamaya çalışmak veya alt veri kümeleri elde edip bunlara göre hipotezler, doktrin çıkarmaktır. Verilere göz gezdirildiğinde bir miktar anlaşılır olması ilk adımdaki problemin formüle edilmesi açısından önemlidir. Verileri anlamlandırabilmek için bu adımda veri görselleştirme ve istatistiksel hesaplama yöntemlerine başvurulabilir.

3.2.1.3. Verinin Hazırlanması

Bir makine öğrenimi modelinin amacı, genelleştirilebilir yapıları belirlemektir. Böylece, belirli bir olayı gözlemleyerek, belirli bir tahmin sonucu elde edilebilir. Veri setindeki verilerin kalitesi çok önemlidir. Kalitesiz veriler var ise, hangi makine öğrenmesi algoritması kullanılırsa kullanılsın elde edilen sonuçların da kalitesiz olmasına yol açabilir.

Verinin hazırlanması adımı, verinin anlaşılması adımıyla devam eder. Toplanan verilerde, istatistiksel hesaplamalar yapıldıktan sonra eksik veri, gürültülü veri ve tutarsız veri var ise bunlar çözülecektir. Veri setinde kategorik veriler var ise bu aşamada sayısal değerlere dönüştürülmesi gerekmektedir. Çizelge 3.1’de verinin hazırlanması adımı yapılan işlemler özetle gösterilmiştir.

Çizelge 3.1. Verinin Hazırlanması Adımında Yapılan İşlemler

Veri Temizleme	Veri Standardizasyonu	Veri Azaltma	Değişken Dönüşümü
Aykırı veri analizi	Normalizasyon	Örnek sayısını azaltma	Sayısal değişken dönüşümü
Eksik veri analizi	Standardizasyon	Değişken sayısını azaltma	Kategorik değişken dönüşümü
Gürültülü veri	Logaritmik Dönüşüm		

3.2.1.3.1. Korelasyon Analizi

Sınıflandırma sürecinde kalitenin artırılması gerekir. Bu nedenle değişkenler arasındaki korelasyon incelenmiştir. Korelasyon analizi, iki veya daha fazla değişkenler arasındaki ilişkidir. Korelasyon analizi, değişkenler arasındaki ilişkiyi, ilişkinin gücü ve yönü açısından tanımlar. Korelasyon kavramını detaylı açıklamak istersek; bir değişken bir köydeki sansarların sayısı olabilir ve diğer değişken köylülerin kümeslerindeki tavuk olabilir. Sansar popülasyonu arttıkça köydeki tavukların sayısı azalır. Bu bir negatif korelasyon örneğidir: bir değişken arttıkça diğeri azalır. Pozitif bir korelasyon, iki değişkenin birlikte artan veya birlikte azalan şekilde tepki vermesidir.

Korelasyon analizi yapılırken genellikle grafiklerden yorumlama yapılır. Çalışmayı yapan kişi, grafikleri gözle inceleyerek, verilerin ne kadar yakın olduğuna bakarak bir korelasyon olup olmadığını anlayabilir. Bir korelasyonun gücü, korelasyon katsayısı ile ölçülür. İki değişkenin değerleri kullanılarak hesaplanan katsayıya Pearson Moment Çarpım Korelasyon Katsayısı denir. Pearson Moment Çarpım Korelasyon Katsayısını geliştiren Karl Pearson'ın anısına konmuştur. Pearson Moment Çarpım Korelasyon Katsayısı r ile ifade edilir.

$$r = \frac{cov_{xy}}{S_x S_y} = \frac{E(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y} \quad (3.1)$$

Denklem 3.1'de r korelasyonunun değerinin nasıl bulunacağını gösterir. COV, kovaryans anlamına gelir. S_x ve S_y , X ve Y vektörlerinin standart sapmalarıdır. $\bar{X}$ ve $\bar{Y}$ ise X ve Y vektörlerinin matematiksel beklentilerini temsil eder. E değişkeni matematiksel beklenti değerini ifade eder. N , X ve Y vektörlerinde bulunan veri sayısıdır. Yanıt sütununu içeren son sütun ile tüm sütunlar korelasyon için analiz edilir. Korelasyon katsayısını yorumlamak istenirse;

$$0.9 < |r| < 1.0, \quad 0.81 < r^2 < 1.00 \quad (3.2)$$

Denklem 3.2'de r değeri 0,9 ile 1,0 arasında olan korelasyon katsayıları, çok yüksek düzeyde ilişkili olarak kabul edilebilecek değişkenleri gösterir.

$$0.7 < |r| < 0.9, \quad 0.49 < r^2 < 0,81 \quad (3.3)$$

Denklem 3.3'te r değeri 0,7 ile 0,9 arasında olan korelasyon katsayıları, yüksek düzeyde ilişkili sayılabilecek değişkenleri gösterir.

$$0.5 < |r| < 0.7, \quad 0.25 < r^2 < 1.00 \quad (3.4)$$

Denklem 3.4’de r değeri 0,5 ile 0,7 arasında olan korelasyon katsayıları, orta düzeyde ilişkili kabul edilebilecek değişkenleri gösterir.

$$0.3 < |r| < 0.5, \quad 0.09 < r^2 < 0.25 \quad (3.5)$$

Denklem 3.5’te r değeri 0,3 ile 0,5 arasında olan korelasyon katsayıları, düşük korelasyona sahip değişkenleri gösterir.

$$0.0 < |r| < 0.3, \quad 0.00 < r^2 < 0.09 \quad (3.6)$$

Denklem 3.6’da r değeri 0,3’ten küçük olan korelasyon katsayıları, varsa çok düşük korelasyonu ifade eder (Calkins, 2005).

Spearman ve Kendall korelasyon yöntemleri de vardır, ancak bu iki yöntem normal dağıtılmamış verilerde kullanılır. Bu nedenle bu çalışmada Pearson yöntemi kullanılacaktır.

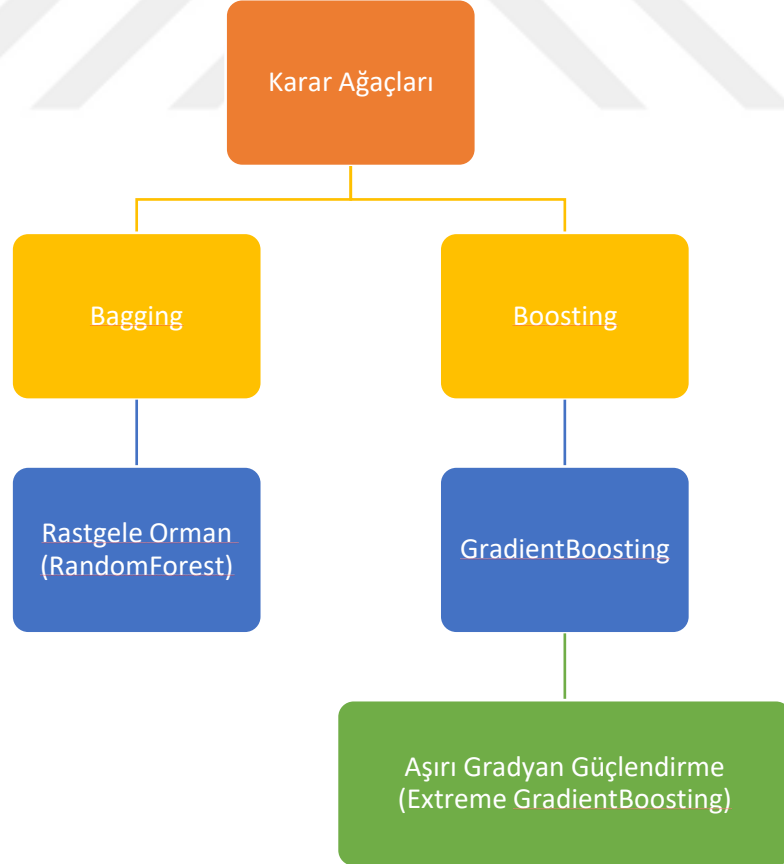
3.2.1.4. Modelleme

Yapılacak uygulamaya ve verilere göre en uygun modeller seçilir. Bu tez kapsamında sınıflandırma çalışması yapılacağı için, sınıflandırma algoritmaları bu aşamada seçilir. Modellere en uygun parametreler seçildikten sonra modelin çalıştırılması gerçekleştirilir. Bu tez çalışmasında, öğrencilerin başarısına en çok etki eden faktörlerin tespiti bulunmak istendiğinden, bir sınıflandırma işlemi yapılacaktır. Sınıflandırma konusunda, en çok tercih edilen sınıflandırma algoritmaları Çizelge 3.2’de verilmiştir.

Çizelge 3.2. Kaynak Araştırmasına Göre Algoritmaların Kullanım Miktarı

Adı	Orijinal Adı	Miktar
Karar Ağacı	Decision Tree (ID3,C4.5, CART)	16
Naive Bayes	Naïve Bayes	10
Çok Katmanlı Algılayıcı	Multi Layer Perceptron (MLP)	5
Azaltılmış Hata Budaması	Reduced Error Pruning (REPTree)	4
Destek Vektör Makineleri	Support Vector Machine (SVM)	4
Yapay Sinir Ağları	Artificial Neural Network (ANN)	4
Sıralı Minimal Optimizasyon	Sequential Minimal Optimization (SMO)	2
Rastgele Orman	Random Forest	2
Sinir Ağı	Neural Network (NN)	1
Karar Tabloları	Decision Table	1
K En Yakın Komşu	K-Nearest Neighbors (KNN)	1

Görüldüğü üzere en çok tercih edilen yöntem Karar Ağacı'dır. Karar ağaçları eski bir yöntemdir ve en büyük problemi aşırı öğrenmedir. Bu durumu aşmak için zamanla yeni karar ağacı yöntemleri ortaya çıkmıştır. Şekil 3.2 'de Karar Ağaçlarının gelişimi verilmiştir.

**Şekil 3.2.** Karar Ağaçlarının Gelişimi

Karar ağaçlarının günümüzdeki son sürümleri olarak görülen Aşırı Gradyan Güçlendirme ve Rastgele Orman algoritmaları modelleme için seçilmiştir. Ayrıca son yıllarda tez ile benzer alanda çalışmalarda kullanılan en çok tercih edilen yöntemlerden olduğu için Destek Vektör Makineleri bu tez çalışması için seçilmiştir.

3.2.1.5. Değerlendirme

Modelin değerlendirmesinde sorunu belirleyen kişinin kendisi ve iş sahibi ile birlikte işin hedefine ulaşıp ulaşılamadığı tespit edilmiştir. Modelleme, en alakalı faktörleri, geliştirilecek yapıya dâhil ederek oluşturulan bir gözden geçirme sürecidir. Gözden geçirme esnasında baştan sona bütün süreç bir daha değerlendirilir.

Karışıklık matrisi, sınıflandırma sisteminin gerçek ve tahmini sınıflandırması hakkında bilgi sağlamaktadır. Model performansı genellikle matristeki veriler kullanılarak hesaplanır (Santra ve Christy, 2012). Şekil 3.3'te karmaşıklık matrisi içeriği gösterilmiştir.

		TAHMİN VERİLERİ	
GERÇEK VERİLER		POZİTİF	NEGATİF
	POZİTİF	GERÇEK POZİTİFLER	YANLIŞ NEGATİFLER
	NEGATİF	YANLIŞ POZİTİFLER	GERÇEK NEGATİFLER

Şekil 3.3. Karmaşıklık Matrisi

Gerçek Pozitifler (TP), gerçek değeri 1 ve tahmin ettiğimiz değerin de 1 olduğu örneklerdir. Gerçek Negatifler (TN), gerçek değeri 0 ve tahmin ettiğimiz değerin de 0 olduğu örneklerdir. Yanlış Pozitifler (FP), gerçek değeri 0 ancak tahmin ettiğimiz değerin 1 olduğu örneklerdir. Yanlış Negatifler (FN), gerçek değeri 1 ancak tahmin ettiğimiz

değerin 0 olduğu örneklerdir. Karışıklık matrisinden hesaplanan ifadeler vardır. TP, TN, FP ve FN'nin birbirleriyle ilişkisini gösteren bu ifadeler aşağıda verilmiştir.

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (3.7)$$

Denklem 3.7 Doğruluk Oranı, Genel olarak, sınıflandırmanın ne kadar sıklıkta doğru tahmin ettiğinin bir ölçüsüdür.

$$\frac{FP + FN}{TP + TN + FP + FN} \quad (3.8)$$

Denklem 3.8 Yanlış Sınıflandırma Oranı, sınıflandırmanın ne kadar sıklıkta yanlış tahmin ettiğinin bir ölçüsüdür.

$$\frac{TP}{TP + FN} \quad (3.9)$$

Denklem 3.9 Gerçek Pozitif Değerlerin Oranı, sınıflandırmanın ne kadar gerçek pozitif değeri doğru tahmin ettiğinin bir ölçüsüdür. Recall olarak da bilinir.

$$\frac{TN}{TN + FP} \quad (3.10)$$

Denklem 3.10 Gerçek Negatif Değerlerin Oranı, sınıflandırmanın ne kadar gerçek negatif değeri doğru tahmin ettiğinin bir ölçüsüdür.

$$\frac{FP}{TN + FP} \quad (3.11)$$

Denklem 3.11 Yanlış Pozitif Değerlerin Oranı, gerçek değeri 0 olmasına karşılık 1 olarak tahmin edilenlerin oranıdır.

$$\frac{FN}{TP + FN} \quad (3.12)$$

Denklem 3.12 Yanlış Negatif Değerlerin Oranı, gerçek değeri 1 olmasına karşılık 0 olarak tahmin edilenlerin oranıdır.

$$\frac{TP}{TP + FP} \quad (3.13)$$

Denklem 3.13 Hassasiyet, Tüm sınıflardan, doğru olarak ne kadar tahmin edildiğinin bir ölçüsüdür. Mümkün olduğu kadar yüksek olmalıdır. Precision olarak da bilinir.

$$\frac{2 * \left(\frac{TP}{TP + FP} \right) * \left(\frac{TP}{TP + FN} \right)}{\left(\frac{TP}{TP + FP} \right) + \left(\frac{TP}{TP + FN} \right)} \quad (3.14)$$

Denklem 3.14 F Score, gerçek pozitif değerlerin oranı ve hassasiyetin harmonik ortalamasıdır. Sınıflandırmanın ne kadar iyi performans gösterdiğinin bir ölçüsüdür ve sınıflandırmaların karşılaştırılmasında kullanılır.

Modellerin karmaşıklık matrisi çıkarılacaktır. Karışıklık Matrisi, mevcut duruma ve sınıflandırma modeline ilişkin doğru ve yanlış tahminlerin sayısını gösterir. Öğrenmenin ne kadar yüksek olduğunu görmemizi sağlar. Matris, tahmindeki hedeflerin sayısına bağlıdır.

3.2.1.6. Konuşlandırma

Bu aşamada en iyi sonuç veren modele ham veriler verilecek ve model izlenmeye başlanacaktır. Bu süreçte başlangıçta ortaya çıkmayan faktörler, yapısal değişiklikler ortaya çıkabilir. Bunlara göre modelde gerekli değişiklikler yapılacaktır. Proje değerlendirmesi yapılır. Proje değerlendirmesinin bir parçası olarak, proje değerlendirmesi gerçekte neyin doğru yapıldığını, neyin yanlış olduğunu ve neyin daha iyi yapılabileceğini değerlendirir. Çizelge 3.3’de CRISP-DM adımları özetlenmiştir.

Çizelge 3.3. CRISP-DM İş Akışına Bir Genel Bakış

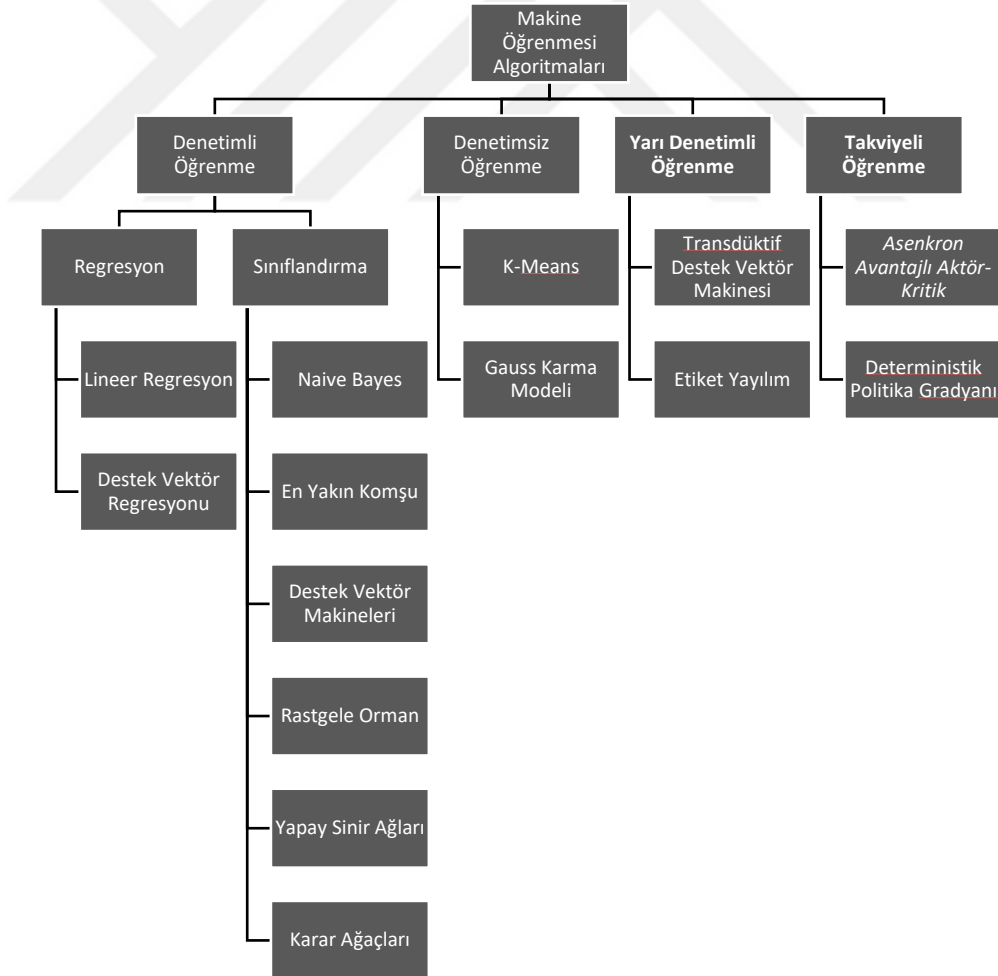
İşin Anlaşılması	Verinin Anlaşılması	Verinin Hazırlanması	Modelleme	Değerlendirme	Konuşlandırma
İş Hedeflerini Belirleme İş hedeflerinin arka planı İş başarısının kriterlerini belirleme Durumu Değerlendirme Kaynak envanteri Gereksinimler, varsayımlar ve kısıtlamalar Riskler ve beklenmedik durumlar Terminoloji Maliyetler ve faydalar Veri Madenciliği Hedeflerini Belirleme Veri madenciliği hedefleri Veri madenciliği başarısı için kriterler Proje Planı Üretme Proje planı üretme Araçlar ve teknikler ile İlk değerlendirme	İlk Verileri Toplama İlk veri toplama raporu Verileri Tanımlama Veri açıklama raporu Verileri Keşfetme Veri keşif raporu Veri Kalitesini Doğrulama Veri kalitesi raporu	Veri Kümesi Veri kümesi açıklaması Veri Seçme Dahil etme / Hariç tutma Gerekçesi Verileri Temizle Veri temizleme raporu Veri Düzenleme Türetilmiş nitelikler oluşturulan kayıtlar Verileri Entegre Et Verileri birleştir Verileri Biçimlendir Birleşmiş verileri yeniden biçimlendir	Model Seçme Modelleme tekniği Modelleme varsayımları Test Tasarımı Oluşturma Test tasarımı Model İnşa Etme Modellerin parametresini ayarlama Model Açıklaması Modeli Değerlendirme Model değerlendirmesi Model parametrelerini revize etme	Sonuçları Değerlendirme Veri madenciliği sonuçlarının değerlendirilmesi, iş başarı kriterleri, onaylı Modeller Gözden Geçirme Süreci Sürecin Gözden Geçirilmesi Sonraki Adımları Belirleme Olası Eylemlerde Aksiyon Almak İçin Liste Tutma	Konuşlandırmayı Planlama Konuşlandırma planı İzleme ve Bakım Onarım Planlama İzleme ve bakım planı Sonra Rapor Hazırlama Son rapor Son sunum Projeyi Gözden Geçir Tecrübeni doküman haline getir.

3.3 Makine Öğrenmesi Yöntemleri

Samuel (1959). makine öğrenmesini “Bilgisayarların açıkça programlanmadan öğrenmesini sağlayan bir çalışma alanı” olarak tanımlamıştır. Mitchell ve McGraw (1997) mühendislik alanında daha anlamlı olan yeni tanım ortaya koymuşlardır;

Bir bilgisayar programı, T görevlerdeki, P performansını E tecrübesi ile artırırsa, bir bilgisayar programı, bazı T görevleri, p performans ölçüsüne göre E deneyimlerinden öğrenerek yerine getirir.

Günümüzde makine öğrenmesi algoritmaları, genellikle araştırmacıların kendi kendilerine öğrenmelerine izin vererek veri kümelerinden modelleri keşfetmek için kullandıkları yöntemlerdir (Asyraf ve ark., 2017). Üstelik büyük ölçekli hesaplama tekniklerinin gelişmesiyle, veri öğrenmesi kolaylaşmaktadır. Makine öğrenimi, denetimli öğrenme ve denetimsiz öğrenme gibi çeşitli öğrenme türleri olarak sınıflandırılabilir. Şekil 3.4’te, makine öğrenmesi algoritmalarının türlerini göstermektedir.



Şekil 3.4. Makine Öğrenmesi Algoritması Türleri

Denetimli öğrenme etiketli verilerle, denetimsiz öğrenme etiketlenmemiş verilerle çalışırken, yarı denetimli öğrenme her iki tür veriyle de çalışır. Denetimli öğrenmede asıl amaç girdi ve çıktı değerlerine dayalı formül oluşturmaktır. Denetimsiz Öğrenmede ise girdi değerleri arasında bir ilişki aranmaktadır. Takviyeli öğrenmede, bir vekil ajanı çevresiyle etkileşime girer ve etkileşimden aldıkları geri bildirimden öğrenir.

Sınıflandırma genelde etiketler ve kategoriler olarak bilinen ayrık değişkenler için uygulanabilir. Kısaca bir sınıf ve kategori, gözlem ile tahmin edilir. Bir sınıftaki öğrencilerin performansı “iyi” veya “kötü” olarak sınıflandırılabilir.

Böyle bir durumda en uygun sınıflandırma algoritmasını seçmek, öğrenci performansını etkileyen faktörlerin tespitinde en faydalı sonuçları sağlayacaktır.

Bu çalışmada;

- Aşırı Gradyan Güçlendirme Sınıflandırma
- Rastgele Orman Sınıflandırma
- Destek Vektör Makineleri Sınıflandırma

algoritmaları kullanılmıştır.

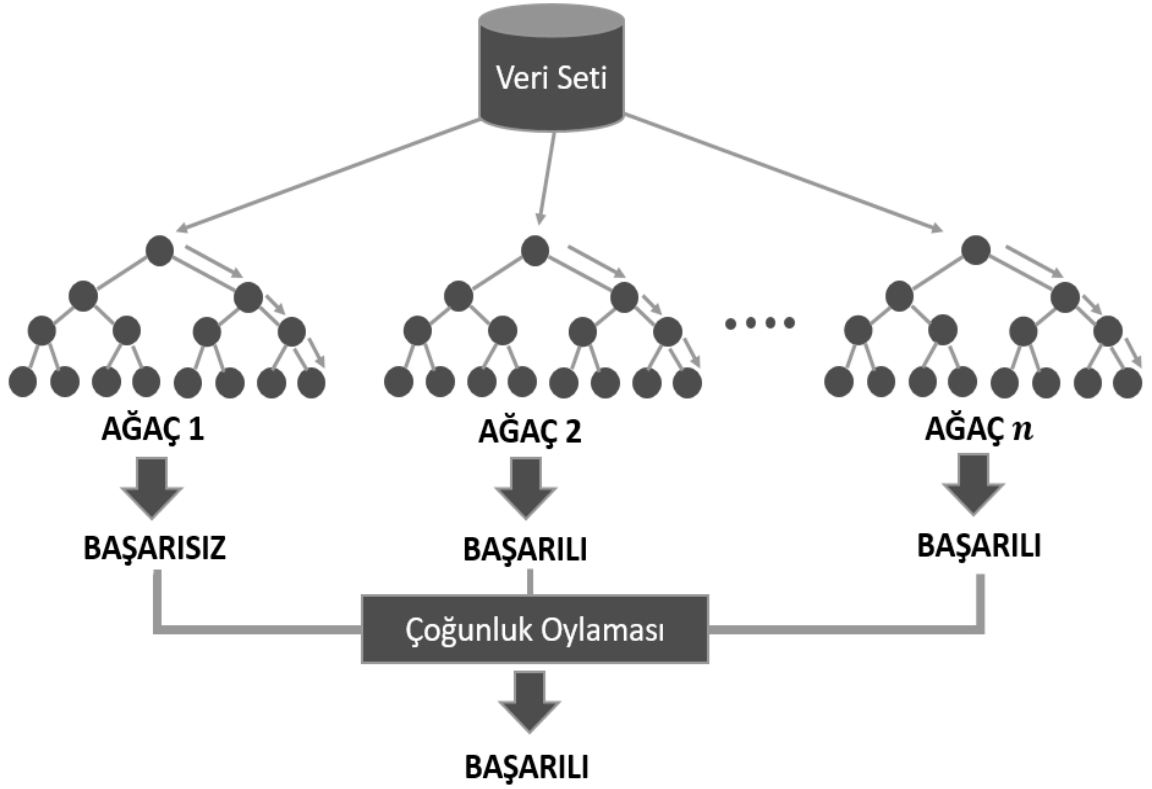
3.3.1. Rastgele Orman Algoritması

Ho (1998), her bir ağacı büyütmek için kullanılacak özelliklerin rastgele bir alt kümesini hesaplayan “Rastgele Altuzay” yöntemi hakkında çalışma yapmıştır.

Amit ve Geman (1997), çok sayıda geometrik özellik üzerinde bir çalışma yürüttüler ve her bir düğüm için ayrılacak en iyi elemanı rastgele seçtiler. Breiman (2001) yapmış olduğu çalışmada, 1996 yılında kendi yaptığını torbalama tekniği ve Ho’nun yapmış olduğu çalışma ile birlikte, Amit ve Geman’ın çalışmasından da ilham alarak Rastgele Orman Sınıflandırma algoritmasını geliştirmiştir.

Algoritma, her sınıflandırıcının girdi vektörüne bağlı olmayan bir karar ağacından oluşmaktadır. Akıştaki her ağaç, girdi vektörünü sınıflandırmak için bir birim oy gerçekleştirir (Pal, 2005). En çok kullanılan geleneksel yöntemlerden biri olan karar ağaçlarının en büyük problemi aşırı öğrenmedir.

Rastgele Orman algoritması, özellik kümesinden rastgele birçok alt küme seçerek bu sorunun üstesinden gelir. Üstelik her bir karar ağacının kullandığı veri seti farklı olduğu için veri aykırı değer sorunları en aza indirilir. Seçilen alt kümeden elde edilen sonuçlardan, en fazla birim oyunu alan değer, Rastgele Orman algoritmasının sonucunu oluşturur. Rastgele örnekleme ve topluluk yöntemlerindeki tekniklerin iyileştirilmiş özelliklerini içermesi nedeniyle Rastgele Orman yöntemi daha iyi genellemeler ortaya koyar ve sağlam tahminlerde bulunur (Qi, 2012). Şekil 3.5’de Rastgele Orman Algoritmasının çalışma mantığı gösterilmiştir.



Şekil 3.5. Rastgele Orman Algoritmasının Çalışma Mantığı

Rastgele Orman algoritmasının sınıflandırma performansı ağaç sayısı arttıkça artmaktadır (Breiman, 2001). Bununla birlikte, yapılan çalışmalar dikkate alındığında, Rastgele Orman algoritması, başlangıçta belirtilen ağaç sayısını aşıyorsa, sınıflandırma performansını önemli ölçüde artırmadığı görülmüştür (Oshiro ve ark., 2012).

Rastgele Orman Algoritması, kategorik özelliklerin olduğu düzensiz veri setlerinde iyi sonuçlar vermektedir. Bu çalışma kapsamında veri setinde kategorik veri oldukça yer aldığından bu yöntem tercih edilmiştir.

$$\{h(x, \theta_k) | k = 1, \dots\} \quad (3.15)$$

Denklem 3.15'te girdi verisi olarak x ve rastgele vektörü ise θ_k ile temsil etmektedir (Breiman, 2001; Pal ve Chaurasia, 2014).

$$\{\sum_j \neq i (f(C_i, T)/|T|)(f(C_j, T)/|T|) \quad (3.16)$$

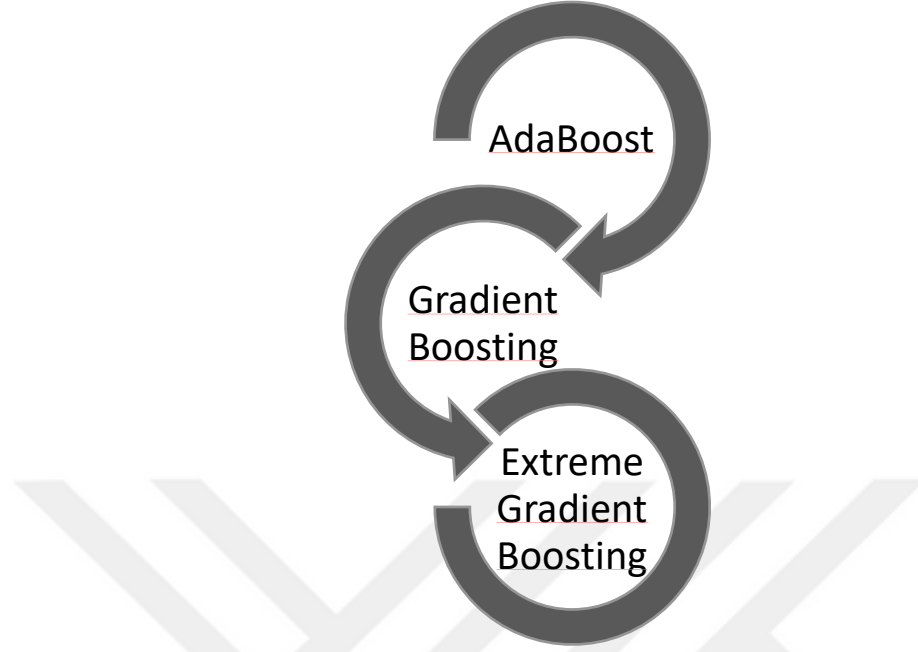
Rastgele orman algoritması mevcut dallar içinde en iyi dalı belirlemek için Denklem 3.16'da verilen Gini indeksini kullanır (Pal, 2005). Her düğümde rastgele seçtiği değişkenler için Gini indeksini hesaplar. Gini indeksinin küçük olduğu değişkeni seçerek diğer düğüme geçer. Gini sıfır olması durumunda ilgili düğüm tamamen homojendir ve dallanma orada biter (Gislason ve ark., 2006). T eğitim veri kümesini, C_i verinin ait olduğu sınıfı, $f(C_i, T)/|T|$ seçilen verinin C_i sınıfına ait olma olasılığını gösterir (Gislason ve ark., 2006; Qi, 2012). Rastgele orman algoritması rastgele örnekleme ve topluluk algoritmaları tekniklerinin iyileştirilmiş bir yapıyı içermesi nedeniyle diğer sınıflandırma algoritmalarına göre daha iyi genelleme ve doğru tahminlerde bulunur (Qi, 2012).

3.3.2. Aşırı Gradyan Güçlendirme Algoritması

Geliştirme, öğrenme algoritmalarının doğruluğunu artırmak için kullanılan yaygın bir tekniktir. Çok doğru veya güçlü sınıflandırma araçlarının, hatalı veya zayıf birçok sınıflandırma aracının doğrusal birleşimi oluşması ilkesine dayanmaktadır (Ong ve Bowden, 2004). Gradyan Güçlendirme yöntemi, eğitilmiş öğrenmede yanlış sınıflandırma hatalarıyla tek bir güçlü sınıflandırma aracı oluşturmak için birden çok sınıflandırma aracını birleştiren bir yaklaşıma dayanır. Veri örnekleri, her bir yineleme dizisinde oluşturulan temel sınıflandırma araçlarının ağırlıklı tahminlerini oylayarak sınıflandırılır (Farid ve ark., 2013).

Güçlendirme sırasında, her eğitim örneğine bir ağırlık atanır ve artırma her tekrarlandığında, ağırlık yeni değere göre değişir. Yeni bir eğitim seti almak için eğitim örneğinin dağılımına bağlı olarak bir değer alır. Sınıflandırma aracı, eğitim veri setindeki değerlerden oluşturulur ve orijinal veri kümesindeki tüm eğitim örneklerini sınıflandırmak için kullanılır. Her geliştirme sonrasında eğitimdeki örneklerin ağırlıkları güncellenir. Yanlış sınıflandırılan eğitim örneklerinin ağırlıkları güçlendirilirken, doğru sınıflandırılan örneklerin ağırlıkları düşürülür. Bu, sınıflandırma aracının sürekli yinelemelerde sınıflandırılması zor olan eğitim örneklerine odaklanmasını sağlar. Artışın

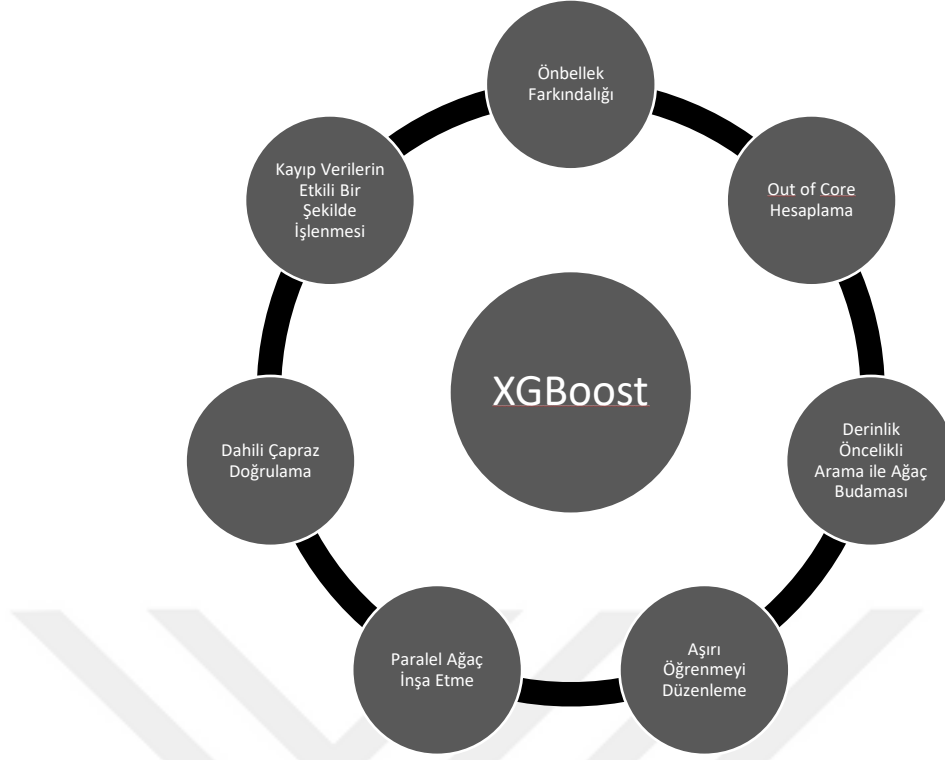
önündeki en büyük engel, aşırı öğrenmenin ortaya çıkmasıdır. Şekil 3.6’da güçlendirme algoritmalarının gelişimi gösterilmektedir.



Şekil 3.6. Güçlendirme Algoritmalarının Gelişimi

Morde ve Brownlee’ye göre Aşırı Gradyan Güçlendirme Sınıflandırma algoritması Gradyan güçlendirme çerçevesini kullanan karar ağacı tabanlı denetimli öğrenme temelli makine öğrenmesi algoritmasıdır (Brownlee, 2016; Morde, 2019). Aşırı Gradyan Güçlendirme, her biri K düğümlerine sahip Sınıflama ve Regresyon Ağacı topluluğu kullanır. Son tahmin, her ağaç için tahmin puanlarının toplamıdır. Aşırı Gradyan Güçlendirme, kullanım kolaylığı ve tahmin etme yönünün güçlü olması sebebiyle, büyük veri problemlerinin çözülmesi kapsamında yarışmaların düzenlendiği Kaggle platformunda son yıllarda en popüler algoritmalarından biri olmuştur (Morde, 2019). Aşırı Gradyan Güçlendirme kısaca XGBoost olarak ifade edilir. Aşırı Gradyan Güçlendirme algoritması, Sınıflandırma ve Regresyon ağaçları algoritmasını kullanır. Aşırı Gradyan Güçlendirme, sistem optimizasyonu ve algoritmik geliştirmeler yoluyla temel Gradyan Güçlendirme Makineleri (GBM) çerçevesini geliştirir (Morde, 2019).

Şekil 3.7’de Aşırı Gradyan Güçlendirme algoritmasının Gradyan Güçlendirme Makinelerini en uygun hale getirdiği yönleri gösterilmiştir.



Şekil 3.7. XGBoost'un GBM'yi Geliştirdiği Yönler

$$L^t = \sum_{i=1}^n l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \quad (3.17)$$

Denklem 3.17'de t'inci iterasyonun amaç fonksiyonunu L^t , i örneğinin gerçek sınıf etiketini y_i , i örneğinin tahmin edilen sınıf etiketini $\hat{y}_i$ karar ağacın fonksiyonunu f_t , eğitim veri setindeki örnek sayısını n ve düzenleme terimini Ω temsil etmektedir.

$$\Omega(f_t) = \gamma T_t + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (3.18)$$

Denklem 3.18'de γ ve λ hiperparametreleri, T ağaçtaki yaprak sayısı ve w her bir yaprağın ağırlığı göstermektedir. $\Omega(f_t)$ fonksiyonu aşırı öğrenmeyi önlemek kullanılır.

$$L^t \cong \sum_{i=1}^n \left[l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (3.19)$$

$$g_i = \partial_{\hat{y}_i^{t-1}}^2 l(y_i, \hat{y}_i^{t-1}) \quad (3.20)$$

$$h_i = \partial_{\hat{y}_i^{t-1}}^2 l(y_i, \hat{y}_i^{t-1}) \quad (3.21)$$

Denklem 3.19'daki hata fonksiyonunun değerini yaklaşık olarak tahmin etmek ikinci dereceden Taylor açılımı uygulanır, burada Denklem 3.20'deki g_i hata fonksiyonun birinci dereceden gradyanını temsil ederken Denklem 3.21'deki h_i , hata fonksiyonun ikinci dereceden gradyanını temsil eder.

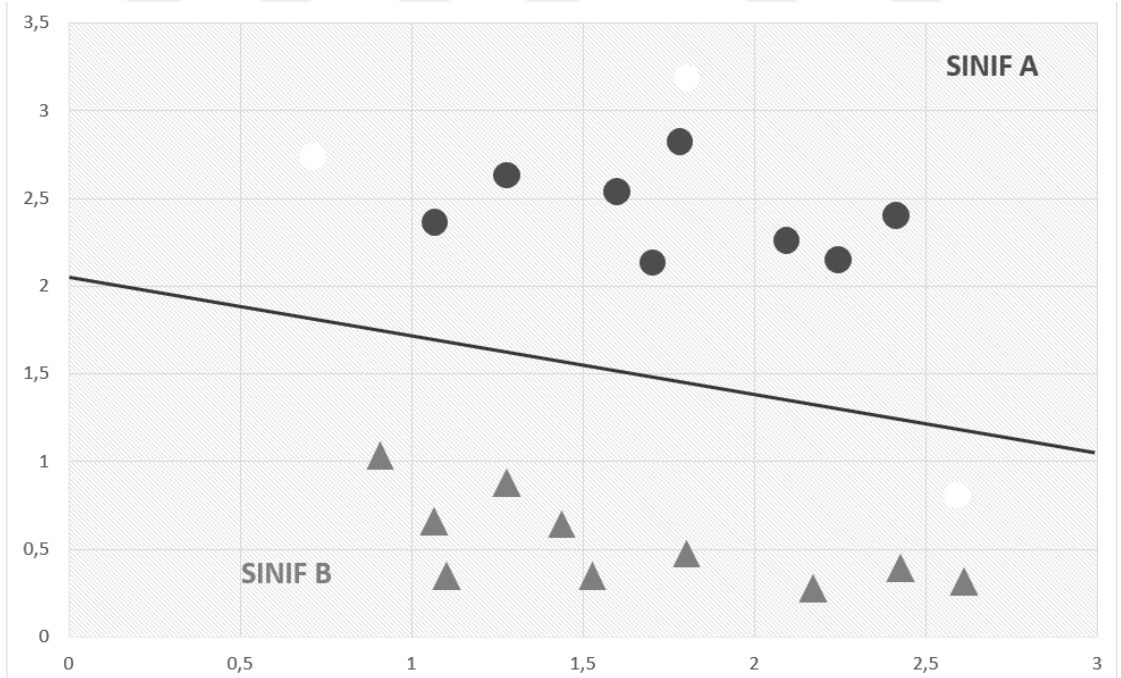
$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (3.22)$$

$$L(q)w_j^* = -\frac{1}{2} \sum_{j=1}^T \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \quad (3.23)$$

Sabit bir q ağaç yapısı için, j .düğümündeki en uygun yaprak ağırlığı ve buna karşılık gelen amaç fonksiyonunun en uygun değeri sırasıyla Denklem 3.22’de ve 3.23’te verilmiştir. I_j , j yaprağının örnek kümesini ifade eder. Denklem 3.23’te ağaç yapısının q ’nun kalitesini değerlendirmek için bir puanlama fonksiyonu olarak kullanılır (Aydın, 2021).

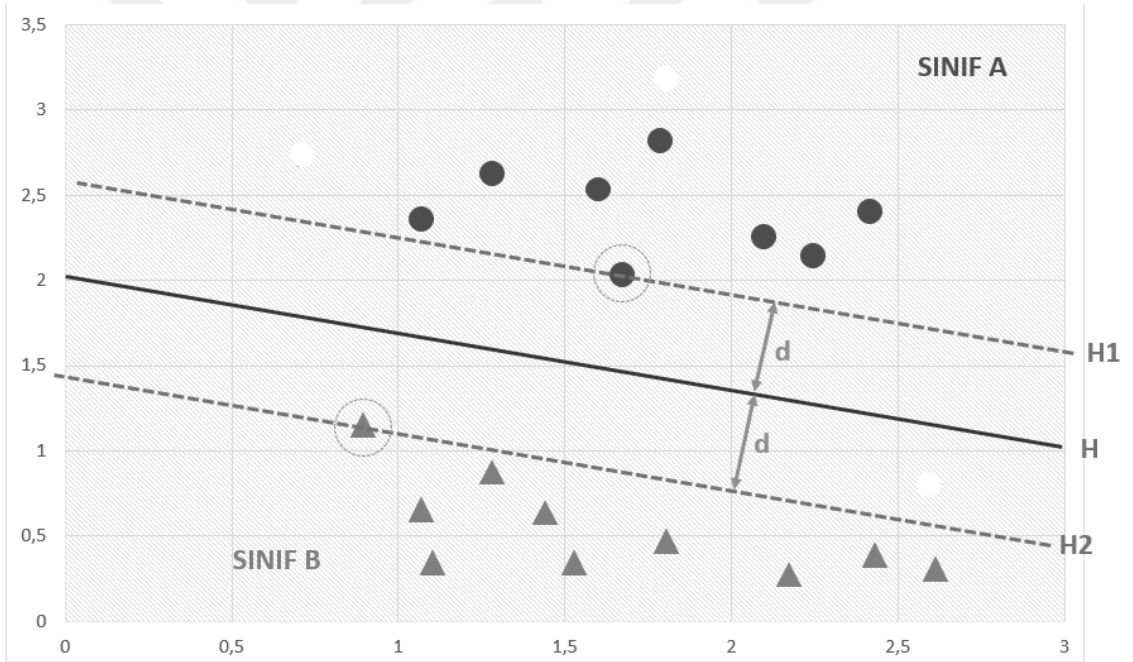
3.3.3. Destek Vektör Makineleri Algoritması

Destek Vektör Makinesi (DVM), öğrenci performansını tahmin etmek için kullanılan ünlü tahmin yöntemlerinden biridir. Destek Vektör Makineleri özellik vektörlerini çok boyutlu bir uzaya eşlemek için doğrusal olmayan bir yöntem kullanmaktadır ve bu yeni uzayda doğrusal bir sınıflandırıcı olarak kullanılmaktadır. Şekil 3.8’de Destek Vektör Makinesi ile iki sınıfa ayrılmış bir düzlem gösterilmiştir.



Şekil 3.8. DVM ile İki Sınıfa Ayrılmış Düzlem

Destek Vektör Makinesi genelde iki sınıflı problem için kullanılır. Şekil 3.8'de gösterildiği gibi giriş verileri, bir destek vektörü olarak tanımlanabilen bir hiper düzlem ile iki kısma ayrılır. Buradaki hedef uygun bir hiper düzlem bularak iki sınıfı birbirinden ayırmaktır. Pek çok düzlem eğitim verileri üzerinde eşdeğer bir şekilde performans gösterse de, yeni veriler üzerindeki genelleme performansı önemli bir şekilde farklılık gösterebilir (Kulkarni ve Harman, 2011). Optimal hiper düzlemi bulmak için, her iki sınıfın optimal hiper düzlemlerini çizmek için o sınıflara ait en yakın veriler bulunur. En yakın noktalar üzerinden geçecek şekilde hiper düzlemler çizilir ve çizilen iki düzlem birbirine paraleldir. Hiper düzlemler aralarındaki mesafe hiper düzlemlerin kalitesini ifade eder. DVM iki sınıf arasındaki mesafe maksimum olacak şekilde yüzeyi belirler. Bulunan en yakın noktalar destek vektörü olarak ifade edilir. Hiper düzlemin bir tarafı +1, diğer tarafında ise -1 olarak etiketlenir.



Şekil 3.9. Doğrusal Olarak İki Hiper Düzleme Ayrılabilen Destek Vektörleri

Şekil 3.9'de gözüktüğü gibi en yakın noktalar bulunarak H1 ve H2 hiper düzlemleri çizilmiştir. H1 hiper düzlemi ile H2 hiper düzlemi arasında H optimal hiper düzlemi çizilmiştir.

$$w \cdot x + \theta_0 = 0 \quad (3.24)$$

H hiper düzleminin denklemi vektörü Denklem 3.24'teki gibi ifade edilebilir.

$$w \cdot x + \theta_0 - d = 0 \quad (3.25)$$

Denklem 3.24'teki H hiper düzlemi formülüne d mesafesini eklersek Denklem 3.25'teki H1 düzlemi elde edilir.

$$w \cdot x + \theta_0 + d = 0 \quad (3.26)$$

Denklem 3.24'teki H hiper düzlemi formülüne d mesafesini çıkarırsak Denklem 3.26'daki H2 düzlemi elde edilir.

H hiper düzlemi ile en kısa mesafenin olduğu +1 etiketli veriler arasındaki mesafe uzunluğu d, H hiper düzlemi ile en kısa mesafenin olduğu -1 etiketli veriler arasındaki mesafe uzunluğu d olduğu için $d_{+1} + d_{-1}$ toplamı 2d marjin olarak kabul edilir. d mesafesi 1 olarak kabul edilirse;

$$w \cdot x + \theta_0 \geq +1 \quad (3.27)$$

$$w \cdot x + \theta_0 \leq -1 \quad (3.28)$$

Denklem 3.27 ve 3.28'den $y = (-1, +1)$ bulunur. İki denklem tek denkleme indirgenmek istenir ise;

$$y(w \cdot x + \theta_0) \geq 1 \quad (3.29)$$

Denklem 3.29 denklemi elde edilerek, formül tek duruma indirgenmiştir.

4. ARAŞTIRILAN YÖNTEM VE DENEYSEL SONUÇLAR

Bu tez çalışmasında, makine öğrenmesi algoritmaları kullanılarak öğrencilerin akademik başarısını etkileyen faktörlerin tespiti yapılmıştır. Çalışmada başarıya etki eden faktörlerin tespitinde kullanılan modeller ile ilgili bilgiler alt başlıklarda verilmiştir. Sınıflandırma işlemi yapılırken öncelikle hedef sınıf belirlenir. Bu sınıf, diğer özelliklerle ilişkili olarak formüle edilir. Çalışma kapsamında verinin hazırlamasından sonra önerilen modeller; Rastgele Orman Algoritması, Aşırı Gradyan Güçlendirme ve Destek Vektör Makineleridir.

Veri setleri ile elde edilen model çalışmaları sonucunda, Ailedeki kişi miktarı, okula gelme süresi ve önceki dönem notları gibi girdi verileri arasından hangisinin yılsonu notlarının tespitinde etkili olduğu konusunda çalışılmıştır. Sonraki bölümlerde, CRISP-DM metodolojisine uygun şekilde uygulama süreci açıklanmıştır.

4.1. Araştırmanın CRISP-DM Yöntemine Göre Gerçekleştirilmesi

4.1.1. İşin Anlaşılması

Üniversite öğrenimi insan hayatının en önemli dönüm noktalarından biridir. Son yıllarda öğrenciler istedikleri bilgiye kolayca ulaşabilmektedir. Ülkemiz 2020-2021 yılı istatistiklerine göre 8 milyon 240 bin 997 üniversite öğrencisine sahiptir. 2019-2021 yılı istatistiklerine göre ise 7 milyon 940 bin 133 üniversite öğrencisi vardır (Yök, 2020) Buradaki veriler göz önünde bulunduğunda 300 bin civarında öğrenci sayısında artış olmuştur. Bu sayılar ülkemiz açısından oldukça yüksektir ve bu durumda öğrencinin aldığı eğitim kalitesi önemlidir.

Gün geçtikçe mezun olmayı başaramayan öğrenci sayıları artmaktadır. Başta aileler ve akademisyenler, akademik anlamda başarıyı arttırmak için, öğrencinin bir an önce başarılı bir şekilde mezun olmasına katkı sağlayacak faktörleri tespit etmelidirler. Üniversiteler, halen öğrenciliği devam ve mezun olmuş öğrencilerine yol göstermelidir. Akademik yapı çerçevesi hangi öğrencilerin okulu bırakma tehlikesiyle karşı karşıya olduğuna bakma konusunda yöneticilere gerçekten yardımcı olmak için oluşturulmamıştır (Márquez-Vera ve ark., 2013).

Bu araştırmanın amacı makine öğrenmesi algoritmaları kullanarak öğrencilerin akademik başarısını etkileyen faktörlerin tespit edilmesi ve mevcut durumu en iyi ele alan makine öğrenmesi modelini tespit etmektir. Bu amaçla aşağıdaki soruların cevabı aranmıştır;

- i) Öğrencilerin akademik başarısını etkileyen önemli faktörler nelerdir?
- ii) Başarıyı etkileyen faktörlerin tespitinde kullanılan modellerin doğrulukları nelerdir?
- iii) Hangi model ne kadar başarılı olmuştur?
- iv) Benzer çalışmalarda elde edilen sonuçlar ile bu çalışmadaki gözlenen sonuçlar ne düzeyde benzerlik göstermektedir?

4.1.2. Verinin Anlaşılması

Araştırma çerçevesinde iki adet veri seti kullanılmıştır. İlk veri seti olarak Dahan ve arkadaşlarının yapmış oldukları Hindistan'ın Assam eyaletinde gerçekleştirilen bir araştırmadan elde edilen veriler kullanılmıştır. Veri tabanındaki veriler üç farklı kolejden toplanmıştır: Duliajan Koleji, Doomdooma Koleji ve Assam'daki Digboi Koleji. İkinci veri setinde Cortez ve Silva' nın yapmış oldukları Portekiz'in Alentejo bölgesinden 2005 yılında iki devlet okulundan toplanmıştır: Gabriel Pereira ve Mousinho da Silveira. Tezin bundan sonraki kısımlarında ilk veri setine Veri Seti I, ikinci veri setine Veri Seti II denilecektir.

4.1.2.1. Veri Seti I

Veri seti 22 özellikten oluşan 300 adet örnek içermektedir. Çizelge 4.1'de verinin özellikleri gösterilmektedir. Veri setinde her öğrenci bir sıra ile temsil edilmektedir.

Çizelge 4.1. Veri Seti I Özellikleri

Sıra	Kolon Adları
1	Cinsiyet
2	Kast Sınıfı
3	10.Sınıf Ortalama Yüzdesi
4	12.Sınıf Ortalama Yüzdesi
5	İç Değerlendirme Ortalama Yüzdesi
6	Dönem Sonu Ortalama Yüzdesi
7	Önceki Dönemden Kalan Dersi
8	Medeni Hal Durumu
9	Şehir Veya Köyde Yaşamı
10	Öğrenci Okula Giriş Tipi
11	Aile Aylık Geliri
12	Aile Genişliği
13	Baba Eğitim Seviyesi
14	Anne Eğitim Seviyesi
15	Baba Mesleği
16	Anne Mesleği
17	Arkadaş Sayısı
18	Ders Çalışma Süresi
19	10.Sınıf'a Geldiğinde Geldiği Okul Tipi
20	Sınav Dili
21	Evden Üniversiteye Seyahat Süresi
22	Sınıfa Devam Yüzdesi

Çizelge 4.2’de cinsiyet kolonunun özellikleri verilmiştir. Cinsiyet kolonu “Erkek” ve “Kadın” değerlerini içermektedir.

Çizelge 4.2. Veri Seti I Cinsiyet Kolonunun Özellikleri

Cinsiyet
Erkek
Kadın

Çizelge 4.3’de kast sınıfı kolonunun özellikleri verilmiştir. Kast sınıfları kategorisi;

- MOBC, Azınlıklar ve diğer geri kalmış topluluk
- OBC, Diğer geri kalmış topluluk
- ST, Planlanmış kabileler topluluğu
- SC, Planlanmış kast topluluğu
- G, Genel topluluk

değerlerini içerir. Kast sınıfı kategorileri Hindistan Anayasasına dayanmaktadır.

Çizelge 4.3. Veri Seti I Kast Sınıfı Kolonunun Özellikleri

Kast Sınıfı
Genel
SC
ST
OBC
MOBC

Çizelge 4.4’de 10.Sınıf ortalaması yüzdesi kolonunun özellikleri verilmiştir. Öğrenci %80 'den daha fazla puan alıyorsa "En İyi", %80 'den az ve %60’dan fazla puan alıyorsa "Çok iyi", %60'den az ve %45’den fazla puan alıyorsa "İyi", %45'den az ve %30’dan fazla puan alıyorsa "Geçer", %30’dan aşağı puanı var ise “Başarısız” olarak ifade edilir.

Çizelge 4.4. Veri Seti I 10.Sınıf Ortalaması Yüzdesi Kolonunun Özellikleri

10.Sınıf Ortalaması Yüzdesi
En İyi
Çok İyi
İyi
Geçer
Başarısız

Çizelge 4.5’de 12.Sınıf ortalaması yüzdesi kolonunun özellikleri verilmiştir. Öğrenci %80 'den daha fazla puan alıyorsa "En İyi", %80 'den az ve %60’dan fazla puan alıyorsa "Çok iyi", %60'den az ve %45’den fazla puan alıyorsa "İyi", %45'den az ve %30’dan fazla puan alıyorsa "Geçer", %30’dan aşağı puanı var ise “Başarısız” olarak ifade edilir.

Çizelge 4.5. Veri Seti I 12.Sınıf Ortalaması Yüzdesi Kolonunun Özellikleri

12.Sınıf Ortalaması Yüzdesi
En İyi
Çok İyi
İyi
Geçer
Başarısız

Çizelge 4.6’da İç Değerlendirme ortalaması yüzdesi kolonunun özellikleri verilmiştir. Öğrenci %80 'den daha fazla puan alıyorsa "En İyi", %80 'den az ve %60’dan fazla puan alıyorsa "Çok iyi", %60'den az ve %45’den fazla puan alıyorsa "İyi", %45'den

az ve %30'den fazla puan alıyorsa "Geçer", %30'dan aşağı puanı var ise "Başarısız" olarak ifade edilir. İç değerlendirme yapılırken okullar ve öğretmenler, öğrencilerin performanslarını değerlendirmektedir.

Çizelge 4.6. Veri Seti I İç Değerlendirme Ortalaması Yüzdesi Kolonunun Özellikleri

İç Değerlendirme Ortalaması Yüzdesi
En İyi
Çok İyi
İyi
Geçer
Başarısız

Çizelge 4.7'de Dönem sonu ortalaması yüzdesi kolonunun özellikleri verilmiştir. Öğrenci %80 'den daha fazla puan alıyorsa "En İyi", %80 'den az ve %60'dan fazla puan alıyorsa "Çok iyi", %60'den az ve %45'den fazla puan alıyorsa "İyi", %45'den az ve %30'den fazla puan alıyorsa "Geçer", %30'dan aşağı puanı var ise "Başarısız" olarak ifade edilir. Dönem sonu ortalaması yüzdesi kolonu bu veri setinde çıktı sınıfıdır.

Çizelge 4.7. Veri Seti I Dönem Sonu Ortalaması Yüzdesi Kolonunun Özellikleri

Dönem Sonu Ortalaması Yüzdesi
En İyi
Çok İyi
İyi
Geçer
Başarısız

Çizelge 4.8'de önceki dönemden kalan dersi kolonunun özellikleri verilmiştir. Önceki dönemden dersi var ise "Evet" yok ise "Hayır" değerlerini içermektedir.

Çizelge 4.8. Veri Seti I Önceki Dönemden Kalan Dersi Kolonunun Özellikleri

Önceki Dönemden Kalan Dersi
Evet
Hayır

Çizelge 4.9'da medeni hal durumu kolonunun özellikleri verilmiştir. Medeni durumu evli ise "Evli" bekâr ise "Bekâr" değerlerini içermektedir.

Çizelge 4.9. Veri Seti I Medeni Hal Durumu Kolonu Özellikleri

Medeni Hal
Evli
Bekâr

Çizelge 4.10’da şehir veya köyde yaşanması durumu kolonunun özellikleri verilmiştir. Yaşanan yer köy ise “Köy”, şehir merkezi ise “Şehir” değerlerini içermektedir.

Çizelge 4.10. Veri Seti I Şehir veya Köyde Yaşanması Durumu Kolonunun Özellikleri

Şehir veya Köyde Yaşanması
Şehir
Köy

Çizelge 4.11’de Öğrenci okula giriş tipi kolonunun özellikleri verilmiştir. Eğer öğrenci koleje girerken ücret ödemişse “Ücretli”, ödememiş ise “Bedava” değerlerini içermektedir.

Çizelge 4.11. Veri Seti I Öğrenci Okula Giriş Tipi Kolonunun Özellikleri

Öğrenci Okula Giriş Tipi
Bedava
Ücretli

Çizelge 4.12’de aile aylık geliri kolonunun özellikleri verilmiştir. Öğrenci ailesi 30000 ₹’den gelir kazanıyorsa "Çok Yüksek", 30000 ₹’den az ve 20000 ₹’den fazla geliri var ise "Yüksek", 20000 ₹’den az ve 10000 ₹’den fazla gelir kazanıyorsa "Ortalamanın Üstü", 10000 ₹’den az ve 5000 ₹’den gelir kazanıyorsa "Ortalama", 5000 ₹’den aşağı gelir kazanıyorsa “Düşük” olarak ifade edilir.

Çizelge 4.12. Veri Seti I Aile Aylık Geliri Kolonunun Özellikleri

Aile Aylık Geliri
Çok Yüksek
Yüksek
Ortalamanın Üstü
Ortalama
Düşük

Çizelge 4.13’de aile genişliği kolonunun özellikleri verilmiştir. Öğrencinin ailesi 12 kişiden fazla ise "Kalabalık", 12 kişiden az ve 6 kişiden fazla ise "Ortalama", 6 kişiden az ise “Düşük” olarak ifade edilir.

Çizelge 4.13. Veri Seti I Aile Genişliği Kolonunun Özellikleri

Aile Genişliği
Kalabalık
Ortalama
Düşük

Çizelge 4.14’de Öğrencinin Babasının Eğitim Seviyesi kolonunun özellikleri verilmiştir. Eğer öğrencinin babası hayatında hiç okula gitmemiş ise “Eğitimsiz”, 10.sınıfa kadar okumuş ise “10.Sınıf Altı”, 10.Sınıfı geçmiş ve mezun olmuş ise “10.Sınıf”, 12.Sınıfı geçmiş ve mezun olmuş ise “12.Sınıf”, lisans eğitimi almış ve mezun olmuş ise “Lisans”, yüksek lisans eğitimi almış ve mezun olmuş ise “Yüksek Lisans” olarak ifade edilir.

Çizelge 4.14. Veri Seti I Baba Eğitim Seviyesi Kolonunun Özellikleri

Baba Eğitim Seviyesi
Eğitimsiz
10.Sınıf Altı
10.Sınıf
12.Sınıf
Lisans
Yüksek Lisans

Çizelge 4.15’de Öğrencinin Annesinin Eğitim Seviyesi kolonunun özellikleri verilmiştir. Eğer öğrencinin annesi hayatında hiç okula gitmemiş ise “Eğitimsiz”, 10.sınıfa kadar okumuş ise “10.Sınıf Altı”, 10.Sınıfı geçmiş ve mezun olmuş ise “10.Sınıf”, 12.Sınıfı geçmiş ve mezun olmuş ise “12.Sınıf”, lisans eğitimi almış ve mezun olmuş ise “Lisans”, yüksek lisans eğitimi almış ve mezun olmuş ise “Yüksek Lisans” olarak ifade edilir.

Çizelge 4.15. Veri Seti I Anne Eğitim Seviyesi Kolonunun Özellikleri

Anne Eğitim Seviyesi
Eğitimsiz
10.Sınıf Altı
10.Sınıf
12.Sınıf
Lisans
Yüksek Lisans

Çizelge 4.16’da Öğrencinin Babasının mesleği kolonunun özellikleri verilmiştir. Eğer öğrencinin babası Hizmet sektöründe çalışıyor ise “Hizmet”, ticaret sektöründe çalışıyorsa “Ticaret”, çiftçilik yapıyor ise “Çiftçi”, eğer çalıştığı meslekten emekli ise “Emekli” olarak ifade edilir. Bunların dışındaki sektörler “Diğer” olarak ifade edilir.

Çizelge 4.16. Veri Seti I Baba Mesleği Kolonunun Özellikleri

Baba Mesleği
Hizmet
Ticaret
Emekli
Çiftçi
Diğer

Çizelge 4.17’de Öğrencinin Annesinin mesleği kolonunun özellikleri verilmiştir. Eğer öğrencinin annesi Hizmet sektöründe çalışıyor ise “Hizmet”, ticaret sektöründe çalışıyorsa “Ticaret”, çalışmıyor ev işi yapıyor ise “Ev Hanımı”, eğer çalıştığı meslekten emekli ise “Emekli” olarak ifade edilir. Bunların dışındaki sektörler “Diğer” olarak ifade edilir.

Çizelge 4.17. Veri Seti I Anne Mesleği Kolonunun Özellikleri

Anne Mesleği
Hizmet
Ticaret
Emekli
Ev Hanımı
Diğer

Çizelge 4.18’de Arkadaş Sayısı kolonunun özellikleri verilmiştir. Öğrencinin yakın çevresinde bulunan arkadaş sayısı fazla ise “Çok”, arkadaş sayısı çok fazla değil ise “Orta”, arkadaş sayısı birkaç kişiden oluyor ise “Az” değerlerini içermektedir.

Çizelge 4.18. Veri Seti I Arkadaş Sayısı Kolonunun Özellikleri

Arkadaş Sayısı
Çok
Orta
Az

Çizelge 4.19’da Ders Çalışma Süresi kolonunun özellikleri verilmiştir. Öğrencin ders çalışma saati fazla ise “Fazla”, ne az ne de fazla ise “Orta”, ders çalışma saati az ise “Az” olarak ifade edilir.

Çizelge 4.19. Veri Seti I Ders Çalışma Süresi Kolonunun Özellikleri

Ders Çalışma Süresi
Fazla
Orta
Az

Çizelge 4.20’de Geldiği Okul Tipi kolonunun özellikleri verilmiştir. Öğrencinin koleje başlamadan önce geldiği okul tipi özel ise “Özel”, devlet okullarından geliyor ise “Devlet” olarak ifade edilir.

Çizelge 4.20. Veri Seti I Geldiği Okul Tipi Kolonunun Özellikleri

Geldiği Okul Tipi
Devlet
Özel

Çizelge 4.21’de Sınav Dili kolonunun özellikleri verilmiştir. Öğrencinin eğitim gördüğü ve sınavda sorulan diller yer alır. Bengalce ve Hintçe modern Hint dilleridir. Assam dili ise Kuzey Hindistan’da kullanılır.

Çizelge 4.21. Veri Seti I Sınav Dili Kolonunun Özellikleri

Sınav Dili
İngilizce
Assam Dili
Hintçe
Bengalce

Çizelge 4.22’de Evden Üniversiteye Seyahat Süresi kolonunun özellikleri verilmiştir. Öğrencinin evden üniversiteye giderken harcadığı süre fazla ise “Fazla”, ne az ne de fazla ise “Orta”, harcadığı az ise “Az” olarak ifade edilir.

Çizelge 4.22. Veri Seti I Evden Üniversiteye Seyahat Süresi Kolonunun Özellikleri

Evden Üniversiteye Seyahat Süresi
Fazla
Orta
Az

Çizelge 4.23’te Sınıfa Devam Oranı kolonunun özellikleri verilmiştir. Öğrencinin dönem boyunca sınıfındaki derslere katılma oranı yüksek ise “İyi”, ortalama ise “Orta”, katılma oranı düşük ise “Az” olarak ifade edilir.

Çizelge 4.23. Veri Seti I Sınıfa Devam Yüzdesi Kolonunun Özellikleri

Sınıfa Devam Oranı
İyi
Orta
Az

Çizelge 4.24’de görüleceği gibi veri setimizde yer alan başarı tiplerinin örnek sayılarının dağılımı dengesizdir. Bu çalışmada “Best” ve “Fail” örneklerin sayısı oldukça azdır. Regresyon çalışması ile not tahmini yapıyor olunsaydı, başarının tespit edilmesinin zor olacağı görülmüştür.

Çizelge 4.24. Veri Seti I Dönem Sonu Ortalama Kolonunda Sınıf Dağılımı

Başarı	Örnek Sayısı
Best	8
Very Good	59
Good	125
Pass	76
Fail	32

Çizelge 4.25’de veri setindeki kolonların kısaltmaları yer almaktadır. Bundan sonraki bölümlerde bir kolondan bahsedilirken ilgili kolonun kısaltması verilecektir.

Çizelge 4.25. Veri Seti I Kolonların Kısaltmaları

Kolon Adları	Kısaltması
Cinsiyet	GE
Kast Sınıfı	CST
10.Sınıf Ortalama Yüzdesi	TNP
12.Sınıf Ortalama Yüzdesi	TWP
İç Değerlendirme Ortalama Yüzdesi	IAP
Dönem Sonu Ortalama Yüzdesi	ESP
Önceki Dönemden Dersi Olup Olmadığı	ARR
Medeni Hal Durumu	MS
Şehir Veya Köyde Yaşaması	LS
Öğrenci Okula Giriş Tipi	AS
Aile Aylık Geliri	FMI
Aile Genişliği	FS
Baba Eğitim Seviyesi	FQ
Anne Eğitim Seviyesi	MQ
Baba Mesleği	FO
Anne Mesleği	MO
Arkadaş Sayısı	NF
Ders Çalışma Süresi	SH
10.Sınıf'a Geldiğinde Geldiği Okul Tipi	SS
Sınav Dili	ME
Evden Üniversiteye Seyahat Süresi	TT
Sınıfa Devam Yüzdesi	ATD

4.1.2.2. Veri Seti II

Veri seti 33 özellikten oluşan 649 adet örnek içermektedir. Çizelge 4.26'da verinin özellikleri gösterilmektedir. Veri setinde her öğrenci bir sıra ile temsil edilmektedir.

Çizelge 4.26. Veri Seti II Öğrenci Veri Setinin Özellikleri

Sıra	Kolon Adları
1	Cinsiyet
2	Yaş
3	Okul
4	Adres
5	Ebeveyn
6	Baba Eğitim Seviyesi
7	Anne Eğitim Seviyesi
8	Baba Mesleği
9	Anne Mesleği
10	Veli
11	Aile Genişliği
12	Aile İlişkisi Kalitesi
13	Okulu Seçme Sebebi
14	Evden Okula Seyahat Süresi
15	Ders Çalışma Süresi
16	Önceki Dönemden Kalan Ders Sayısı
17	Okul Desteği
18	Ailenin Eğitim Desteği
19	Etkinliklere Katılma
20	Ekstra Ücretli Ders
21	Evde İnternet Erişimi
22	Anaokuluna Gidip Gitmediği
23	Yüksek Eğitim Almak İstemesi
24	Romantik İlişki
25	Boş Zaman
26	Dışarı Çıkma
27	Hafta Sonu Alkol Tüketimi
28	Hafta İçi Alkol Tüketimi
29	Sağlık Durumu
30	Devamsızlık Sayısı
31	Birinci Dönem Notu
32	İkinci Dönem Notu
33	Final Notu

Çizelge 4.27’de Cinsiyet kolonunun özellikleri verilmiştir. Cinsiyet kolonu “Erkek” ve “Kadın” değerlerini içermektedir.

Çizelge 4.27. Veri Seti II Cinsiyet Kolonunun Özellikleri

Cinsiyet
Erkek
Kadın

Çizelge 4.28’de Yaş kolonunun özellikleri verilmiştir. Yaş kolonu 15 ile 22 yaş arasındaki değerlerini içermektedir.

Çizelge 4.28. Veri Seti II Yaş Kolonunun Özellikleri

Yaş
15
16
17
18
19
20
21
22

Çizelge 4.29’da okul kolonunun özellikleri verilmiştir. Okul kolonu öğrencinin okuduğu okulu “Gabriel Pereira” veya “Mousinho da Silveira ” olarak ifade etmektedir.

Çizelge 4.29. Veri Seti II Okul Kolonunun Özellikleri

Okul
Gabriel Pereira
Mousinho da Silveira

Çizelge 4.30’da adres kolonunun özellikleri verilmiştir. Adres kolonu öğrencinin yaşadığı yerin konumunu “Kentsel” veya “Kırsal” değerlerini içermektedir.

Çizelge 4.30. Veri Seti II Adres Kolonunun Özellikleri

Adres
Kentsel
Kırsal

Çizelge 4.31’de Öğrencinin Babasının Eğitim Seviyesi kolonunun özellikleri verilmiştir. Eğer öğrencin babası hayatında hiç okula gitmemiş ise “Eğitimsiz”, 4.Sınıfı geçmiş ve mezun olmuş ise “4.Sınıf”, 5.Sınıfı geçmiş ve ise 9.Sınıfa kadar okumuş ise “5.Sınıf İle 9.Sınıf Arası”, ortaöğretim eğitimi almış ve mezun olmuş ise “Ortaöğretim”, lisans eğitimi almış ve mezun olmuş ise “Lisans” olarak ifade edilir.

Çizelge 4.31. Veri Seti II Baba Eğitim Seviyesi Kolonunun Özellikleri

Baba Eğitim Seviyesi
Eğitimsiz
4.Sınıf
5.Sınıf ile 9.Sınıf arası
Ortaöğretim
Lisans

Çizelge 4.32’de Öğrencinin Annesinin Eğitim Seviyesi kolonunun özellikleri verilmiştir. Eğer öğrencinin annesi hayatında hiç okula gitmemiş ise “Eğitimsiz”, 4.Sınıfı geçmiş ve mezun olmuş ise 4.Sınıf, 5.Sınıfı geçmiş ve ise 9.Sınıfa kadar okumuş ise “5.Sınıf İle 9.Sınıf Arası”, ortaöğretim eğitimi almış ve mezun olmuş ise “Ortaöğretim”, lisans eğitimi almış ve mezun olmuş ise “Lisans” olarak ifade edilir.

Çizelge 4.32. Veri Seti II Anne Eğitim Seviyesi Kolonunun Özellikleri

Anne Eğitim Seviyesi
Eğitimsiz
4.Sınıf
5.Sınıf ile 9.Sınıf arası
Ortaöğretim
Lisans

Çizelge 4.33’te aile genişliği kolonunun özellikleri verilmiştir. Öğrencinin ailesi 3 kişiden daha az ise "3’ten Az", 3 kişiden fazla ise “3’ten Büyük” değerlerini içermektedir.

Çizelge 4.33. Veri Seti II Aile Genişliği Kolonunun Özellikleri

Aile Genişliği
3’ten Az
3’ten Büyük

Çizelge 4.34’te Öğrencinin Babasının mesleği kolonunun özellikleri verilmiştir. Eğer öğrencinin babası Öğretmen ise “Öğretmen”, Amir, polis gibi bir konumda kamu sektöründe çalışıyorsa “Kamu Hizmetleri”, evde çalışıyor ise “Ev Erkeği”, sağlık sektöründe çalışıyor ise “Sağlık Sektörü”, olarak ifade edilir. Bunların dışındaki sektörler “Diğer” olarak ifade edilir.

Çizelge 4.34. Veri Seti II Baba Mesleği Kolonunun Özellikleri

Baba Mesleği
Öğretmen
Sağlık Sektörü
Kamu Hizmetleri
Ev Erkeği
Diğer

Çizelge 4.35’te Öğrencinin Annesinin mesleği kolonunun özellikleri verilmiştir. Eğer öğrencinin babası Öğretmen ise “Öğretmen”, Amir, polis gibi bir konumda kamu sektöründe çalışıyorsa “Kamu Hizmetleri”, evde çalışıyor ise “Ev Hanımı”, sağlık sektöründe çalışıyor ise “Sağlık Sektörü”, olarak ifade edilir. Bunların dışındaki sektörler “Diğer” olarak ifade edilir.

Çizelge 4.35. Veri Seti II Anne Mesleği Kolonunun Özellikleri

Anne Mesleği
Öğretmen
Sağlık Sektörü
Kamu Hizmetleri
Ev Kadını
Diğer

Çizelge 4.36’da Veli kolonunun özellikleri verilmiştir. Veli kolonu öğrencinin velisini “Anne”, “Baba” bunlarında dışında bir kişi ise “Diğer” olarak ifade etmektedir.

Çizelge 4.36. Veri Seti II Veli Kolonunun Özellikleri

Veli
Anne
Baba
Diğer

Çizelge 4.37’de Aile İlişkisi Kalitesi kolonunun özellikleri verilmiştir. Aile ilişkisi kalitesi kolonu öğrencinin ailesindeki ilişkiyi 1 ile 5 arası puanlar ile ifade edilmektedir.

Çizelge 4.37. Veri Seti II Aile İlişkisi Kalitesi Kolonunun Özellikleri

Aile İlişkisi Kalitesi
1
2
3
4
5

Çizelge 4.38’de Okulu seçme sebebi kolonunun özellikleri verilmiştir. Okulu seçme sebebi kolonu öğrencinin okulu seçme nedenini “Eve Yakın”, ”Okul Şöhreti”, “Ders Seçimi” ve “Diğer” değerlerini içermektedir.

Çizelge 4.38. Veri Seti II Okulu Seçme Sebebi Kolonunun Özellikleri

Okulu Seçme Sebebi
Eve Yakın
Okul Şöhreti
Ders Seçimi
Diğer

Çizelge 4.39’da Evden Üniversiteye Seyahat Süresi kolonunun özellikleri verilmiştir. Öğrencinin evden üniversiteye giderken harcadığı süre 1 saatten fazla ise “4”, 30 dakika ile 1 saat arasında ise “3”, 15 dakika ile 30 dakika arasında ise “2”, 15 dakikadan az ise “1” olarak ifade edilmektedir.

Çizelge 4.39. Veri Seti II Evden Üniversiteye Seyahat Süresi Kolonunun Özellikleri

Evden Üniversiteye Seyahat Süresi
4
3
2
1

Çizelge 4.40'ta Ders Çalışma Süresi kolonunun özellikleri verilmiştir. Öğrencin haftalık ders çalışma saati 10 saatten fazla ise “4”, 5 saat ile 10 saat arasında ise “3”, 2 saat ile 5 saat arasında ise “2”, 2 saatten az ise “1” olarak ifade edilmektedir.

Çizelge 4.40. Veri Seti II Ders Çalışma Süresi Kolonunun Özellikleri

Ders Çalışma Süresi
4
3
2
1

Çizelge 4.41'de Önceki dönemlerden kalan ders sayısı özellikleri verilmiştir. Önceki dönemlerden kalan ders sayısı kolonu öğrencinin öndeki dönemden kaldığı derslerin sayısını azdan çoğa doğru 0 ile 5 arası puanlar ile ifade edilmektedir.

Çizelge 4.41. Veri Seti II Önceki Dönemden Kalan Ders Sayısı Kolonunun Özellikleri

Önceki Dönemden Kalan Ders Sayısı
1
2
3
4
5

Çizelge 4.42'de Okul desteği kolonunun özellikleri verilmiştir. Okul desteği kolonu öğrencinin ekstra eğitim için okuldan aldığı desteği “Evet” veya “Hayır” olarak ifade etmektedir.

Çizelge 4.42. Veri Seti II Okul Desteği Kolonunun Özellikleri

Okul Desteği
Evet
Hayır

Çizelge 4.43'de Ailenin eğitim desteği kolonunun özellikleri verilmiştir. Ailenin eğitim desteği verme durumu “Evet” veya “Hayır” olarak ifade etmektedir.

Çizelge 4.43. Veri Seti II Ailenin Eğitim Desteği Kolonunun Özellikleri

Ailenin Eğitim Desteği
Evet
Hayır

Çizelge 4.44'te Etkinliklere katılma kolonunun özellikleri verilmiştir. Etkinliklere katılma kolonu öğrencinin okuldan sonraki yapılan etkinliklere katılma durumunu “Evet” veya “Hayır” olarak ifade etmektedir.

Çizelge 4.44. Veri Seti II Etkinliklere Katılma Kolonunun Özellikleri

Etkinliklere Katılma
Evet
Hayır

Çizelge 4.45'te Ekstra ücretli ders kolonunun özellikleri verilmiştir. Ekstra ücretli ders kolonu öğrencinin okul dışında ekstra ücretli ders alıp almadığını “Evet” veya “Hayır” olarak ifade etmektedir.

Çizelge 4.45. Veri Seti II Ekstra Ücretli Ders Kolonunun Özellikleri

Ekstra Ücretli Ders
Evet
Hayır

Çizelge 4.46'da Evde internet erişimi kolonunun özellikleri verilmiştir. Evde internet erişimi kolonu öğrencinin evindeki internet erişimini “Evet” veya “Hayır” olarak ifade etmektedir.

Çizelge 4.46. Veri Seti II Evde İnternet Erişimi Kolonunun Özellikleri

Evde İnternet Erişimi
Evet
Hayır

Çizelge 4.47'de Anaokuluna gidip gitmediği kolonunun özellikleri verilmiştir. Anaokuluna gidip gitmediği kolonu öğrencinin anaokuluna gidip gitmediğini “Evet” veya “Hayır” olarak ifade etmektedir.

Çizelge 4.47. Veri Seti II Anaokuluna Gidip Gitmediği Kolonunun Özellikleri

Anaokuluna Gidip Gitmediği
Evet
Hayır

Çizelge 4.48'de Yüksek eğitim almak istemesi kolonunun özellikleri verilmiştir. Yüksek eğitim almak istemesi kolonu öğrencinin ekstra eğitim için okuldan aldığı desteği “Evet” veya “Hayır” olarak ifade etmektedir.

Çizelge 4.48. Veri Seti II Yüksek Eğitim Almak İstemesi Kolonunun Özellikleri

Yüksek Eğitim Almak İstemesi
Evet
Hayır

Çizelge 4.49’da Romantik ilişki kolonunun özellikleri verilmiştir. Romantik ilişki kolonu öğrencinin romantik ilişkisinin olup olmadığını “Evet” veya “Hayır” olarak ifade etmektedir.

Çizelge 4.49. Veri Seti II Romantik İlişki Kolonunun Özellikleri

Romantik İlişki
Evet
Hayır

Çizelge 4.50’de Boş zaman kolonunun özellikleri verilmiştir. Öğrencinin okuldaki sonraki boş zamanını azdan çoğa doğru 0 ile 5 arası puanlar ile ifade edilmektedir.

Çizelge 4.50. Veri Seti II Boş Zaman Kolonunun Özellikleri

Boş Zaman
1
2
3
4
5

Çizelge 4.51’de Dışarı çıkma kolonunun özellikleri verilmiştir. Öğrencinin okuldaki sonraki arkadaşları ile dışarı çıkma zamanını azdan çoğa doğru 0 ile 5 arası puanlar ile ifade edilmektedir.

Çizelge 4.51. Veri Seti II Dışarı Çıkma Kolonunun Özellikleri

Dışarı Çıkma
1
2
3
4
5

Çizelge 4.52’de Hafta sonu alkol tüketimi kolonunun özellikleri verilmiştir. Öğrencinin hafta sonu alkol tüketimini azdan çoğa doğru 0 ile 5 arası puanlar ile ifade edilmektedir.

Çizelge 4.52. Veri Seti II Hafta Sonu Alkol Tüketimi Kolonunun Özellikleri

Hafta Sonu Alkol Tüketimi
1
2
3
4
5

Çizelge 4.53'te Hafta içi alkol tüketimi kolonunun özellikleri verilmiştir. Öğrencinin hafta sonu alkol tüketimini azdan çoğa doğru 0 ile 5 arası puanlar ile ifade edilmektedir.

Çizelge 4.53. Veri Seti II Hafta İçi Alkol Tüketimi Kolonunun Özellikleri

Hafta İçi Alkol Tüketimi
1
2
3
4
5

Çizelge 4.54'te Sağlık durumu kolonunun özellikleri verilmiştir. Öğrencinin sağlık durumunu azdan çoğa doğru 0 ile 5 arası puanlar ile ifade edilmektedir.

Çizelge 4.54. Veri Seti II Sağlık Durumu Kolonunun Özellikleri

Sağlık Durumu
1
2
3
4
5

Çizelge 4.55'te Ebeveyn durumu kolonunun özellikleri verilmiştir. Ebeveyn durumu kolonu öğrencinin ebeveynlerinin durumunu “Birlikte” veya “Ayrı” olarak ifade etmektedir.

Çizelge 4.55. Veri Seti II Ebeveyn Durumu Kolonunun Özellikleri

Ebeveyn Durumu
Birlikte
Ayrı

Çizelge 4.56'da Devamsızlık sayısı kolonunun özellikleri verilmiştir. Devamsızlık sayısı azdan çoğa doğru 0 ile 93 arası puanlar ile ifade edilmektedir.

Çizelge 4.56. Veri Seti II Devamsızlık Sayısı Kolonunun Özellikleri

Devamsızlık Sayısı
0
1
...
...
92
93

Çizelge 4.57’de Birinci dönem kolonunun özellikleri verilmiştir. Birinci dönem notunu 0 ile 20 arası puanlar ile ifade edilmektedir.

Çizelge 4.57. Veri Seti II Birinci Dönem Notu Kolonunun Özellikleri

Birinci Dönem Notu
0
1
...
...
19
20

Çizelge 4.58’de İkinci dönem kolonunun özellikleri verilmiştir. İkinci dönem notunu 0 ile 20 arası puanlar ile ifade edilmektedir.

Çizelge 4.58. Veri Seti II İkinci Dönem Notu Kolonunun Özellikleri

İkinci Dönem Notu
0
1
...
...
19
20

Çizelge 4.59’da Final notu kolonunun özellikleri verilmiştir. Final notunu 0 ile 20 arası puanlar ile ifade edilmektedir.

Çizelge 4.59. Veri Seti II Final Notu Kolonunun Özellikleri

Final Notu
0
1
...
...
19
20

Çizelge 4.60’de veri setindeki kolonların kısaltmaları yer almaktadır. Bundan sonraki bölümlerde bir kolondan bahsedilirken ilgili kolonun kısaltması ile bahsedilecektir.



Çizelge 4.60. Veri Seti II Kolonların Kısaltmaları

Kolon Adları	Kısaltması
Cinsiyet	Sex
Yaş	Age
Okul	School
Adres	Address
Ebeveyn	Pstatus
Baba Eğitim Seviyesi	Medu
Anne Eğitim Seviyesi	Mjob
Baba Mesleği	Fedu
Anne Mesleği	Fjob
Veli	Guardian
Aile Genişliği	Famsize
Aile İlişkisi Kalitesi	Famrel
Okulu Seçme Sebebi	Reason
Evden Okula Seyahat Süresi	Traveltime
Ders Çalışma Süresi	Studytime
Önceki Dönemlerden Kalan Ders Sayısı	Failures
Okul Desteği	Schoolsup
Ailenin Eğitim Desteği	Famsup
Etkinliklere Katılma	Activities
Ekstra Ücretli Ders	Paidclass
Evde İnternet Erişimi	İnternet
Anaokuluna Gidip Gitmediği	Nursery
Yüksek Eğitim Almak İstemesi	Higher
Romantik İlişki	Romantic
Boş Zaman	Freetime
Dışarı Çıkma	Goout
Hafta Sonu Alkol Tüketimi	Walc
Hafta İçi Alkol Tüketimi	Dalc
Sağlık Durumu	Health
Devamsızlık Sayısı	Absences
Birinci Dönem Notu	G1
İkinci Dönem Notu	G2
Final Notu	G3

4.1.3. Verinin Hazırlanması

4.1.3.1 Veri Seti I Hazırlanması

Veri kümesi Pandas ve Scipy.io kütüphaneleri yardımıyla çalışma ortamına aktarılmıştır. Araştırma için kullanılan veri seti “ARFF” formatındadır. WEKA üzerinde yapılan çalışmalar için Özellik İlişki Dosya Formatı, ARFF kullanılmaktadır.

```
@RELATION Sapfile1

@ATTRIBUTE ge {M,F}
@ATTRIBUTE cst {G,ST,SC,OBC,MOBC}
@ATTRIBUTE tnp {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE twp {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE iap {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE esp {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE arr {Y,N}
@ATTRIBUTE ms {Married,Unmarried}
@ATTRIBUTE ls {T,V}
@ATTRIBUTE as {Free,Paid}
@ATTRIBUTE fmi {Vh,High,Am,Medium,Low}
@ATTRIBUTE fs {Large,Average,Small}
@ATTRIBUTE fq {1l,Um,10,12,Degree,Pg}
@ATTRIBUTE mq {1l,Um,10,12,Degree,Pg}
@ATTRIBUTE fo {Service,Business,Retired,Farmer,Others}
@ATTRIBUTE mo {Service,Business,Retired,Housewife,Others}
@ATTRIBUTE nf {Large,Average,Small}
@ATTRIBUTE sh {Good,Average,Poor}
@ATTRIBUTE ss {Govt,Private}
@ATTRIBUTE me {Eng,Asm,Hin,Ben}
@ATTRIBUTE tt {Large,Average,Small}
@ATTRIBUTE atd {Good,Average,Poor}

@DATA
F,G,Good,Good,Vg,Good,Y,Unmarried,V,Paid,Medium,Average,Um,10,Farmer,Housewife,Large,Poor,Govt,Asm,Small,Good
M,OBC,Vg,Vg,Vg,Vg,N,Unmarried,V,Paid,Low,Average,Um,1l,Service,Service,Small,Poor,Govt,Asm,Average,Average
F,OBC,Good,Good,Vg,Good,N,Unmarried,V,Paid,Am,Average,12,10,Service,Housewife,Average,Average,Govt,Asm,Large,Good
M,MOBC,Pass,Good,Vg,Good,N,Unmarried,V,Paid,Medium,Small,12,Um,Business,Business,Large,Poor,Govt,Asm,Average,Average
M,G,Good,Good,Vg,Vg,N,Unmarried,V,Paid,Am,Average,10,12,Service,Housewife,Large,Poor,Private,Asm,Small,Good
M,OBC,Vg,Good,Vg,Vg,N,Unmarried,V,Paid,Medium,Average,10,Um,Farmer,Housewife,Small,Poor,Private,Eng,Average,Average
F,OBC,Good,Vg,Good,Good,N,Unmarried,V,Paid,High,Average,12,Degree,Service,Service,Average,Poor,Govt,Asm,Small,Good
F,OBC,Good,Good,Good,Good,N,Unmarried,V,Paid,Medium,Small,1l,Um,Farmer,Housewife,Small,Poor,Private,Asm,Average,Good
F,MOBC,Vg,Good,Good,Vg,Y,Unmarried,V,Paid,Low,Small,12,10,Others,Housewife,Small,Good,Govt,Asm,Average,Good
```

Şekil 4.1. ARFF Formatında Veri Setinin Görünümü

Şekil 4.1.’de çalışmada kullanılan ARFF formatındaki veri setinin içeriği gösterilmiştir. ARFF veri formatında, veri noktaları sayısal ve kategorik olarak yer alır. ARFF formatını python dilinde okuyabilmek için “scipy.io” kütüphanesi kullanılmıştır. Diziler üzerinde hesaplama ve dönüşüm işlemleri için “NumPy”, veri seti analizi için “Pandas” kullanılmıştır. “ARFF” formatlı dosya okunduktan sonra ekranda gösterilmiştir. Çizelge 4.61’de dosya okunduktan sonra veri setinin ilk 3 transpoze hali gösterilmiştir.

Çizelge 4.61. Veri Setinin İlk 3 Satırı

	1.Satır	2.Satır	3.Satır
GE	b'F'	b'M'	b'F'
CST	b'G'	b'OBC'	b'OBC'
TNP	b'Good'	b'Vg'	b'Good'
TWP	b'Good'	b'Vg'	b'Good'
IAP	b'Vg'	b'Vg'	b'Vg'
ESP	b'Good'	b'Vg'	b'Good'
ARR	b'Y'	b'N'	b'N'
MS	b'Unmarried'	b'Unmarried'	b'Unmarried'
LS	b'V'	b'V'	b'V'
AS	b'Paid'	b'Paid'	b'Paid'
FMI	b'Medium'	b'Low'	b'Am'
FS	b'Average'	b'Average'	b'Average'
FQ	b'Um'	b'Um'	b'12'
MQ	b'10'	b'11'	b'10'
FO	b'Farmer'	b'Service'	b'Service'
MO	b'Housewife'	b'Service'	b'Housewife'
NF	b'Large'	b'Small'	b'Average'
SH	b'Poor'	b'Poor'	b'Average'
SS	b'Govt'	b'Govt'	b'Govt'
ME	b'Asm'	b'Asm'	b'Asm'
TT	b'Small'	b'Average'	b'Large'
ATD	b'Good'	b'Average'	b'Good'

Görüldüğü üzere veri seti satırlarında fazla karakterler ortaya çıkmıştır. Bunun sebebi ARFF formatının karakter standardından dolayıdır. Sınıflandırma modeli oluştururken hesaplamalarda yaratılacak muhtemel sorunlardan dolayı veri setinin karakter standardı “UTF-8” olarak düzeltilmiştir. Çizelge 4.62’de veri setinin karakter kodlaması değiştirildikten sonra veri setinin ilk 3 transpoze hali gösterilmiştir.

Çizelge 4.62. Karakter Seti Değiştirilmiş Veri Setinin İlk 3 Satırı

	1.Satır	2.Satır	3.Satır
GE	F	M	F
CST	G	OBC	OBC
TNP	Good	Vg	Good
TWP	Good	Vg	Good
IAP	Vg	Vg	Vg
ESP	Good	Vg	Good
ARR	Y	N	N
MS	Unmarried	Unmarried	Unmarried
LS	V	V	V
AS	Paid	Paid	Paid
FMI	Medium	Low	Am
FS	Average	Average	Average
FQ	Um	Um	12
MQ	10	Il	10
FO	Farmer	Service	Service
MO	Housewife	Service	Housewife
NF	Large	Small	Average
SH	Poor	Poor	Average
SS	Govt	Govt	Govt
ME	Asm	Asm	Asm
TT	Small	Average	Large
ATD	Good	Average	Good

Veri setinin istatistiksel olarak analizi Pandas kütüphanesinin “.describe()” metodu ile yapılabilir. Çizelge 4.63’te veri setinin temel analizi gösterilmiştir. Count, eksik olmayan verilerin toplam sayısını ifade eder. Veri setinde eksik veriler içermediği bu veri setinden anlaşılmaktadır. Unique, o kolon içerisinde birbirinden farklı kaç adet özellik olduğunu gösterir. Veri setinde kirli verinin olmaması gerekir. Gerçek hayat verisinde kirli verinin karşımıza çıkma ihtimali yüksektir. Bir öğrencinin medeni hali incelediğinde bu durum Çizelge 4.9’da gösterildiği gibi “Evli” ve “Bekâr” olarak ifade edilebilir. Veri toplanırken bu alana “Dul” veya “Boşanmış” gibi ifadeler gelebilir. Böyle bir durum yaşanınca kirli veri temizlenmelidir. Veri temizlenirken konunun uzmanı kişilere danışarak nasıl temizlenmesi gerektiği konusunda karar verilmelidir. Unique kolonu incelendiğinde herhangi bir veri kirliliği durumu yaşanmamıştır. Top, o

kolon içerisinde satırlarda en çok yer alan özelliği ifade eder. Freq ise o kolon içerisinde satırlarda en çok yer alan özelliğin bulunma sıklığını ifade eder.

Çizelge 4.63. Veri Seti I Temel Analizi

	Count	Unique	Top	Freq
ge	300	2	M	150
cst	300	5	OBC	118
tnp	300	4	Good	154
twp	300	4	Good	162
iap	300	4	Good	165
esp	300	5	Good	125
arr	300	2	Y	152
ms	300	2	Unmarried	299
ls	300	2	V	240
as	300	2	Free	191
fmi	300	5	Medium	124
fs	300	3	Small	202
fq	300	6	Um	135
mq	300	6	Um	138
fo	300	5	Farmer	117
mo	300	5	Housewife	269
nf	300	3	Average	108
sh	300	3	Average	168
ss	300	2	Govt	221
me	300	4	Asm	193
tt	300	3	Small	141
atd	300	3	Average	132

Sınıflandırma yapabilmek için sayısal veya kategorik verilere ihtiyaç bulunmaktadır. Kategorik veriler sınırlı sayıda belli bir alanda bulunan veri tipidir. Kategorik veriyi örnekler ile açıklarsak;

- Bir arabanın rengi Mavi, Kırmızı, Beyaz, Siyah olabilir,
- Bir kitabın içindeki kağıt türü Ivory, Şamua, Bristol, Kuşe olabilir,
- Bir kişinin medeni durumu Evli veya Bekar olabilir.

Sıcaklık, fiyat, maliyet, miktar, kalem gibi ifadeler tamsayı veya sürekli sayı olarak ifade edilebilir. Genel olarak bakıldığında bir arabanın rengi Yeşil veya Gri de olabilir. Ama bir araba üreten bir fabrika açısından baktığımızda yukarıda verildiği gibi Mavi, Kırmızı, Beyaz, Siyah renklerinde arabalar ürettiği durumda, sadece 4 adet renk olduğu için renk kategorik bir değişkendir. Başka bir durumda renk yüzlerce değeri temsil edebilir. Ancak bu durumda rengin kategorik veri olarak değerlendirilmesi iyi bir seçim

olmaz. Kategorik verilerin kaç adet değeri olabileceğine dair bir sınıflandırma yoktur. Bunun veri setinden edilen analize göre değerlendirilmesi gerekir. Çizelge 4.64’de veri setinde bulunan kolonların veri tipi verilmiştir. Tüm kolonların veri tipi object tipindedir.

Çizelge 4.64 Veri Seti I Tüm Kolonların Veri Tipi

	Kolon Veri Tipi
ge	object
cst	object
tnp	object
twp	object
iap	object
esp	object
arr	object
ms	object
ls	object
as	object
fmi	object
fs	object
fq	object
mq	object
fo	object
mo	object
nf	object
sh	object
ss	object
me	object
tt	object
atd	object

Çizelge 4.63 ve Çizelge 4.64’de verilen durumlar incelendiğinde bu verilerin kategorik verilere dönüştürülmesi gerekmektedir. Veri seti üzerindeki tüm kolonlar tip dönüşümü kullanılarak kategori tipine dönüştürülmüştür. Çıktı kolonu olan ESP kolonu tekrar analiz edildiğinde kolon hiyerarşisi Şekil 4.2’de gösterilmiştir.



Şekil 4.2. ESP Kolonunun Hiyerarşisi

Tüm kolonlar kategorik verilere dönüşmüştür. Makinenin anlayacağı hale gelmiştir. Fakat Şekil 4.2’deki hiyerarşi incelendiği zaman yanlış olduğu ortaya çıkmıştır. Artık yeni bir veri sorununuz ortaya çıkmıştır. Bir kıyafetin beden boyutlarını, “X-

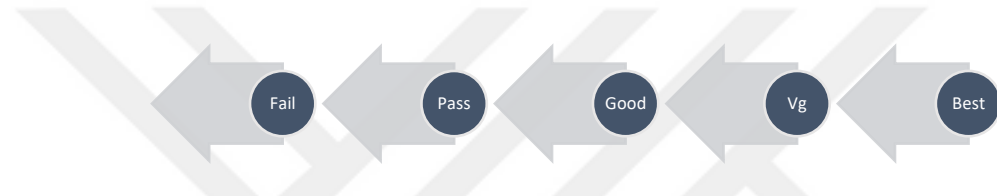
Small”, “Small”, “Medium”, “Large” ve “X-Large” olarak ifade edebiliriz. Bir programlama dilinde bu bedenleri sıralamaya sokarsak;

- Large < Medium < Small < X-Large < X-Small

olarak sonuç elde edilir. Ama asıl beden sıralaması;

- X-Small < Small < Medium < Large < X-Large

olmalıdır. Şekil 4.2’deki görüldüğü üzere benzer bir durumla karşılaşmıştır. Verinin anlaşılması durumunda ESP kolonundaki verilerin bir hiyerarşisi olduğu bilinmektedir. Bu bilgiden yola çıkarak kolon hiyerarşisini “set_categories” metodu ile tekrar tekrar düzenlenmiştir. ESP, TNP, TWP, IAP, FMI, FS, FQ, MQ, FO, MO, NF, SH, TT ve ATD kolonlarında hiyerarşi düzenlemesi yapılmıştır. Çıktı kolonu olan ESP kolonu tekrar analiz edildiğinde düzenlenmiş kolon hiyerarşisi Şekil 4.3’te gösterilmiştir.



Şekil 4.3. ESP Kolonunun Düzenlenmiş Hiyerarşisi

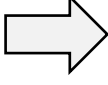
Yapay Sinir Ağları gibi algoritmalar veri girdisi olarak sayısal veriler kabul etmektedir. Günümüzde makine öğrenmesi yöntemleri veri girdisi olarak kategori tipinde de verileri kabul etmektedir. WEKA gibi bazı programlar kendi içerisinde otomatik olarak, kategorik verileri sayısal veri haline dönüştürür. Bu çalışmada tüm makine öğrenmesi sınıflandırma modellerinde aynı veriyi kullanabilmek için tüm veriler sayısal olarak dönüştürülmektedir. Sayısal değerlere çevirme işlemine encoding denir. Kategorik veriler ölçülemeyen, üzerinden sayısal işlem yapılamayan nesnel veriler içerir.

Python üzerinde kategorik veriyi sayısal veriye dönüştürmek için sık kullanılan üç yöntem vardır;

1. Label Encoder
2. Ordinal Encoder
3. One Hot Encoder

Label Encoder, okuma sırasına değer ataması yapmaktadır. Ordinal Encoder, verilen sayısal eşlenik ile veri dönüşümü gerçekleştirmektedir. One Hot Encoder ise kategorik veriyi yeni kolonlar oluşturarak binary olarak göstermektedir. Çizelge 4.65’te One Hot Encoder yöntemi gösterilmiştir. Bu yöntem Naive Bayes sınıflandırması işlemlerinde oldukça sık kullanılmaktadır.

Çizelge 4.65 One Hot Encoder Örneği

İçecek Tipi	Fayda	Kalsiyum		İçecek Tipi	Fayda Zararlı	Fayda Zararsız	Kalsiyum
Soda	Zararsız	191		Soda	0	1	191
Kola	Zararlı	150		Kola	1	0	150
Gazoz	Zararlı	175		Gazoz	1	0	175
Süt	Zararsız	180		Süt	0	0	180

Kolon hiyerarşi düzenlemesi kategorik veriyi sayısal veriye dönüştürme adımında kullanılmıştır. Hiyerarşiyi göz önünde bulundurarak her bir farklı veri için, sayısal değer belirlenmiştir. Label Encoder yöntemi ile kategorik veri sayısal verilere dönüştürülmüştür. Çizelge 4.66’da, kategorik veriler sayısallaştırılmış hale getirilmiştir.

Çizelge 4.66 Kategorik Verilerin Dönüşümü

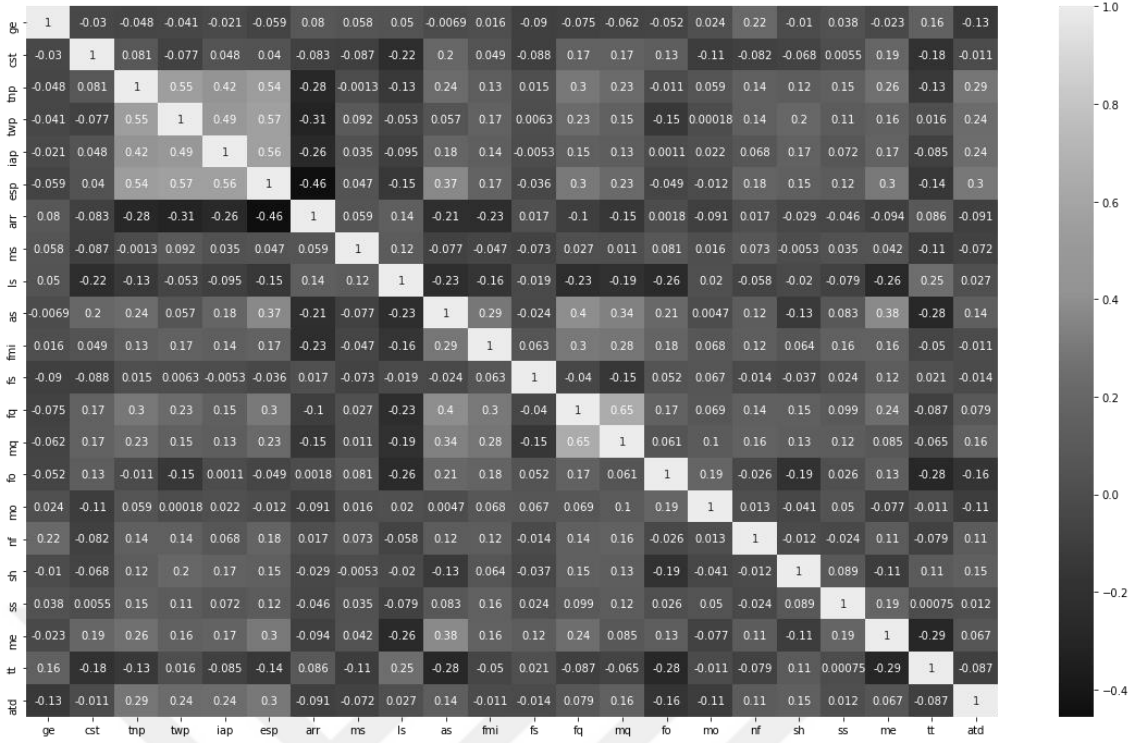
GE		FQ		ARR		SH		MO	
F	1	Um	2	Y	2	Average	2	Housewif	2
M	2	Il	1	N	1	Poor	1	Service	5
CST	10	3	MS	Good	3	Others	1		
OBC	2	12	4	Unmarried	2	ss	Business	3	
G	5	Degree	5	Married	1	Govt	1	Retired	4
MOBC	1	Pg	6	LS	Private	2	NF		
ST	3	MQ	V	2	ME	Average	2		
SC	4	Um	2	T	1	Asm	1	Large	3
TNP	Il	1	AS	Eng	3	Small	1		
Good	3	10	3	Free	1	Hin	4	TWP	
Pass	2	12	4	Paid	2	Ben	2	Good	3
Vg	4	Degree	5	FMI	TT	Vg	4		
Best	5	Pg	6	Medium	3	Small	1	Pass	2
IAP	FO	Low	2	Average	2	Best	5		
Good	3	Farmer	2	Am	4	Large	3	ESP	
Vg	2	Service	5	High	5	ATD	Good	3	
Pass	4	Business	4	Vh	6	Average	2	Pass	2
Best	5	Others	1	FS	Good	3	Vg	4	
	Retired	3	Small	1	Poor	1	Fail	1	
			Average	2			Best	5	
			Large	3					

Çizelge 4.67’da kolonlar arasındaki standart sapma değerleri verilmiştir. Veri setinde standart sapması 0 olan özellikler kullanılmayacaktır. Standart sapmalar göz önünde bulunduğunda 0 olan özellik tespit edilememiştir.

Çizelge 4.67. Kolonlar Arasındaki Standart Sapma Değerleri

Kolonlar	Standart Sapma
GE	0.500835
CST	1.461429
TNP	0.760471
TWP	0.726094
IAP	0.630018
ARR	0.500791
MS	0.057735
LS	0.400668
AS	0.481763
FMI	1.026402
FS	0.517088
FQ	1.293406
MQ	1.082464
FO	1.470160
MO	0.837393
NF	0.801309
SH	0.661710
SS	0.441178
ME	0.996868
TT	0.709067
ATD	0.745974

Standart sapma analizinden sonra korelasyonu yüksek olan özneliliklerin tespiti için korelasyon analizi yapılır. Modeli eğitmek için kullanılan kolonlar arasındaki korelasyon, makine öğrenmesi tekniklerinde önemlidir. Korelasyon analizi sırasında yüksek ve düşük korelasyonlara göre özellik seçimi yapılır. Modeli hedef sınıf kolonu ile ilgili olmayan alanlarla eğitmek, eğitim ve tahmin maliyetlerini gereksiz yere artıracak ve modelin kalitesini düşürecektir. Korelasyon matrisinde korelasyon değeri arttıkça renk git gide açılmaktadır. Veri setinde yüksek ilişkili özellikler var ise veri kümesinde yalnızca bir özelliği tutmamız önerilir (Srinivasan, 2019; Aytekin, 2021). Korelasyon matrisinde +0.9'dan yüksek, -0.9'dan düşük değerler yüksek düzeyde ilişkili olduğundan çıkarılmalıdır. Matris incelendiğinde yüksek korelasyona sahip değer tespit edilmemiştir. Şekil 4.4'de korelasyon matrisi gösterilmiştir.



Şekil 4.4. Veri Seti I Korelasyon Matrisi

TNP, TWP, IAP özellikleri orta düzeyde ilişkili değişkenlerdir. ARR, AS, FQ, ME, özellikleri düşük düzeyde ilişkili değişkenlerdir. GE, FMI, FS, MS, LS, MQ, FO, MO, TT, ATD, NF, SH, SS, CST özellikleri orta düzeyde ilişkili değişkenlerdir.

Son olarak veri seti “Feather” formatına dönüştürülerek local bilgisayara kaydedilmiştir. CSV veya ARFF formatı ile de kaydedip kullanılabilirdi. Fakat Feather dosyalarını okumak ve yazmak bellek için daha hızlıdır. Feather formatında veri binary şeklinde tutulur. Genellikle Python veya R projelerinde tercih edilir. Diğer programlarda da kullanılabilir.

4.1.3.2 Veri Seti II Hazırlanması

Veri kümesi Pandas ve Scipy.io kütüphaneler yardımıyla çalışma ortamı aktarılmıştır. Araştırma için kullanılan veri seti “CSV” formatıdır. Şekil 4.5.’de çalışmada kullanılan CSV formatındaki veri setinin içeriği gösterilmiştir. CSV veri formatında, veri virgülle ayrılmış olarak veriler yer alır.

```

school;sex;age;address;famsize;Pstatus;Medu;Fedu;Mjob;Fjob;reason;guardian;traveltime;studytime;
failures;schoolsup;famsup;paid;activities;nursery;higher;internet;romantic;famrel;freetime;goout;Dalc;
Walc;health;absences;G1;G2;G3
"GP";"F";18;"U";"GT3";"A";4;4;"at_home";"teacher";"course";"mother";2;2;0;"yes";"no";"no";"no";"yes";
"yes";"no";"no";4;3;4;1;1;3;4;"0";"11";11
"GP";"F";17;"U";"GT3";"T";1;1;"at_home";"other";"course";"father";1;2;0;"no";"yes";"no";"no";"no";"yes";
"yes";"no";5;3;3;1;1;3;2;"9";"11";11
"GP";"F";15;"U";"LE3";"T";1;1;"at_home";"other";"other";"mother";1;2;0;"yes";"no";"no";"no";"yes";"yes";
"yes";"no";4;3;2;2;3;3;6;"12";"13";12
"GP";"F";15;"U";"GT3";"T";4;2;"health";"services";"home";"mother";1;3;0;"no";"yes";"no";"yes";"yes";
"yes";"yes";"yes";3;2;2;1;1;5;0;"14";"14";14
"GP";"F";16;"U";"GT3";"T";3;3;"other";"other";"home";"father";1;2;0;"no";"yes";"no";"no";"yes";"yes";
"no";"no";4;3;2;1;2;5;0;"11";"13";13
"GP";"M";16;"U";"LE3";"T";4;3;"services";"other";"reputation";"mother";1;2;0;"no";"yes";"no";"yes";
"yes";"yes";"yes";"no";5;4;2;1;2;5;6;"12";"12";13
"GP";"M";16;"U";"LE3";"T";2;2;"other";"other";"home";"mother";1;2;0;"no";"no";"no";"no";"yes";"yes";
"yes";"no";4;4;4;1;1;3;0;"13";"12";13
"GP";"F";17;"U";"GT3";"A";4;4;"other";"teacher";"home";"mother";2;2;0;"yes";"yes";"no";"no";"yes";"yes";
"no";"no";4;1;4;1;1;1;2;"10";"13";13
"GP";"M";15;"U";"LE3";"A";3;2;"services";"other";"home";"mother";1;2;0;"no";"yes";"no";"no";"yes";"yes";
"yes";"no";4;2;2;1;1;1;0;"15";"16";17
"GP";"M";15;"U";"GT3";"T";3;4;"other";"other";"home";"mother";1;2;0;"no";"yes";"no";"yes";"yes";"yes";
"yes";"no";5;5;1;1;1;5;0;"12";"12";13
"GP";"F";15;"U";"GT3";"T";4;4;"teacher";"health";"reputation";"mother";1;2;0;"no";"yes";"no";"no";"yes";
"yes";"yes";"no";3;3;3;1;2;2;2;"14";"14";14
"GP";"F";15;"U";"GT3";"T";2;1;"services";"other";"reputation";"father";3;3;0;"no";"yes";"no";"yes";
"yes";"yes";"yes";"no";5;2;2;1;1;4;0;"10";"12";13
"GP";"M";15;"U";"LE3";"T";4;4;"health";"services";"course";"father";1;1;0;"no";"yes";"no";"yes";"yes";
"yes";"yes";"no";4;3;3;1;3;5;0;"12";"13";12
"GP";"M";15;"U";"GT3";"T";4;3;"teacher";"other";"course";"mother";2;2;0;"no";"yes";"no";"no";"yes";
"yes";"yes";"no";5;4;3;1;2;3;0;"12";"12";13
"GP";"M";15;"U";"GT3";"A";2;2;"other";"other";"home";"other";1;3;0;"no";"yes";"no";"no";"yes";"yes";

```

Şekil 4.5. CSV Formatında Veri Setinin Görünümü

CSV formatını python dilinde okuyabilmek için “Pandas” kütüphanesi kullanılmıştır. Diziler üzerinde hesaplama ve dönüşüm işlemleri için “NumPy”, veri seti analizi için “Pandas” kullanılmıştır. “CSV” formatlı dosya okunduktan sonra ekranda gösterilmiştir. Veri Seti I de olduğu gibi herhangi bir karakter sorunu ortaya çıkmamıştır. Çizelge 4.68’de dosya okunduktan sonra veri setinin ilk 3 satırın transpoze hali gösterilmiştir.

Çizelge 4.68. Veri Seti II'nin İlk 3 Satırı

	1.Satır	2.Satır	3.Satır
SCHOOL	GP	GP	GP
SEX	F	F	F
AGE	18	17	15
ADDRESS	U	U	U
FAMSIZE	GT3	GT3	LE3
PSTATUS	A	T	T
MEDU	4	1	1
FEDU	4	1	1
MJOB	at_home	at_home	at_home
FJOB	teacher	other	other
REASON	course	course	other
GUARDIAN	mother	father	mother
TRAVELTIME	2	1	1
STUDYTIME	2	2	2
FAILURES	0	0	0
SCHOOLSUP	yes	no	yes
FAMSUP	no	yes	no
PAID	no	no	no
ACTIVITIES	no	no	no
NURSERY	yes	no	yes
HIGHER	yes	yes	yes
INTERNET	no	yes	yes
ROMANTIC	no	no	no
FAMREL	4	5	4
FREETIME	3	3	3
GOOUT	4	3	2
DALC	1	1	2
WALC	1	1	3
HEALTH	3	3	3
ABSENCES	4	2	6
G1	0	9	12
G2	11	11	13
G3	11	11	12

Çizelge 4.69’da veri setinin temel analizi gösterilmiştir. Count, eksik olmayan verilerin toplam sayısını ifade eder. Unique kolonu metin kategorik değişkenlerin olduğu kolonlar içerisinde birbirinden farklı adet özellik olduğunu gösterir. Nümerik değişken yer alan kolonlarda ise sayısal istatistikler yer almaktadır. Veri setinin eksik veriler içermediği bu veri setinden anlaşılmaktadır. Çıktı kolonu olan nümerik değerler içeren G3 kolonunun aldığı değerlerin ortalaması 11 olarak hesaplanmıştır.

Çizelge 4.69. Veri Seti II Temel Analizi

	Count	Unique	Top	Freq	Mean	Std	Min	25%	50%	75%	Max
SCHOOL	649	2	GP	423							
SEX	649	2	F	383							
AGE	649				16,74422	1,21813	15.0	16.0	17.0	18.0	22.0
ADDRESS	649	2	U	452							
FAMSIZE	649	2	GT3	457							
PSTATUS	649	2	T	569							
MEDU	649				2,51463	1,13455	0.0	2.0	2.0	4.0	4.0
FEDU	649				2,30662	1,09993	0.0	1.0	2.0	3.0	4.0
MJOB	649	5	OTHER	258							
FJOB	649	5	OTHER	367							
REASON	649	4	COURSE	285							
GUARDIAN	649	3	MOTHER	455							
TRAVELTIME	649				1,56856	0,74866	1.0	1.0	1.0	2.0	4.0
STUDYTIME	649				1,93066	0,82951	1.0	1.0	2.0	2.0	4.0
FAILURES	649				0,22188	0,59323	0.0	0.0	0.0	0.0	3.0
SCHOOLSUP	649	2	NO	581							
FAMSUP	649	2	YES	398							
PAID	649	2	NO	610							
ACTIVITIES	649	2	NO	334							
NURSERY	649	2	YES	521							
HIGHER	649	2	YES	580							
INTERNET	649	2	YES	498							
ROMANTIC	649	2	NO	410							
FAMREL	649				3,9306	0,95571	1.0	4.0	4.0	5.0	5.0
FREETIME	649				3,1802	1,05109	1.0	3.0	3.0	4.0	5.0
GOOUT	649				3,1849	1,17576	1.0	2.0	3.0	4.0	5.0
DALC	649				1,5023	0,92483	1.0	1.0	1.0	2.0	5.0
WALC	649				2,2801	1,28438	1.0	1.0	2.0	3.0	5.0
HEALTH	649				3,5362	1,44625	1.0	2.0	4.0	5.0	5.0
ABSENCES	649				3,6594	4,64075	0.0	0.0	2.0	6.0	32.0
G1	649				11,3990	2,74526	0.0	10.0	11.0	13.0	19.0
G2	649				11,5701	2,91363	0.0	10.0	11.0	13.0	19.0
G3	649				11,9060	3,23065	0.0	10.0	12.0	14.0	19.0

Çizelge 4.70’de veri setinde bulunan kolonların veri tipi verilmiştir. Tüm kolonlar incelendiğinde 17 kolon object tipinde, 16 kolon Int64 tipinde tespit edilmiştir.

Çizelge 4.70. Veri Seti II Tüm Kolonların Veri Tipi

	Kolon Tipi
SCHOOL	object
SEX	object
AGE	Int64
ADDRESS	object
FAMSIZE	object
PSTATUS	object
MEDU	Int64
FEDU	Int64
MJOB	object
FJOB	object
REASON	object
GUARDIAN	object
TRAVELTIME	Int64
STUDYTIME	Int64
FAILURES	Int64
SCHOOLSUP	object
FAMSUP	object
PAID	object
ACTIVITIES	object
NURSERY	object
HIGHER	object
INTERNET	object
ROMANTIC	object
FAMREL	Int64
FREETIME	Int64
GOOUT	Int64
DALC	Int64
WALC	Int64
HEALTH	Int64
ABSENCES	Int64
G1	Int64
G2	Int64
G3	Int64

Tez çalışmasında sınıflandırma yapılacağı için çıktı kolonundaki sürekli değerler, ayrıık sınıf değerlerine dönüştürülmüştür. Bunun için Erasmus değişim programı dönüştürüm sistemi kullanılmıştır. Çizelge 4.71’de Erasmus’a göre beş sınıflı dönüştürüm sistemi verilmiştir.

Çizelge 4.71. Beş Sınıflı Dönüştürüm Sistemi

	I	II	III	IV	V
Şehir	Çok İyi	İyi	Yeterli	Uygun	Başarısız
Portekiz / Fransa	16-20	14-15	12-13	10-11	0-9
İrlanda	A	B	C	D	F

G1,G2 ve G3 kolonları 0 ile 20 arasında değer almaktadır. Çizelge 4.70’de yer alan sisteme göre G1,G2 ve G3 kolonların kategorik verilere dönüştürülmesi gerekir. Çizelge 4.68 ve Çizelge 4.69’da verilen durumlar incelendiğinde bu verilerin de kategorik verilere dönüştürülmesi gerekmektedir. G3, G2, G1, FAMSIZE, SCHOOLSUP, FAMSUP, PAID, ACTIVITIES, NURSERY, HIGHER, INTERNET, ROMANTIC, ACTIVITIES kolonları kategorik verilere dönüştürülmüştür. 4.1.2.2. kısımda yer alan bilgilere göre kategorik verilere dönüşen kolonlarda hiyerarşi düzenlemesi yapılmıştır. Hiyerarşiyi göz önünde bulundurarak her bir farklı veri için, sayısal değer belirlenmiştir. Label Encoder yöntemi ile kategorik veriler sayısal verilere dönüştürülmüştür. Çizelge 4.72’de kategorik veriler sayısallaştırılmış hale getirilmiştir

Çizelge 4.72. Kategorik Verilerin Dönüşümü

SCHOOL		SCHOOLSUP		NURSERY		ACTIVITIES	
GP	1	No	1	Yes	2	No	1
MS	2	Yes	2	No	1	Yes	2
SEX		FAMSUP		HIGHER		ACTIVITIES	
F	1	Yes	2	Yes	2	No	1
M	2	No	1	No	1	Yes	2
FAMSIZE		PAID		INTERNET		ROMANTIC	
GT3	2	No	1	Yes	2	No	1
LE3	1	Yes	2	No	1	Yes	2
G1		G2		G3		GUARDIAN	
Good	3	Good	3	Good	3	Mother	2
Vg	4	Vg	4	Vg	4	Father	1
Pass	2	Pass	2	Pass	2	Other	3
Best	5	Best	5	Best	5		

Veri Seti II deki kolon sayımız oldukça fazladır. Dalc kolonu hafta içi kullanılan alkol miktarını göstermektedir. Walc ise hafta sonu içi kullanılan alkol miktarını göstermektedir. Dalc ve Walc kolonları birleştirilmiştir.

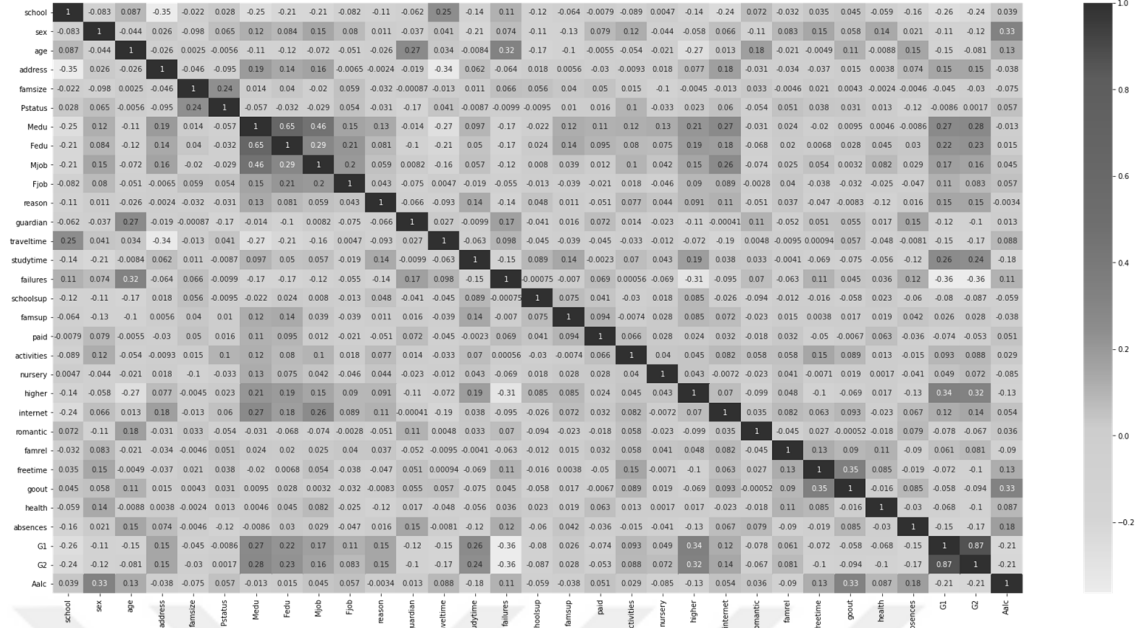
$$Aalc = (Dalc * 5 + Walc * 2)/7 \quad (4.1)$$

(4.1) deki formülü yazarak iki kolonu tek kolona indirilmiştir. Hesaplamalar sonucu yeni elde edilen kolona “Aalc” adı verilmiştir. Çizelge 4.73’da kolonlar arasındaki standart sapma değerleri verilmiştir. Veri setinde standart sapması 0 olan özellikler atılmalıdır. Standart sapmalar göz önünde bulunduğunda 0 olan özellik tespit edilememiştir.

Çizelge 4.73. Kolonlar Arasındaki Standart Sapma Değerleri

Kolonlar	Standart Sapma
SCHOOL	0.476776
SEX	0.492187
AGE	1.218138
ADDRESS	0.460143
FAMSIZE	0.456771
PSTATUS	0.328996
MEDU	1.134552
FEDU	1.099931
MJOB	1.248317
FJOB	0.863487
REASON	1.192045
GUARDAN	0.519187
TRAVELTIME	0.748660
STUDYTIME	0.829510
FAILURES	0.593235
SCHOOLSUP	0.306502
FAMSUP	0.487381
PAID	0.237841
ACTIVITIES	0.500171
NURSERY	0.398212
HIGHER	0.308481
INTERNET	0.422857
ROMANTIC	0.482704
FAMREL	0.955717
FREETIME	1.051093
GOOUT	1.175766
HEALTH	1.446259
ABSENCES	4.640759
G1	1.219359
G2	1.236414
G3	1.252728
AALC	0.932727

Standart sapma analizinden sonra korelasyonu yüksek olan öz niteliklerin tespiti için korelasyon analizi yapılır. Şekil 4.6’da korelasyon matrisi gösterilmiştir.



Şekil 4.6. Veri Seti II Korelasyon Matrisi

G2, G1 özellikleri yüksek düzeyde ilişkili değişkenlerdir. PSTATUS, FAMSUP, FAMSIZE, ROMANTIC, AGE, NURSERY özellikleri düşük düzeyde ilişkili değişkenlerdir. G2 ve G3 özelliklerinin korelasyonu çok yüksek çıktığından veri setinden çıkarılmıştır. Dalc ve Walc kolonlarında Aalc kolonu oluşturulduğundan dolayı kaldırılmıştır. Son olarak kolon sayımız 29 sayıya düşürülmüştür.

4.1.4. Modelleme

CRISP-DM metodolojisinin verinin hazırlanması adımı hazırlanan veri seti modele aktarılmaktadır. Tez çalışmasında Rastgele Orman, Aşırı Gradyan Güçlendirme ve Destek Vektör Makineleri algoritmaları kullanılmıştır. Bu üç yöntem karşılaştırılarak, Öğrencilerin akademik başarısını etkileyen faktörlerin tespit edilmesi için uygun yöntem bulunmaya çalışılmıştır. Tüm veri eğitilmiş, eğitilmiş verilerin %80'i doğrulama için kullanılmıştır. Elimizdeki makine öğrenmesine eğitim için gönderilir, Model elindeki veriyi öğrenir ve öğrendiği verinin üstünden tekrar tekrar geçerek kendisini geliştirir. Model eğitildikten sonra modelin kendisi için eğitimde kullandığı veri ile kendisi ölçülür. Eğer skor yüksek çıkar ise model başarılıdır.

Model tasarlanırken model için kullanılacak parametrelerin belirlenmesi gerekmektedir. Parametre seçimi yapılırken bir dizi parametre belirlenir. Bu parametre dizisine hiper parametre denir. Modelin en iyi sonuçları vermesi için bu parametrelerin hangi birleşimlerde olması gerektiğinin belirlenmesi gerekmektedir. Grid Search algoritması en iyi hiper parametre değerlerinin seçilebilmesi için kullanılan algoritmalarından biridir (Anonymous, 2021). Grid Search algoritması kullanılırken her bir algoritmanın hiper parametre değerleri için bir değer aralığı oluşturulur ve bu değerlerin aralığı k kat çapraz doğrulama ile değerlendirilir. Bu tez çalışmasında Scikit-Learn kütüphanesinde bulunan GridSearchCV metodu yardımıyla veri setleri kullanılarak hiper-parametreler belirlenmiştir. K kat çapraz doğrulama değeri 10 olarak tercih edilmiştir. Hiperparametre seçimi yaparken Rastgele Orman üzerinde RandomizedSearchCV, Aşırı Gradyan Güçlendirme ve Destek Vektör Makineleri için GridSearchCV kullanılmıştır. RandomizedSearchCV deki fark, modelin birleşimleri rastgele seçmesidir.

4.1.1.1. Rastgele Orman

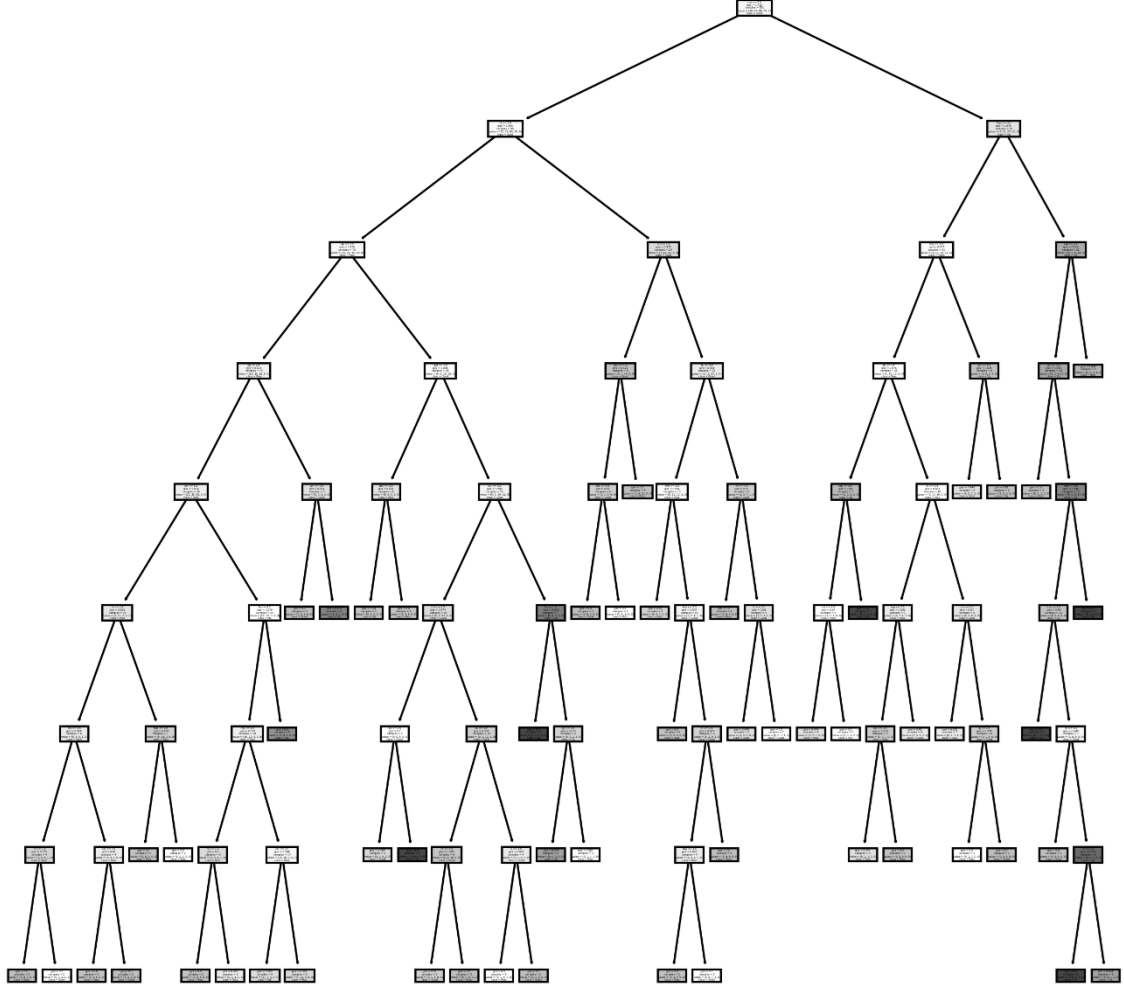
Rastgele Orman sınıflandırılmasını uygulamak için; öncelikle sınıflandırma için uygun parametreler seçilmiştir. Rastgele Orman yönteminin hiperparametre seçimi ve optimizasyonu 10 Kat Çapraz Doğrulama ve RandomizedSearchCV yöntemi ile yapılmıştır. Test edilen hiperparametreler aşağıda verilmiştir;

- max_depth = 2, 4, 6, 8, 10, 20, 30, 40
- max_features = 1, 2, 3, 4, 5, 6, 7, 8
- min_samples_split = 2, 4, 6, 8, 10
- min_samples_leaf = 1, 2, 4
- bootstrap = True, False
- criterion = gini, entropy
- n_estimators = 130, 180, 230

Test edilen hiperparametreler içinden n_estimators için 130, min_samples_split için 4, min_samples_leaf için 2, max_features için 6, max depth için 8, criterion için gini ve bootstrap için True değerleri en iyi sonuçlar verdiği için seçilmiştir. “n_estimator”, ormandaki ağaç sayısını belirtmektedir. “criterion”, ağacın kalitesini ölçmek için kullanılmaktadır. Gini alt ağacın saflık derecesini seçmek için kullanılmaktadır. “max_depth”, ağacın maksimum derinliğidir. “min_samples_leaf” bir yaprakta olması

gereken minimum örnek sayısıdır. “max_features” maksimum bölünmeyi ararken göz önünde bulundurulması gereken özelliklerin sayısıdır. Bootstrap ağaçlar inşa edilirken tüm veri setinin kullanılıp kullanılmayacağını ifade etmektedir. Ayrıca bootstrap yöntemiyle oluşturulan alt ağaçlarda aykırı değer içerme ihtimali düşmektedir.

Şekil 4.7’de oluşturulan ilk alt ağaç gösterilmiştir.



Şekil 4.7. Rastgele Orman Sınıflandırması İçin İlk Alt Ağaç

4.1.1.2. Aşırı Gradyan Güçlendirmesi

Aşırı Gradyan Güçlendirmesini uygulamak için; öncelikle sınıflandırma için uygun parametreler seçilmiştir. Aşırı Gradyan Güçlendirmesi hiperparametre seçimi ve optimizasyonu 10-Kat Çapraz Doğrulama ve GridSearchCV yöntemi ile yapılmıştır. Test edilen hiperparametreler aşağıda verilmiştir;

- max_depth = 3, 4, 5, 6
- learning_rate = 0.1, 0.01, 0.05
- gamma = 0, 0.25, 1
- reg_lambda = 0, 1, 10
- nthread = 4
- n_estimators = 50, 100, 150, 200, 250, 300, 350, 400

Test edilen hiperparametreler içinden n_estimators için 150, min_learning_rate için 0.1, gamma için 0, reg_lambda için 10, max_depth için 5 değerleri en iyi sonuçlar verdiği için seçilmiştir. “nthread” aynı anda paralel kaç thread ile işlem yapacağını göstermektedir. Algoritma çok yavaş çalıştığından dolayı thread şartı eklenmiştir. Hiperparametre seçimi 5.30 saat boyunca çalışmıştır. “n_estimator”, ormandaki ağaç sayısını belirtmektedir. “learning_rate”, yeni oluşacak ağaçların ağırlığını kontrol etmek için kullanılmaktadır. “max_depth”, ağacın maksimum derinliğidir. “reg_lambda” modeli ridge regresyonu için düzenleme terimidir. “gamma” ise bir yaprak düğümünde diğer araçlar arası dengeyi sağlayarak ağacı düzenli hale getirmek için kullanılmaktadır. Gamma değeri arttıkça model dağılımı küçük olmaktadır.

4.1.1.3. Destek Vektör Makineleri

Destek Vektör Makineleri uygulamak için; öncelikle sınıflandırma için uygun parametreler seçilmiştir. Destek Vektör Makineleri uygulamak hiperparametre seçimi ve optimizasyonu 10-Kat Çapraz Doğrulama ve GridSearchCV yöntemi ile yapılmıştır.

Test edilen hiperparametreler aşağıda verilmiştir;

- C = 0.1, 1, 10, 100
- gamma = 0.05, 0.005
- kernel = linear

Test edilen hiperparametreler içinden C için 100, gamma için 0,05, kernel için linear, değerleri en iyi sonuçlar verdiği için seçilmiştir. “C”, parametresi sınıflar arasındaki dengeyi düzenlemek için kullanılmaktadır. C ne kadar büyürse sınıflar arasındaki margin de daralır. ”Gamma” ise modeldeki sınıfların dağılımı kontrol etmek için kullanılmaktadır. Eğer modelimiz aşırı öğrenirse gamma değerini düşürmemiz gerekir. Eğer modelimiz eksik öğrenirse gamma değerini yükseltmek gerekir.

Kernel modelin çekirdeğidir. Linear ise doğrusal olduğunu göstermektedir. Karmaşık veri setlerinde doğrusal çekirdek veri setlerini açıklamak için yeterli olmayabilir. Çekirdek doğrusal dışında polinomal ve radyal temel fonksiyonu da olabilir. Fakat özelliklere atanan katsayıları tespit edebilmek için doğrusal çekirdek kullanılabilir. Diğer çekirdek tiplerinde veriler farklı boyutlarda yer aldığı için özellikler arasındaki önemi tespit etmek oldukça zordur. Bundan dolayı çalışmada sadece doğrusal çekirdek kullanılmıştır.

4.1.5. Değerlendirme

Denetimli makine öğrenmesi yöntemlerinde eğitim verilerinin doğru sonuçlarla eşleşmesi gerekir. Eğitim verilerinin kalitesi modelin doğruluğunu yansıtacağı ve etkileyeceği için önemlidir. Eğitim verilerinin doğruluğu ne kadar yüksek olursa modelin performansı o kadar iyidir. Tüm modellerin genel değerlendirmesi sonuçlar bölümünde verilmiştir.

4.2.5.1. Rastgele Orman Sınıflandırılması Modeli Değerlendirmesi

Rastgele Orman sınıflandırması ile yapılan çalışmalarda en uygun yapılan parametrelerle Veri Seti I için en iyi sonuç 0.93, Veri Seti II için en iyi sonuç 0.89 elde edilmiştir. Başarım sonuçları Veri Seti I için Çizelge 4.74’de, Veri Seti II için Çizelge 4.75’de verilmiştir.

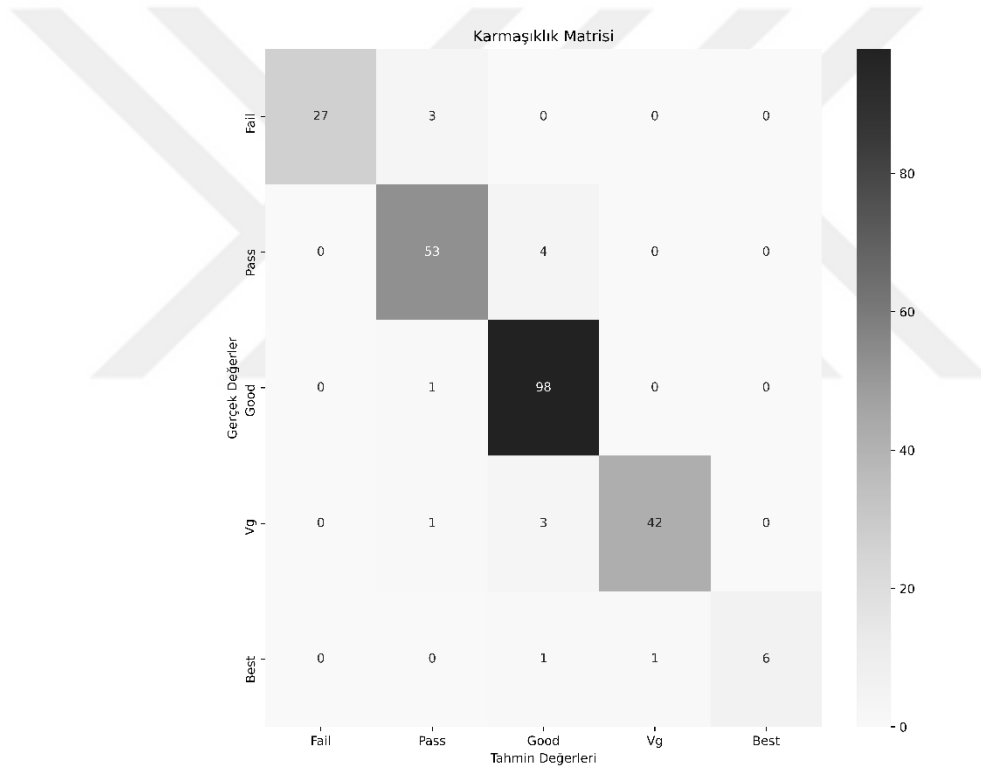
Çizelge 4.74. Veri Seti I İçin Rastgele Orman Sınıflandırmasının Başarımları

Doğruluk Oranı	ACC	0.93
Düzeltilmiş R Kare	R^2	0.91
Ortalama Mutlak Hata	MAE	0.07
Ortalama Kare Hatası	MSE	0.08
Kök Ortalama Kare Hatası	RMSE	0.29
Bağıl Mutlak Hata	RAE	0.08
Bağıl Standart Hata	RSE	0.08

Çizelge 4.75. Veri Seti II İçin Rastgele Orman Sınıflandırmasının Başarımları

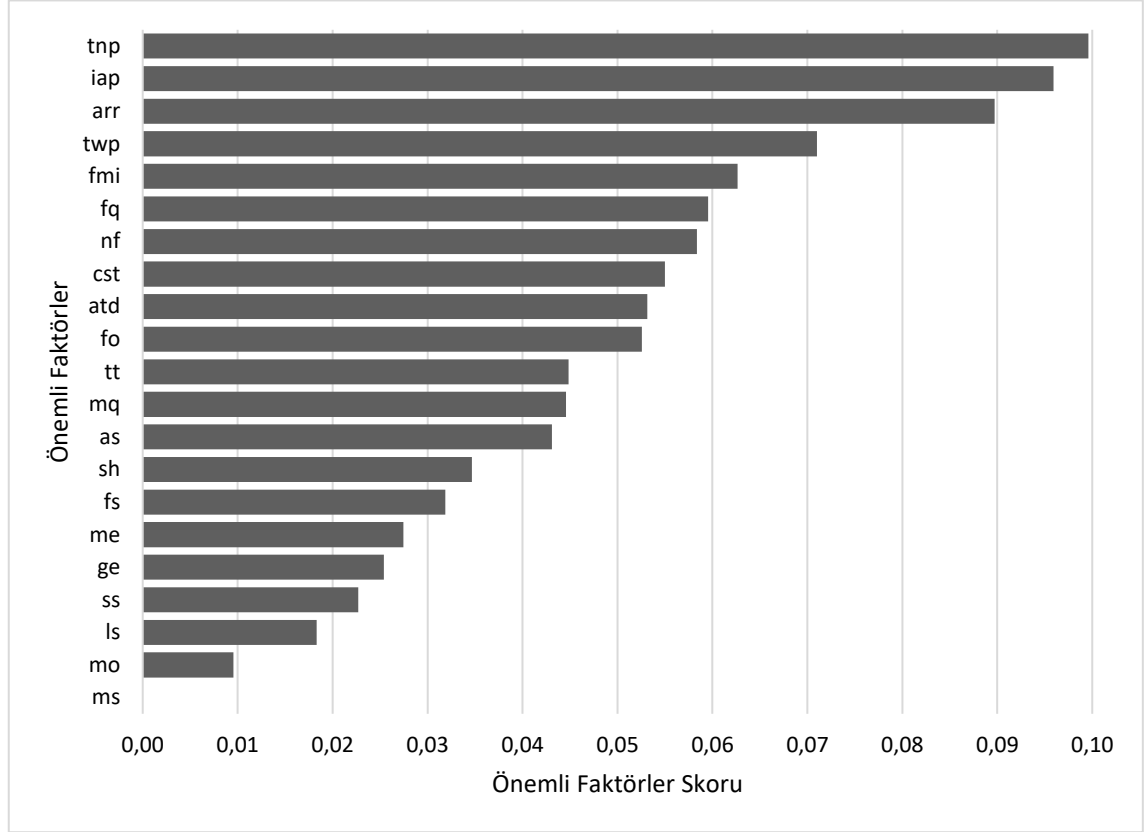
Doğruluk Oranı	ACC	0.89
Düzeltilmiş R Kare	R^2	0.79
Ortalama Mutlak Hata	MAE	0.16
Ortalama Kare Hatası	MSE	0.20
Kök Ortalama Kare Hatası	RMSE	0.55
Bağıl Mutlak Hata	RAE	0.15
Bağıl Standart Hata	RSE	0.20

Şekil 4.8’de Rastgele orman sınıflandırmasının Veri Seti I için karmaşıklık matrisi verilmiştir. Şekil 4.8’e göre elde edilen Precision, Recall ve F1 skor değerleri Çizelge 4.76.’de gösterilmiştir.

**Şekil 4.8.** Veri Seti I İçin Rastgele Orman Sınıflandırması Modelinin Karmaşıklık Matrisi**Çizelge 4.76.** Veri Seti I İçin Rastgele Orman Sınıflandırması Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri

	Precision	Recall	F1-Score
Fail	1.00	0.90	0.94
Pass	0.91	0.93	0.92
Good	0.92	0.99	0.95
Vg	0.97	0.91	0.94
Best	1.00	0.75	0.85

Fail sınıfı için F-1 skoru 0.94 çıkmıştır. Pass sınıfı için F-1 skoru 0.92 çıkmıştır. Good sınıfı için F-1 skoru 0.95 çıkmıştır. Vg sınıfı için F-1 skoru 0.94 çıkmıştır. Best sınıfı için F-1 skoru 0.94 çıkmıştır. Tüm sınıflar incelendiğinde en yüksek başarı Good sınıfında, en düşük başarı Best sınıfında elde edilmiştir. Coefficient katsayıları göz önüne bulundurulduğunda her bir özelliğin önemi ortaya çıkmaktadır. Şekil 4.9.'de katsayılar ile kolonlar bar grafiği şeklinde gösterilmiştir.



Şekil 4.9. Veri Seti I İçin Rastgele Orman Sınıflandırması İle Önemli Faktörler

Şekil 4.9 incelendiğinde en önemli Veri Seti I için iki faktörün TNP ve IAP olduğu ortaya çıkmıştır. TNP ve IAP faktörü öğrencinin eski değerlendirmeleri üzerinden elde edilmektedir. Sosyal Faktörler incelediğinde FMI, FQ, NF en önemli faktörler olarak tespit edilmiştir. Model sonucunda ortaya çıkan Veri Seti I için önemli özelliklerin sıralaması;

MS < MO < LS < SS < GE < ME < FS < SH < AS < MQ < TT < FO < ATD < CST < NF < FQ < FMI < TWP < ARR < IAP < TNP
olarak bulunmuştur.

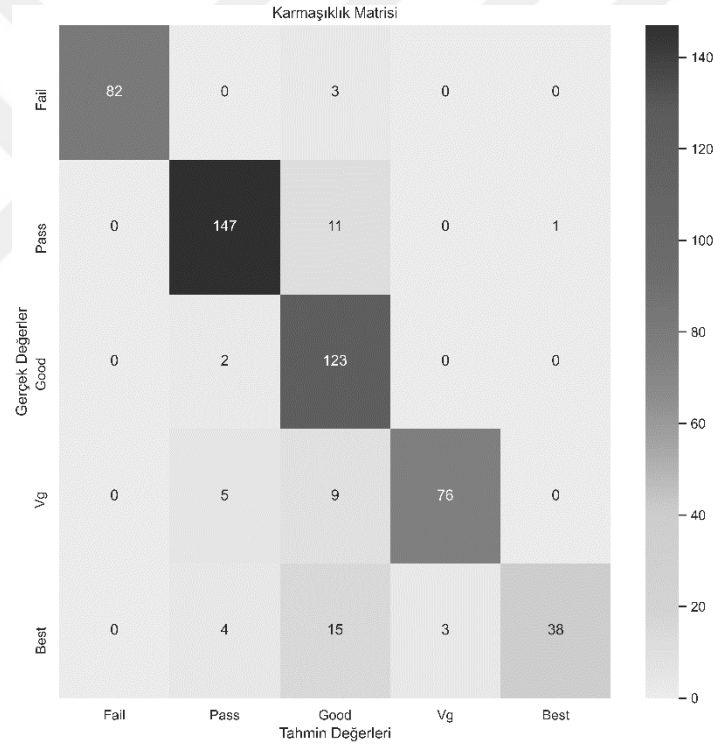
Şekil 4.9'a göre Veri Seti I için öğrencinin akademik başarısına etki eden en önemli on faktörler sırasıyla;

1. 10.Sınıf Ortalama Yüzdesi,
2. İç Değerlendirme Ortalama Yüzdesi,

3. Önceki Dönemden Dersi Olup Olmadığı,
4. 12.Sınıf Ortalama Yüzdesi,
5. Aile Aylık Geliri
6. Baba Eğitim Seviyesi
7. Arkadaş Sayısı
8. Kast Sınıfı
9. Sınıfa Devam Yüzdesi
10. Baba Mesleği

faktörleridir.

Şekil 4.10’da Rastgele orman sınıflandırmasının Veri Seti II için karmaşıklık matrisi verilmiştir. Şekil 4.10’a göre elde edilen Precision, Recall ve F1 skor değerleri Çizelge 4.77.’de gösterilmiştir.



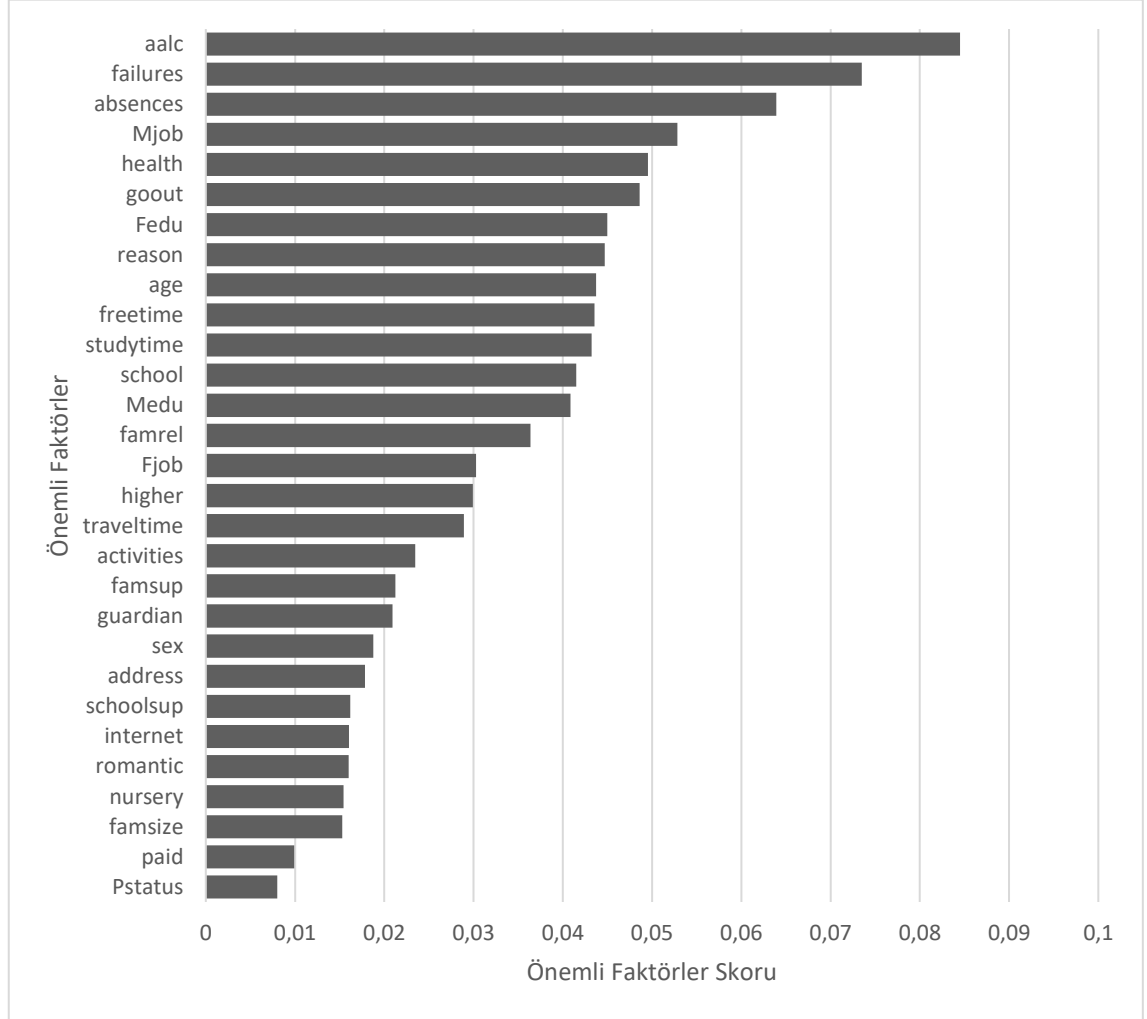
Şekil 4.10. Veri Seti II İçin Rastgele Orman Sınıflandırması Modelinin Karmaşıklık Matrisi

Çizelge 4.77. Veri Seti II İçin Rastgele Orman Sınıflandırması Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri

	Precision	Recall	F1-Score
Fail	1.00	0.96	0.98
Pass	0.93	0.92	0.92
Good	0.76	0.98	0.86
Vg	0.96	0.84	0.89
Best	0.97	0.63	0.76

Fail sınıfı için F-1 skoru 0.98 çıkmıştır. Pass sınıfı için F-1 skoru 0.92 çıkmıştır. Good sınıfı için F-1 skoru 0.86 çıkmıştır. Vg sınıfı için F-1 skoru 0.89 çıkmıştır. Best sınıfı için F-1 skoru 0.76 çıkmıştır.

Tüm sınıflar incelendiğinde en yüksek başarı Fail sınıfında, en düşük başarı Best sınıfında elde edilmiştir. Coefficient katsayıları göz önüne bulundurulduğunda her bir özelliğin önemi ortaya çıkmaktadır. Şekil 4.11.'de katsayılar ile kolonlar bar grafiği şeklinde gösterilmiştir.



Şekil 4.11. Veri Seti II İçin Rastgele Orman Sınıflandırması İle Önemli Faktörler

Şekil 4.11 incelendiğinde Veri Seti II için en önemli faktörün AALC olduğu ortaya çıkmıştır. Diğer faktörler incelendiğinde FAILURES, ABSENCES, MJOB önemli faktörler olarak tespit edilmiştir. Model sonucunda ortaya çıkan Veri Seti II için önemli özelliklerin sıralaması;

PSTATUS < PAID < FAMSIZE < NURSERY < ROMANTIC < INTERNET < SCHOOLSUB < ADRESS < SEX < GUARDIAN < FAMSUP < ACTIVITIES <

TRAVELTIME < GE < HIGHER < FJOB < FAMREL < MEDU < SCHOOL < STUDYTIME < FREETIME < AGE < REASON < FEDU < GOOUT < HEALTH < MJOB < ABSENCES < FAILURES < AALC

olarak bulunmuştur.

Şekil 4.11'e göre Veri Seti II için öğrencinin akademik başarısına etki eden en önemli on faktörler sırasıyla;

1. Alkol Tüketimi,
2. Önceki Dönemlerden Kalan Ders Sayısı,
3. Devamsızlık Sayısı,
4. Anne Eğitim Seviyesi,
5. Sağlık Durumu,
6. Dışarı Çıkma
7. Baba Mesleği
8. Okulu Seçme Sebebi
9. Yaş
10. Boş Zaman

faktörleridir.

4.2.5.2. Aşırı Gradyan Güçlendirmesi Sınıflandırılması Modeli Değerlendirmesi

Aşırı Gradyan Güçlendirmesi sınıflandırması ile yapılan çalışmalarda en uygun parametrelerle Veri Seti I için en iyi sonuç 0.97, Veri Seti II için en iyi sonuç 0.97 olmuştur. Başarım sonuçları Veri Seti I için Çizelge 4.78'de, Veri Seti II için Çizelge 4.79'da verilmiştir.

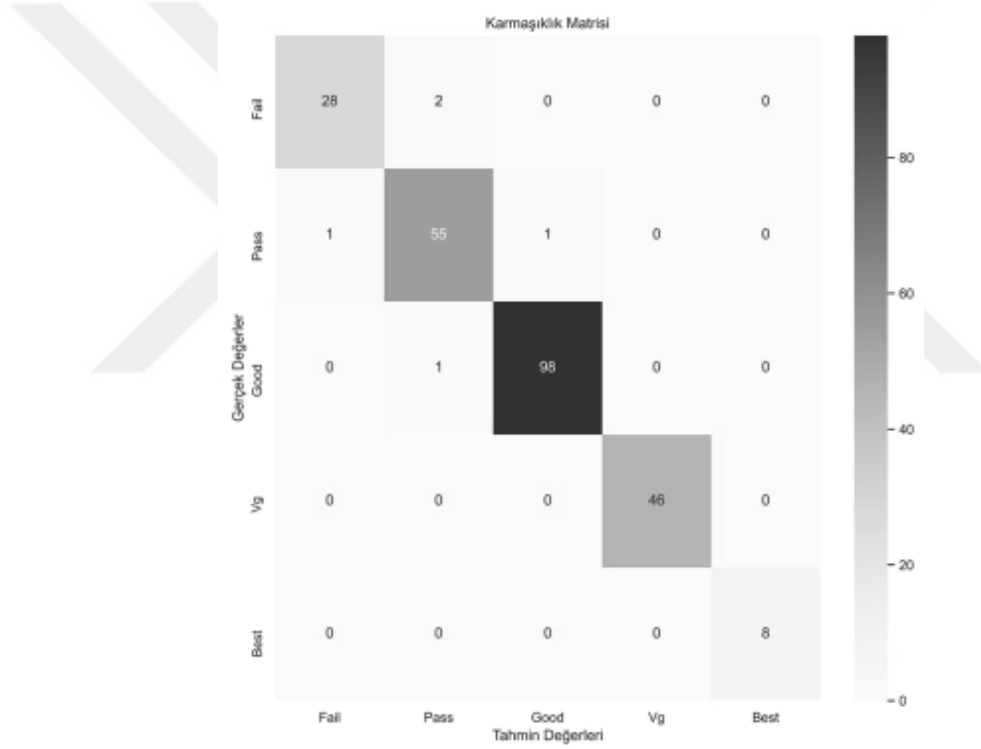
Çizelge 4.78. Veri Seti I İçin Aşırı Gradyan Güçlendirmesi Sınıflandırmasının Başarımları

Doğruluk Oranı	ACC	0.97
Düzeltilmiş R Kare	R2	0.97
Ortalama Mutlak Hata	MAE	0.02
Ortalama Kare Hatası	MSE	0.02
Kök Ortalama Kare Hatası	RMSE	0.14
Bağıl Mutlak Hata	RAE	0.02
Bağıl Standart Hata	RSE	0.02

Çizelge 4.79. Veri Seti II İçin Aşırı Gradyan Güçlendirmesi Sınıflandırmasının Başarımları

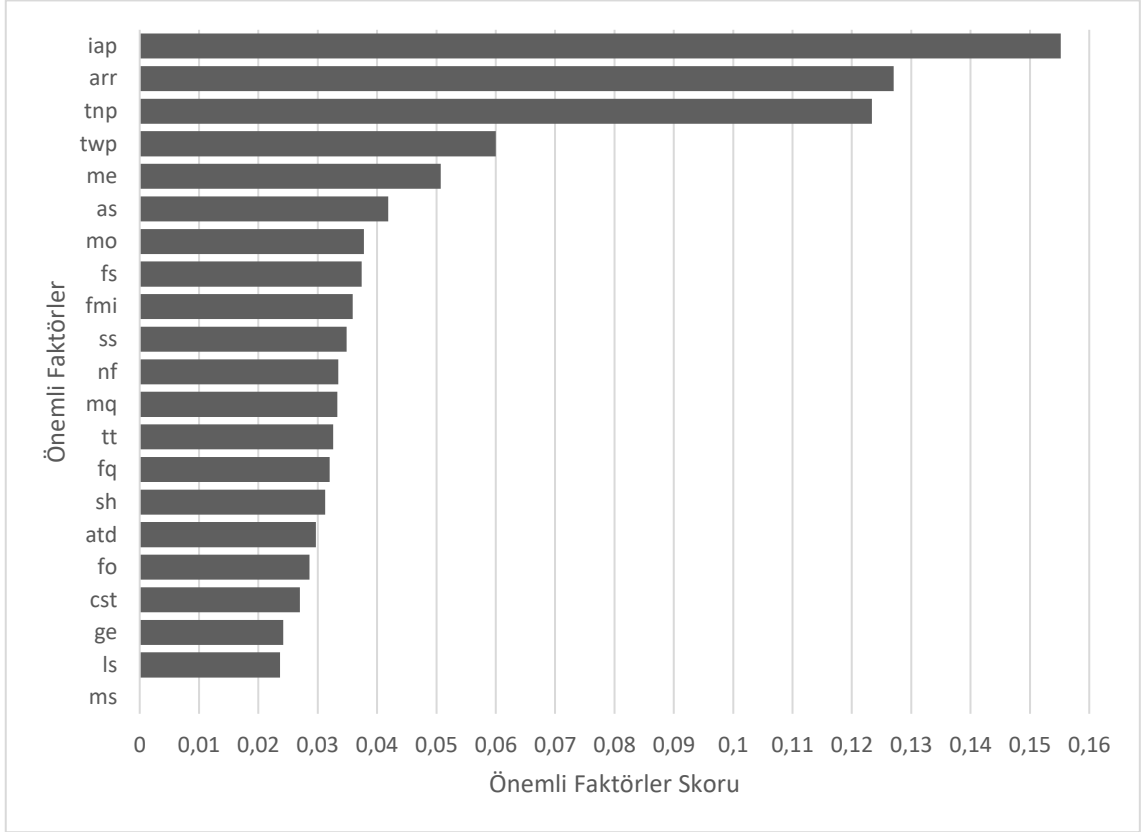
Doğruluk Oranı	ACC	0.97
Düzeltilmiş R Kare	R2	0.95
Ortalama Mutlak Hata	MAE	0.03
Ortalama Kare Hatası	MSE	0.04
Kök Ortalama Kare Hatası	RMSE	0.25
Bağıl Mutlak Hata	RAE	0.03
Bağıl Standart Hata	RSE	0.04

Şekil 4.12’de göre Aşırı Gradyan Güçlendirmesi sınıflandırmasının Veri Seti I için karmaşıklık matrisi verilmiştir. Şekil 4.12’ye göre elde edilen Precision, Recall ve F1 skorları Çizelge 4.80.’de gösterilmiştir.

**Şekil 4.12.** Veri Seti I İçin Aşırı Gradyan Güçlendirmesi Modelinin Karmaşıklık Matrisi**Çizelge 4.80.** Veri Seti I İçin Aşırı Gradyan Güçlendirmesi Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri

	Precision	Recall	F1-Score
Fail	0,93	0.97	0.95
Pass	0,96	0.95	0.96
Good	0,99	0.99	0.99
Vg	1	1	1
Best	1	1	1

Fail sınıfı için F-1 skoru 0.95 çıkmıştır. Pass sınıfı için F-1 skoru 0.96 çıkmıştır. Good sınıfı için F-1 skoru 0.99 çıkmıştır. Vg sınıfı için F-1 skoru 0.99 çıkmıştır. Best sınıfı için F-1 skoru 1 çıkmıştır. Tüm sınıflar incelendiğinde en yüksek başarı Vg ve Best sınıflarında, en düşük başarı Fail sınıfında elde edilmiştir. Coefficient katsayıları göz önüne bulundurulduğunda her bir özelliğin önemi ortaya çıkmaktadır. Şekil 4.13.'te katsayılar ile kolonlar bar grafiği şeklinde gösterilmiştir.



Şekil 4.13. Veri Seti I İçin Aşırı Gradyan Güçlendirmesi Sınıflandırması İle Önemli Faktörler

Şekil 4.13 incelediğinde Veri Seti I için en önemli faktörün IAP olduğu ortaya çıkmıştır. Ardından ARR ve TNP faktörlerinin yakın derecede etki ettiği görülmektedir. IAP, ARR ve TNP faktörü öğrencinin eski değerlendirmeleri üzerinden elde edilmektedir. Sosyal Faktörler incelediğinde ME en önemli faktör olarak tespit edilmiştir. ME'nin ardından ise AS, MO, FS önemli olarak gelmektedir. Model sonucunda ortaya çıkan önemli özelliklerin sıralaması;

MS < LS < GE < CST < FO < ATD < SH < FQ < TT < MQ < NF < SS < FMI < FS < MO < AS < ME < TWP < TNP < ARR < IAP

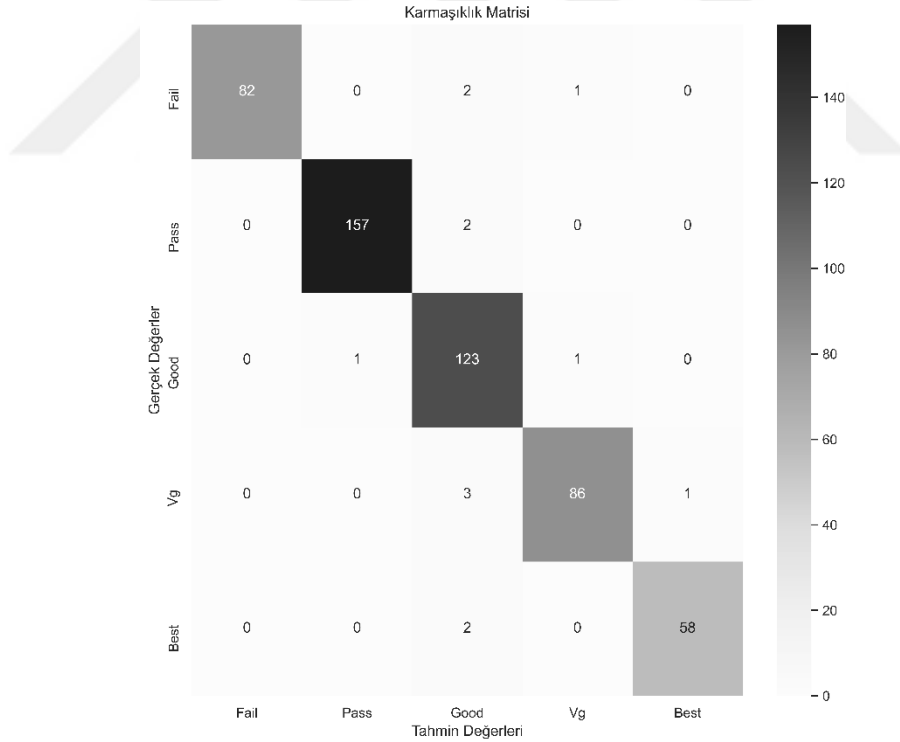
olarak bulunmuştur.

Şekil 4.13'e göre Veri Seti I için öğrencinin akademik başarısına etki eden en önemli on faktörler sırasıyla;

1. İç Değerlendirme Ortalama Yüzdesi,
2. Önceki Dönemden Dersi Olup Olmadığı,
3. 10.Sınıf Ortalama Yüzdesi,
4. 12.Sınıf Ortalama Yüzdesi,
5. Sınav Dili,
6. Öğrenci Okula Giriş Tipi,
7. Anne Mesleği,
8. Aile Genişliği,
9. Aile Aylık Geliri,
10. 10.Sınıf'a Geldiğinde Geldiği Okul Tipi

faktörleridir.

Şekil 4.14'de Aşırı Gradyan Güçlendirmesi sınıflandırmasının Veri Seti II için karmaşıklık matrisi verilmiştir. Şekil 4.14'e göre elde edilen Precision, Recall ve F1 skorları Çizelge 4.81.'de hesaplamalar gösterilmiştir.

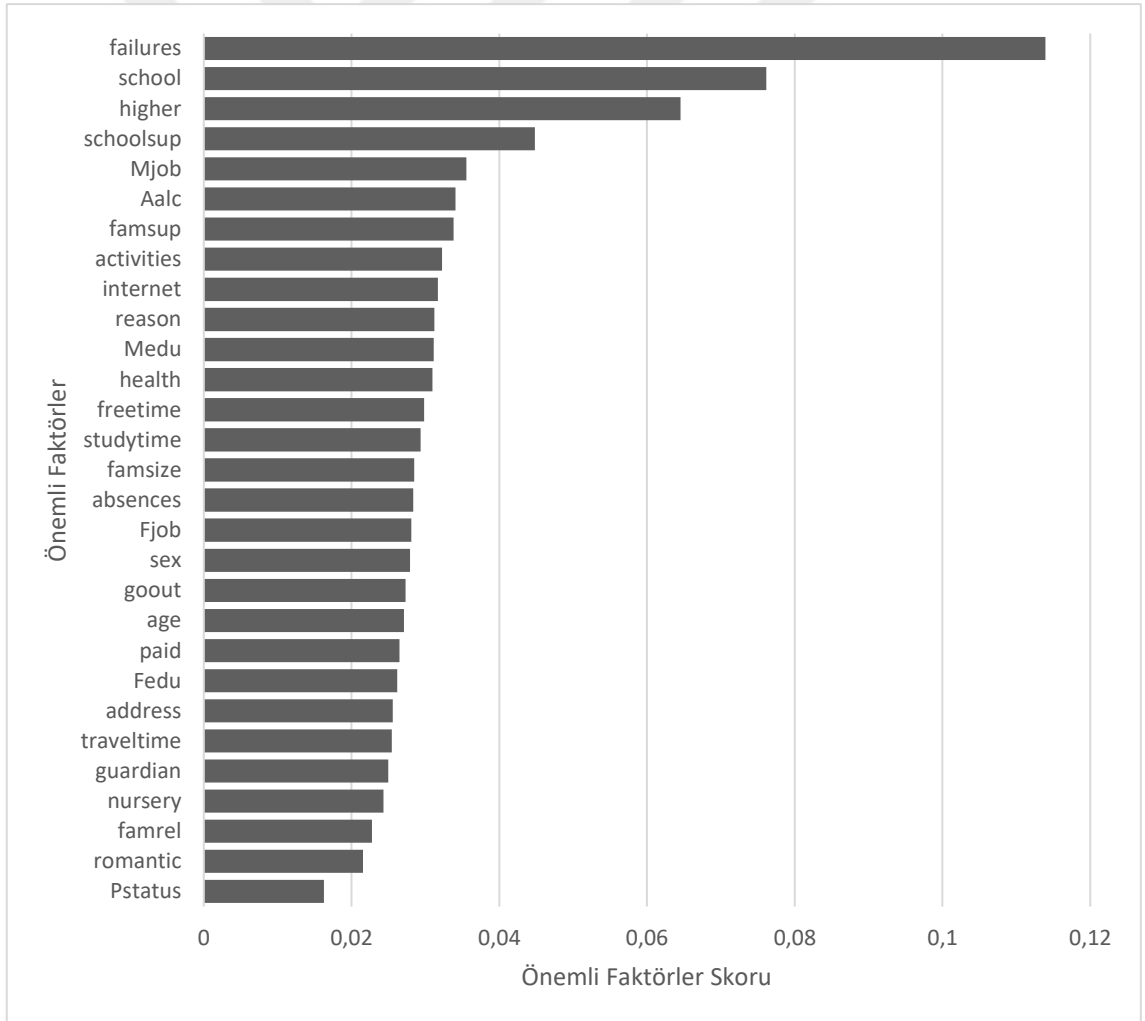


Şekil 4.14. Veri Seti II İçin Aşırı Gradyan Güçlendirmesi Modelinin Karmaşıklık Matrisi

Çizelge 4.81. Veri Seti II İçin Aşırı Gradyan Güçlendirme Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri

	Precision	Recall	F1-Score
Fail	1.00	0.96	0.98
Pass	0.99	0.98	0.99
Good	0.93	0.98	0.95
Vg	0.97	0.95	0.96
Best	0.98	0.96	0.97

Fail sınıfı için F-1 skoru 0.98 çıkmıştır. Pass sınıfı için F-1 skoru 0.99 çıkmıştır. Good sınıfı için F-1 skoru 0.95 çıkmıştır. Vg sınıfı için F-1 skoru 0.96 çıkmıştır. Best sınıfı için F-1 skoru 0,97 çıkmıştır. Tüm sınıflar incelendiğinde en yüksek başarı Pass sınıfında, en düşük başarı Good sınıfında elde edilmiştir. Coefficient katsayıları göz önünde bulundurulduğunda her bir özelliğin önemi ortaya çıkmaktadır. Şekil 4.15.'de katsayılar ile kolonlar bar grafiği şeklinde gösterilmiştir.



Şekil 4.15. Veri Seti II İçin Aşırı Gradyan Güçlendirme Sınıflandırması İle Önemli Faktörler

Şekil 4.15'i incelendiğinde Veri Seti II İçin en önemli faktörün FAILURES olduğu ortaya çıkmıştır. Ardından SCHOOL ve HIGHER faktörlerinin yüksek derecede etki ettiği görülmektedir. Model sonucunda ortaya çıkan önemli özelliklerin sıralaması;

PSTATUS < ROMANTIC < FAMREL < NURSERY < GUARDIAN < TRAVELTIME < ADRESS < FEDU< PAID < AGE < GOOUT < SEX < FJOB < ABSENCES < FAMSIZE < STUDYTIME < FREETIME < HEALTH < MEDU < REASON < INTERNET < ACTIVITES < FAMSUP < AALC < MJOB < SCHOOLSUP < HIGHER < SCHOOL < FAILURES

olarak bulunmuştur.

Şekil 4.15'e göre Veri Seti II için öğrencinin akademik başarısına etki eden en önemli on faktörler sırasıyla;

1. Önceki Dönemlerden Kalan Ders Sayısı,
2. Okul,
3. Yüksek Eğitim Almak İstemesi,
4. Okul Desteği,
5. Anne Eğitim Seviyesi,
6. Alkol Tüketimi,
7. Ailenin Eğitim Desteği,
8. Etkinliklere Katılma,
9. Evde İnternet Erişimi,
10. Okulu Seçme Sebebi

faktörleridir.

4.2.5.3. Destek Vektör Makineleri Modeli Değerlendirmesi

Destek Vektör Makineleri sınıflandırması ile yapılan çalışmalarda en uygun parametrelerle yapılan çalışmada Veri Seti I için en iyi sonuç 0.71, Veri Seti II için en iyi sonuç 0.49 olmuştur. Başarım sonuçları Veri Seti I için Çizelge 4.82'de, Veri Seti II için Çizelge 4.83'de verilmiştir.

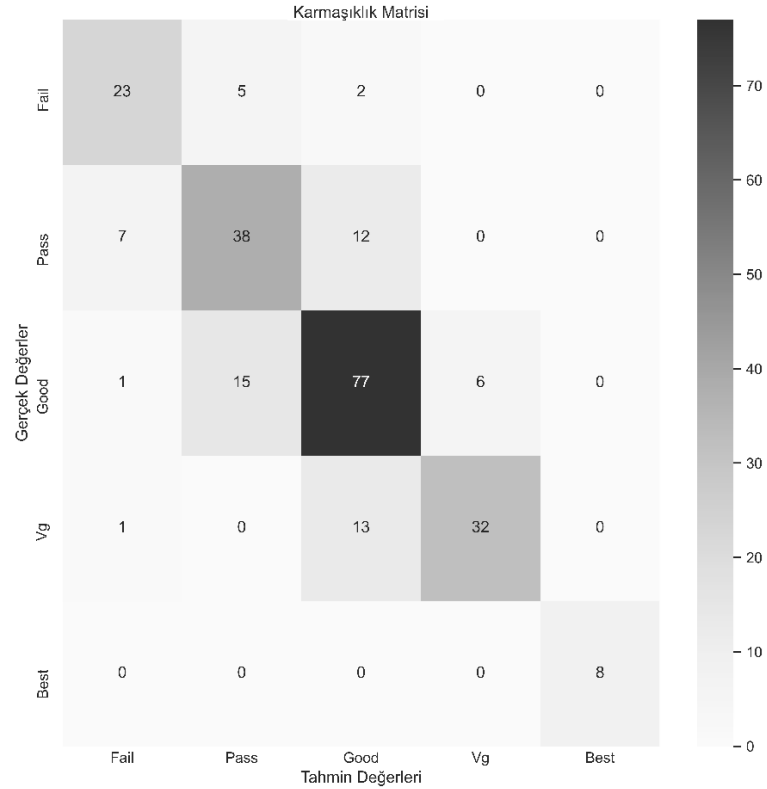
Çizelge 4.82. Veri Seti I İçin Destek Vektör Makineleri Sınıflandırmasının Başarımları

Doğruluk Oranı	ACC	0.71
Düzeltilmiş R Kare	R2	0.67
Ortalama Mutlak Hata	MAE	0.27
Ortalama Kare Hatası	MSE	0.32
Kök Ortalama Kare Hatası	RMSE	0.57
Bağıl Mutlak Hata	RAE	0.34
Bağıl Standart Hata	RSE	0.32

Çizelge 4.83. Veri Seti II İçin Destek Vektör Makineleri Sınıflandırmasının Başarımları

Doğruluk Oranı	ACC	0.49
Düzeltilmiş R Kare	R2	0.27
Ortalama Mutlak Hata	MAE	0.67
Ortalama Kare Hatası	MSE	0.72
Kök Ortalama Kare Hatası	RMSE	1.06
Bağıl Mutlak Hata	RAE	0.64
Bağıl Standart Hata	RSE	0.72

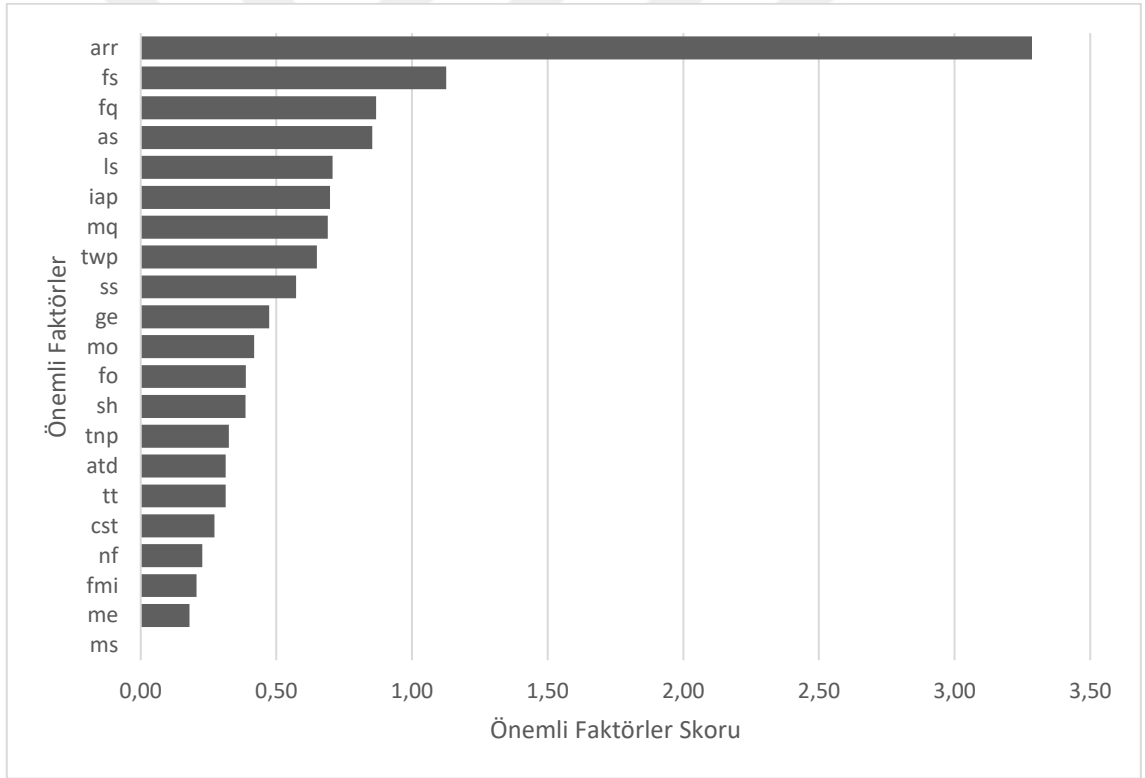
Şekil 4.16’de Destek Vektör Makineleri sınıflandırmasının Veri Seti I için karmaşıklık matrisi verilmiştir. Şekil 4.16’e göre elde edilen Precision, Recall ve F1 skorları Çizelge 4.84.’de hesaplamalar gösterilmiştir.

**Şekil 4.16.** Veri Seti I İçin Destek Vektör Makineleri Modelinin Karmaşıklık Matrisi

Çizelge 4.84. Veri Seti I İçin Destek Vektör Makineleri Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri

	Precision	Recall	F1-Score
Fail	0.77	0.72	0.74
Pass	0.67	0.66	0.66
Good	0.78	0.74	0.76
Vg	0.70	0.84	0.76
Best	1.00	1.00	1.00

Fail sınıfı için F-1 skoru 0.74 çıkmıştır. Pass sınıfı için F-1 skoru 0.66 çıkmıştır. Good sınıfı için F-1 skoru 0.76 çıkmıştır. Vg sınıfı için F-1 skoru 0.76 çıkmıştır. Best sınıfı için F-1 skoru 1 çıkmıştır. Tüm sınıflar incelendiğinde en yüksek başarı Best sınıfında, en düşük başarı Fail sınıfında elde edilmiştir. Coefficient katsayıları göz önüne bulundurulduğunda her bir özelliğin önemi ortaya çıkmaktadır. Şekil 4.17.'de katsayılar ile kolonlar bar grafiği şeklinde gösterilmiştir.



Şekil 4.17. Veri Seti II İçin Destek Vektör Makineleri Sınıflandırması İle Önemli Faktörler

Şekil 4.17 incelendiğinde en önemli faktörün ARR olduğu ortaya çıkmıştır. Diğer Faktörler incelendiğinde FS en önemli faktör olarak tespit edilmiştir. FS ardından ise FQ ve FQ önemli olarak gelmektedir. Model sonucunda ortaya çıkan önemli özelliklerin sıralaması;

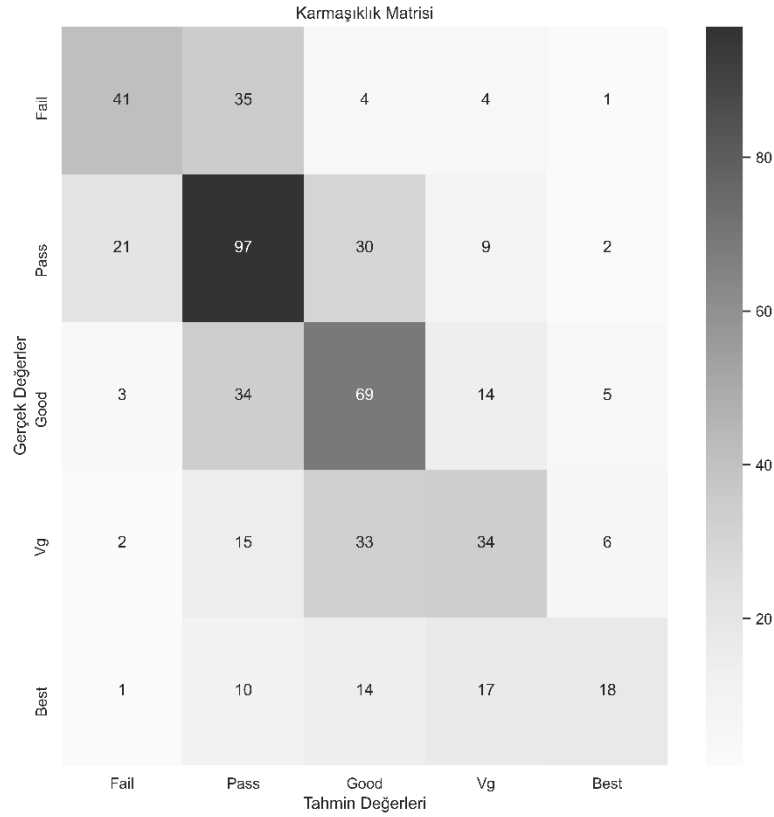
MS < ME < FMI < NF < CST < TT < ATD < TNP < SH < FO < MO < GE < SS
< TWP < MQ < IAP < LS < AS < FQ < FS < ARR

Şekil 4.17'ye göre Veri Seti I için öğrencinin akademik başarısına etki eden en önemli on faktörler sırasıyla;

1. Önceki Dönemden Dersi Olup Olmadığı,
2. Aile Genişliği,
3. Baba Eğitim Seviyesi,
4. Öğrenci Okula Giriş Tipi,
5. Şehir Veya Köyde Yaşaması,
6. İç Değerlendirme Ortalama Yüzdesi,
7. Anne Eğitim Seviyesi,
8. 12.Sınıf Ortalama Yüzdesi,
9. 10.Sınıf'a Geldiğinde Geldiği Okul Tipi,
10. Cinsiyet

faktörleridir.

Şekil 4.18'de Destek Vektör Makineleri sınıflandırmasının Veri Seti II için karmaşıklık matrisi verilmiştir. Şekil 4.18'e göre elde edilen Precision, Recall ve F1 skorları Çizelge 4.85'de gösterilmiştir.

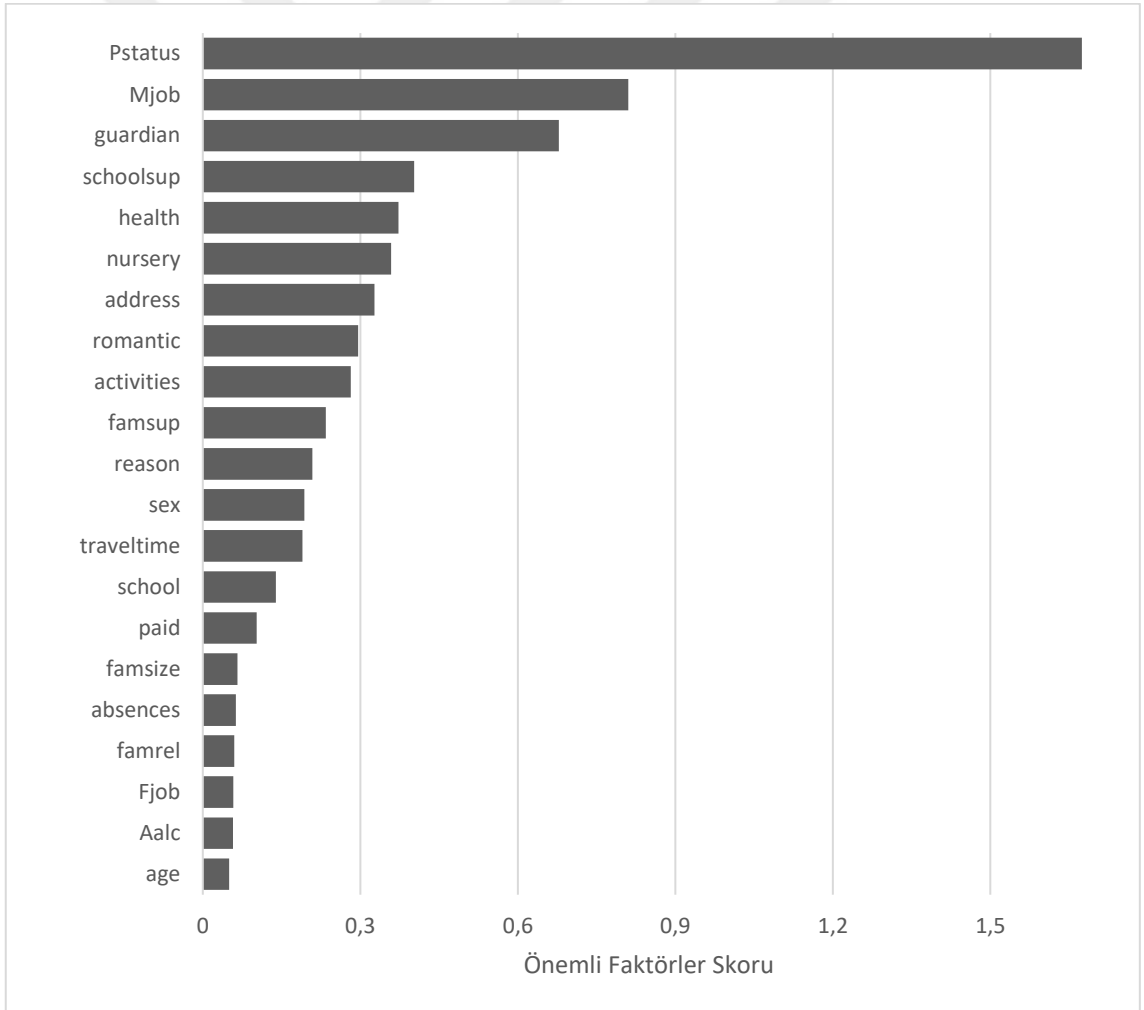


Şekil 4.18. Veri Seti II İçin Destek Vektör Makineleri Modelinin Karmaşıklık Matrisi

Çizelge 4.85. Veri Seti II İçin Destek Vektör Makineleri Modelinin Hedef Sınıfların Precision, Recall ve F1 Skor Değerleri

	Precision	Recall	F1-Score
Fail	0.48	0.60	0.53
Pass	0.61	0.51	0.55
Good	0.55	0.46	0.50
Vg	0.38	0.44	0.40
Best	0,30	0.56	0.39

Fail sınıfı için F-1 skoru 0.53 çıkmıştır. Pass sınıfı için F-1 skoru 0.55 çıkmıştır. Good sınıfı için F-1 skoru 0.50 çıkmıştır. Vg sınıfı için F-1 skoru 0.40 çıkmıştır. Best sınıfı için F-1 skoru 0.39 çıkmıştır. Tüm sınıflar incelendiğinde en yüksek başarı Pass sınıfında, en düşük başarı Best sınıfında elde edilmiştir. Coefficient katsayıları göz önüne bulduğumuzda her bir özelliğin önemi ortaya çıkmaktadır. Şekil 4.19.'de katsayılar ile kolonlar bar grafiği şeklinde gösterilmiştir.



Şekil 4.19. Veri Seti II İçin Destek Vektör Makineleri Sınıflandırması İle Önemli Faktörler

Şekil 4.19 incelediğinde en önemli faktörün PSTATUS olduğu ortaya çıkmıştır. Diğer Faktörler incelediğinde MJOB en önemli faktör olarak tespit edilmiştir. MJOB ardından ise GUARDIAN önemli olarak gelmektedir.

Model sonucunda ortaya çıkan önemli özelliklerin sıralaması;
 AGE < AALC < FJOB < FAMREL < ABSENCES < FAMSIZE < PAID < SCHOOL < TRAVELTIME < SEX < REASON < FAMSUP < ACTIVITIES < ADDRESS < NURSERY < HEALTH < SCHOOLSUP < GUARDIAN < MJOB < PSTATUS

Şekil 4.15'e göre Veri Seti II için öğrencinin akademik başarısına etki eden en önemli on faktörler sırasıyla;

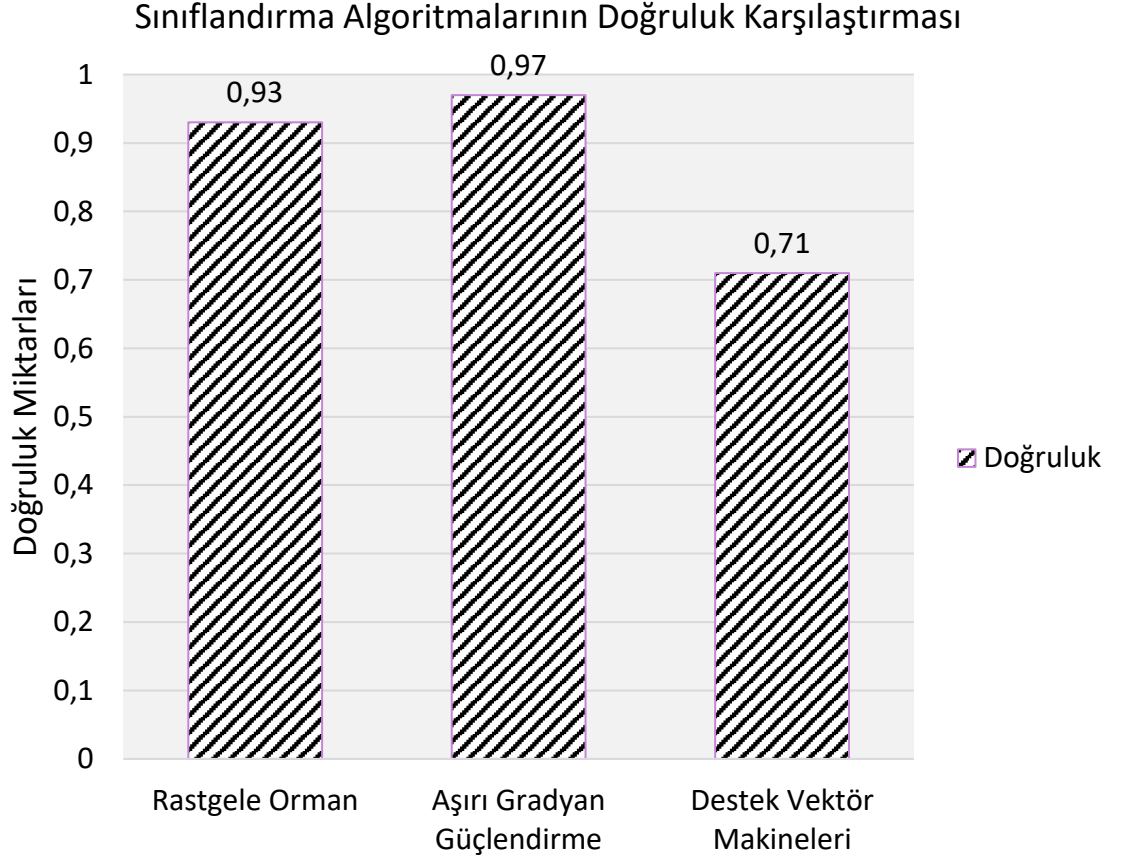
1. Ebeveyn'leri Birlikte olma durumu,
2. Anne Eğitim Seviyesi,
3. Veli,
4. Okul Desteği,
5. Sağlık Durumu,
6. Anaokuluna Gidip Gitmediği,
7. Adres,
8. Etkinliklere Katılma,
9. Ailenin Eğitim Desteği,
10. Okulu Seçme Sebebi

faktörleridir.

4.2.5.4. Genel Değerlendirme

4.2.5.4.1. Veri Seti I Değerlendirmesi

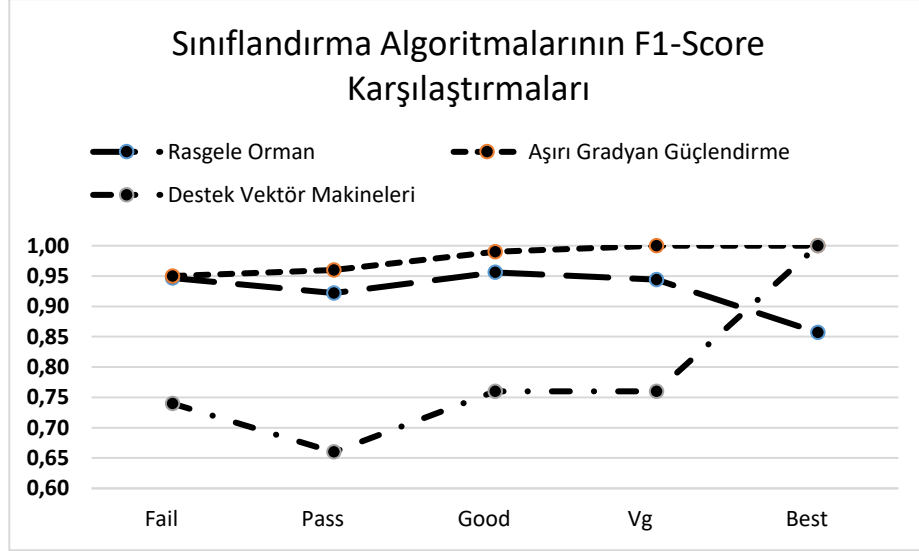
Değerlendirme adımında, bu çalışmanın ikinci araştırma sorusu olan “Başarıyı etkileyen faktörlerin tespitinde kullanılan modellerin doğrulukları nelerdir?” sorusunun cevabı Veri Seti I için Şekil 4.20’de gösterilmiştir.



Şekil 4.20. Veri Seti I İçin Sınıflandırma Algoritmalarının Karşılaştırılması

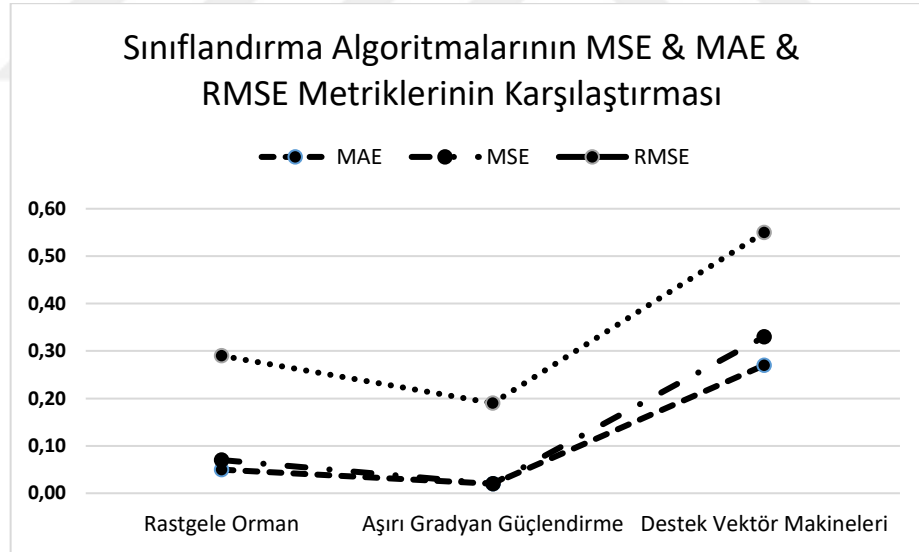
Üç sınıflandırma modelinin doğruluğu karşılaştırıldığında, Aşırı Gradyan Güçlendirme % 97 doğruluk ile diğer sınıflandırma yöntemlerinden daha başarılı olarak sınıflandırma yapmıştır.

Veri Seti I için Karmaşıklık matrisleri incelendiği zaman F1 Score kriterlerinde en yüksek değeri Aşırı Gradyan Güçlendirme vermiştir. Şekil 4.21’de F1 Score karşılaştırma grafiği gösterilmiştir

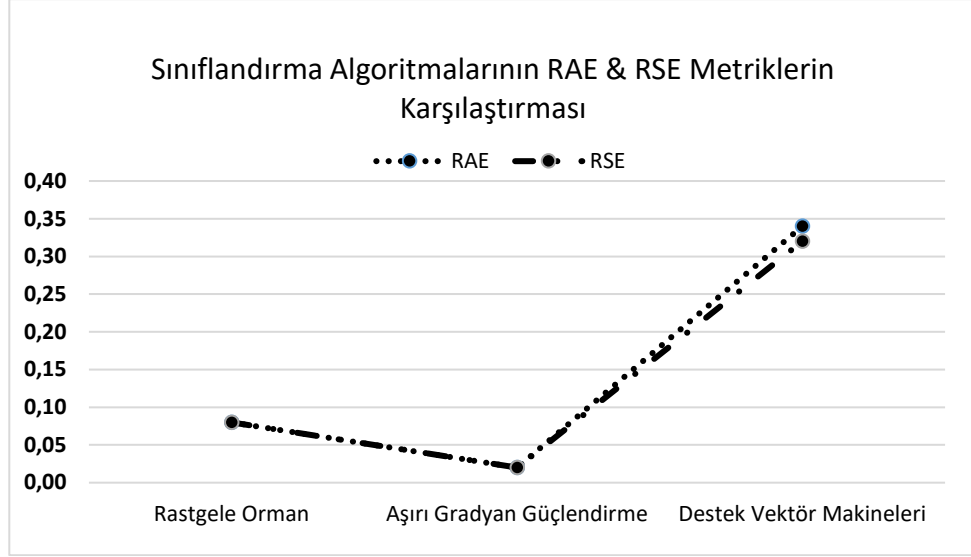


Şekil 4.21. Veri Seti I İçin Sınıflandırma Algoritmalarının F1-Score Karşılaştırmaları

Üçüncü soru olan “Hangi model ne kadar başarılı olmuştur?” sorusunun Veri Seti I için cevabı Şekil 4.22 ve 4.23’te gösterilmiştir. Aşırı Gradyan Güçlendirme sınıflandırma modelinin Veri Seti I için, diğerleriyle karşılaştırıldığında, Ortalama Mutlak Hata (MAE), Ortalama Kare Hatası (MSE), Kök Ortalama Kare Hatası (RMSE), Bağlı Mutlak Hata (RAE) ve Bağlı Standart Hata (RSE) açılarından minimum hatalara sahip olduğunu görülmektedir.



Şekil 4.22. Veri Seti I İçin Sınıflandırma Algoritmalarının MSE,MAE ve RMSE Karşılaştırmaları

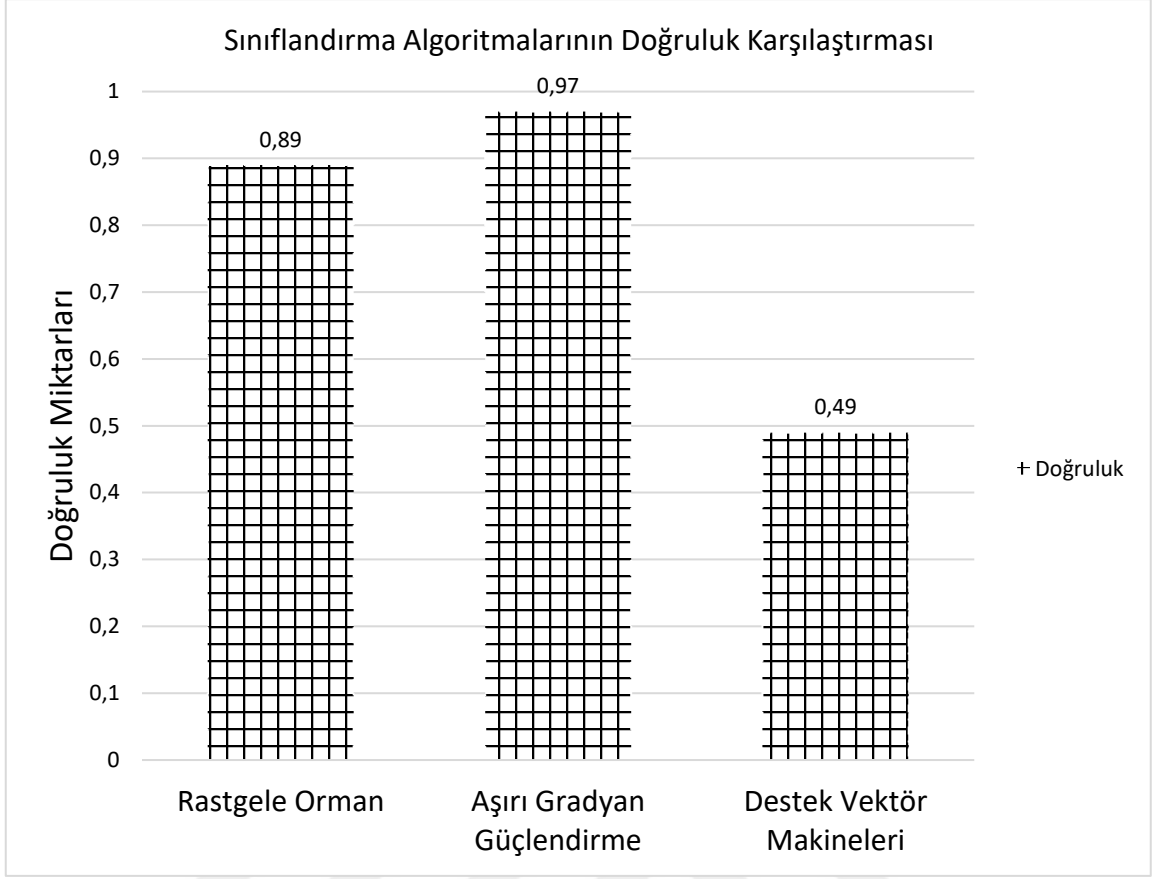


Şekil 4.23. Veri Seti I İçin Sınıflandırma Algoritmalarının RAE, RSE Karşılaştırmaları

Bu çalışmada Veri Seti I üzerinde öğrencilerin başarısına etki eden faktörlerin tespitinde, makine öğrenmesi sınıflandırma yaklaşımları ile faktörlerin önemlerinin tespit edilebileceği önerilmiştir. Amaç doğrultusunda Rastgele Orman Sınıflandırması, Aşırı Gradyan Güçlendirmesi ve Destek Vektör Makineleri denenmiş olup, modeller oluştururken CRISP-DM metodoloji adımları ile uygulama geliştirilmiştir. Elde edilen araştırma sonuçları analiz edildiğinde Aşırı Gradyan Güçlendirmesi 0,97 başarı oranı, Rastgele Orman 0,93 başarı oranı ve Destek Vektör Makineleri'nin 0,71 başarı oranına sahip olduğu tespit edilmiştir. Aşırı Gradyan Güçlendirmesi'nin daha 0,97 başarı oranı ile en yüksek başarıya sahip olduğu tespit edilmiştir.

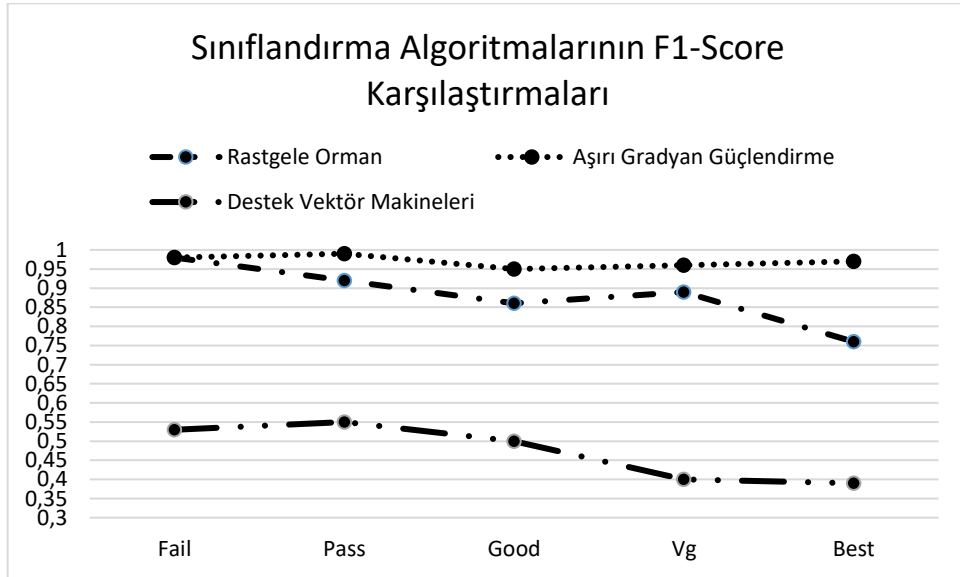
4.2.5.4.2. Veri Seti II Değerlendirilmesi

Değerlendirme adımı, bu çalışmanın ikinci araştırma sorusu olan “Başarıyı etkileyen faktörlerin tespitinde kullanılan modellerin doğrulukları nelerdir?” sorusunun cevabı Veri Seti II için Şekil 4.24’te gösterilmiştir.



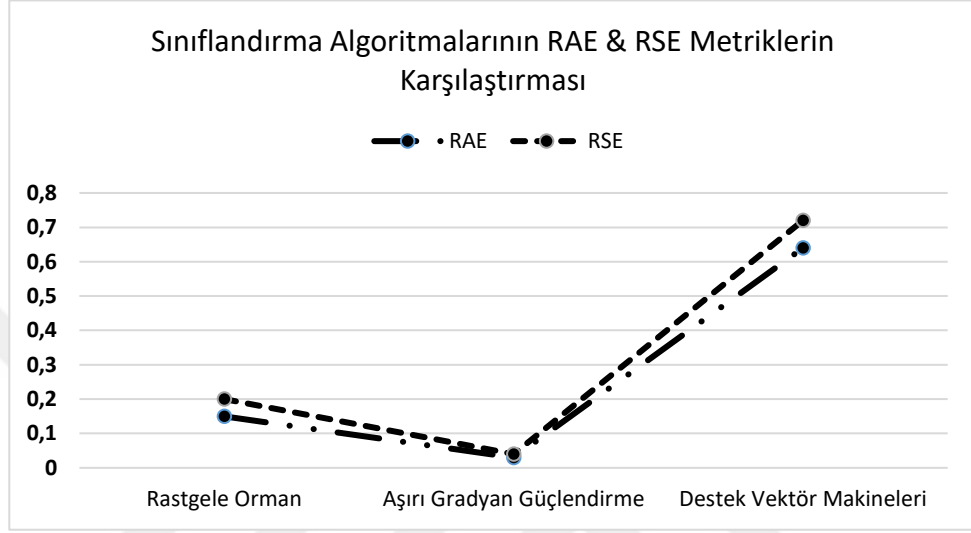
Şekil 4.24. Veri Seti II İçin Sınıflandırma Algoritmalarının Karşılaştırılması

Üç sınıflandırma modelinin doğruluğu karşılaştırıldığında, Aşırı Gradyan Güçlendirme % 97 doğruluk ile diğer sınıflandırma yöntemlerinden daha başarılı olarak sınıflandırma yapmıştır. Veri Seti II için Karmaşıklık matrisleri incelendiği zaman F1 Score kriterlerinde en yüksek değeri Aşırı Gradyan Güçlendirme vermiştir. Şekil 4.25 de F1 Score karşılaştırma grafiği gösterilmiştir.

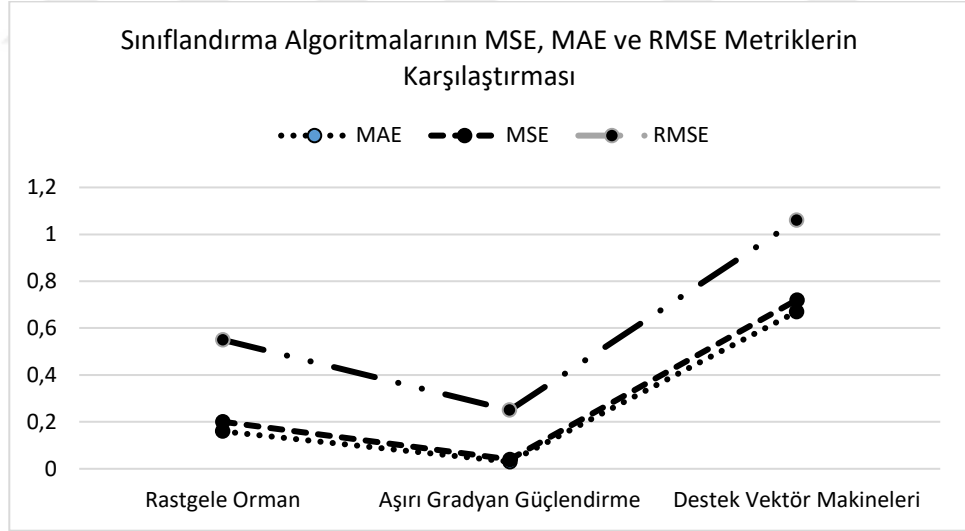


Şekil 4.25. Veri Seti II İçin Sınıflandırma Algoritmalarının F1-Score Karşılaştırmaları

Üçüncü soru olan “Hangi model ne kadar başarılı olmuştur?” sorusunun Veri Seti I için cevabı Şekil 4.26 ve 4.27’de gösterilmiştir. Aşırı Gradyan Güçlendirme sınıflandırma modelinin Veri Seti II için, diğerleriyle karşılaştırıldığında, Ortalama Mutlak Hata (MAE), Ortalama Kare Hatası (MSE), Kök Ortalama Kare Hatası (RMSE), Bağlı Mutlak Hata (RAE) ve Bağlı Standart Hata (RSE) açılarından minimum hatalara sahip olduğu görülmektedir.



Şekil 4.26. Veri Seti II İçin Sınıflandırma Algoritmalarının MSE, MAE ve RMSE Karşılaştırmaları



Şekil 4.27. Veri Seti II İçin Sınıflandırma Algoritmalarının RAE, RSE Karşılaştırmaları

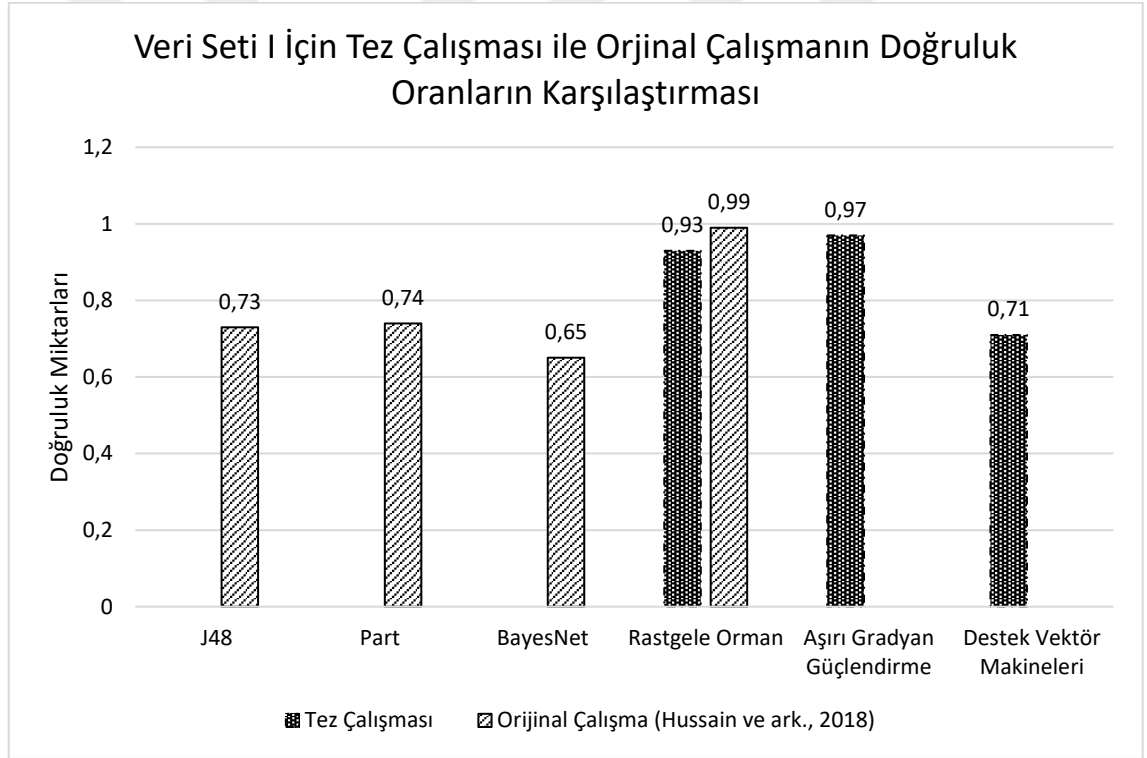
Bu çalışmada Veri Seti II üzerinde öğrencilerin başarısına etki eden faktörlerin tespitinde, makine öğrenmesi sınıflandırma yaklaşımları ile faktörlerin önemlerinin tespit edilebileceği önerilmiştir. Amaç doğrultusunda Rastgele Orman Sınıflandırması, Aşırı Gradyan Güçlendirmesi ve Destek Vektör Makineleri denenmiş olup, modeller oluştururken CRISP-DM metodoloji adımları ile uygulama geliştirilmiştir. Aşırı Gradyan Güçlendirmesi 0.97 başarı oranı, Rastgele Orman 0.89 başarı oranı ve Destek Vektör

Makineleri'nin 0,49 başarı oranına sahip olduğu tespit edilmiştir. Aşırı Gradyan Güçlendirmesi'nin daha 0,97 başarı oranı ile en yüksek başarıya sahip olduğu tespit edilmiştir.

Üçüncü soru olan “Hangi model ne kadar başarılı olmuştur?” sorusunun cevabı bu başlık altında iki veri setinde de yapılan karşılaştırmalar sonucu en iyi sonucu Aşırı Gradyan Güçlendirmesi olarak tespit edilmiştir.

4.2.5.4.3. Elde Edilen Sonuçların Literatürdeki Sonuçlar İle Değerlendirilmesi

Dördüncü soru olan “Benzer çalışmalarda elde edilen sonuçlar ile bu çalışmadaki gözlenen sonuçlar ne düzeyde benzerlik göstermektedir?” sorusunun cevabı bu başlık altında cevaplanmıştır. Şekil 4.28’de Veri Seti I için tez çalışması ile veri setinin alındığı orijinal çalışmanın sonuçları karşılaştırmıştır.



Şekil 4.28. Veri Seti I İçin Tez Çalışması ile Orijinal Çalışmanın Doğruluk Oranların Karşılaştırması

Şekil 4.28’a göre Orijinal çalışmadaki Rastgele Orman doğruluk oranı 0,99 iken tez çalışmasında doğruluk oranı 0,93 olarak tespit edilmiştir.

Veri Seti I için orijinal makalede oluşturulan modeller sonucu elde edilen algoritmalar paylaşılmamıştır. Fakat öznelite seçme adımında korelasyon analizine göre hesaplanan önemli 15 faktörü paylaşmışlardır; ARR, IAP, TNP, AS, TWP, SH, ME, FS, NF, ATD, FO, FMI, FQ, TT, SS (Hussain ve ark., 2018). Bu çalışmada verinin

hazırlanması aşamasında korelasyon analizine göre en önemli 15 faktör TWP, IAP, TNP, ARR, AS, FQ, ME, ATD, MQ, NF, FMI, SH, LS, TT, SS olarak elde edilmiştir.

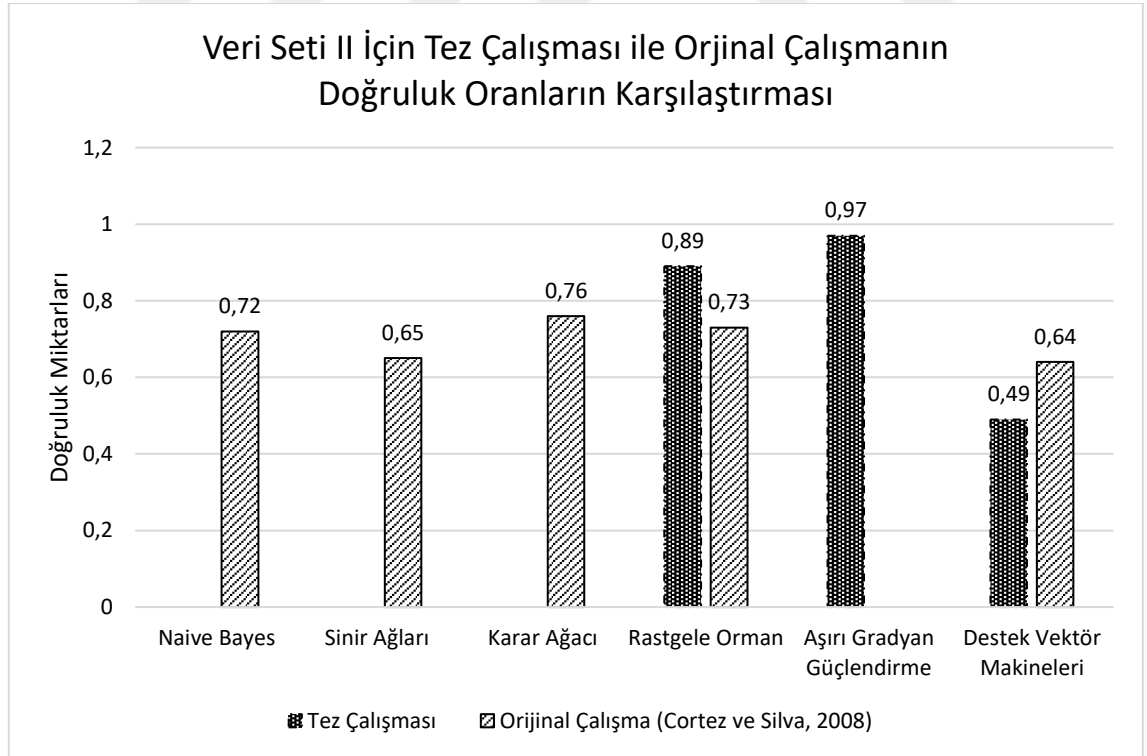
Çizelge 4.86’da Veri Seti I için orijinal çalışma ile bu tez çalışmasındaki, korelasyon analizi ile elde edilen en önemli 15 özellik verilmiştir.

Çizelge 4.86. Veri Seti I İçin Orijinal Çalışma ile Bu Çalışmadaki Faktörlerin Korelasyon Önem Sıralarının Karşılaştırılması

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Tez Çalışması	TWP	IAP	TNP	ARR	AS	FQ	ME	ATD	MQ	NF	FMI	SH	LS	TT	SS
(Hussain ve ark., 2018)	ARR	IAP	TNP	AS	TWP	SH	ME	FS	NF	ATD	FO	FMI	FQ	TT	SS

Çizelge 4.86’ya göre bu tez çalışmasında korelasyon analizi sırasında anne eğitim seviyesi, şehir veya köyde yaşaması en önemli 15 faktör içerisinde yer alırken orijinal çalışmada yer almamıştır. Orijinal çalışmada korelasyon analizi sırasında aile genişliği ve baba mesleği önemli faktörler içerisinde yer alırken bu çalışmada yer almadığı görülmüştür.

Şekil 4.29’da Veri Seti II için tez çalışması ile veri setinin alındığı orijinal çalışmanın sonuçları karşılaştırmıştır.



Şekil 4.29. Veri Seti II İçin Tez Çalışması ile Orijinal Çalışmanın Doğruluk Oranların Karşılaştırması

Şekil 4.29'a göre Orijinal makaledeki tek ortak algoritmamız Rastgele Orman'dır. Orijinal makaledeki doğruluk oranı 0,73 iken bu çalışma sonucunda 0,89 olarak bulunmuştur.

Veri Seti II için orijinal makalede Rastgele Orman için en önemli 5 faktör verilmiştir. Önceki dönemlerden kalan ders sayısı, yüksek eğitim almak istemesi, okul, okul desteği, hafta sonu alkol tüketimidir (Cortez ve Silva, 2008). Bu çalışmada Rastgele Orman algoritmasına göre bulunan en önemli 5 faktör; sağlık durumu, anne eğitim seviyesi, devamsızlık sayısı, önceki dönemlerden kalan ders sayısı, alkol tüketimidir. Hafta sonu alkol tüketimi ve hafta içi alkol tüketimi verinin hazırlanması aşamasında birleştirilmiştir. İki çalışmadaki Rastgele Orman yöntemi ile elde edilen modeller karşılaştırıldığında; önceki dönemlerden kalan ders sayısı ve hafta sonu alkol tüketimi faktörlerinin sadece iki çalışmada yer aldığı tespit edilmiştir. Aşırı Gradyan Güçlendirme algoritmasının en önemli 5 faktörü ise; anne eğitim seviyesi, okul desteği, yüksek eğitim almak istemesi, okul, önceki dönemlerden kalan ders sayısıdır.

Veri Seti II için orijinal makalede Rastgele Orman algoritmasına göre en önemli 5 faktör ile bu çalışmadaki en iyi algoritma seçilen Aşırı Gradyan Güçlendirme algoritmasının en önemli 5 faktörünün kıyaslanması Çizelge 4.87 de gösterilmiştir.

Çizelge 4.87. Veri Seti II İçin Orijinal Makaledeki En İyi Sonuç Veren Algoritmaların En Önemli Faktörlerinin Karşılaştırılması.

	1	2	3	4	5
Tez Çalışması	FAILURES	HIGHER	SCHOOL	SCHOOLSUP	WALC
(Cortez ve Silva, 2008)	MJOB	SCHOOLSUP	HIGHER	SCHOOL	FAILURES

Çizelge 4.87'ye göre Okul desteği, okul, önceki dönemlerden kalan ders sayısı, yüksek eğitim almak istemesi iki çalışmadaki en iyi algoritmalara göre en önemli beş faktör içerisinde yer aldığı görülmüştür.

Çizelge 4.88'de Literatürdeki diğer çalışmalar ile bu tez çalışmasının karşılaştırılması verilmiştir.

Çizelge 4.88. Veri Seti II'yi Kullanan Diğer Çalışmalar İle Tez Çalışmasının Karşılaştırılması

	Yapay Sinir Ağı	Karar Ağacı	Naive Bayes	Çok Algılayıcı Katmanlar	Lojistik Regresyon	Sıralı Minimal Optimizasyon	Rastgele Orman	Gradyan Güçlendirme	Destek Vektör Makineleri	Aşırı Gradyan Güçlendirme
Tez Çalışması							0,89		0,49	0,97
(Senthil ve Lin, 2017)			0,91	0,94	0,93	0,92	0,95			
(Ragab ve ark., 2021)	0,83	0,9	0,84							
(Pal ve Chaurasia, 2014)		0,79				0,73				
(Ahmed ve Sadiq, 2018)							0,83			
(Ali ve ark., 2021)					0,92			0,93		
(Ng ve ark., 2021)			0,54	0,7					0,7	
(Özkan ve ark., 2018)		0,86			0,82		0,84		0,78	
(Güldal, 2018)		0,66								
(Agrawal ve ark., 2017)		0,88	0,82				0,84			
(Aghalarova ve Keser, 2021)							0,87			

Veri Seti II için Agrawal ve ark. (2017) yapmış oldukları çalışmada, G3 özelliğini dört sınıfa indirerek 100 örnek üzerinden bir çalışma gerçekleştirmişlerdir. Yaptıkları çalışmada en iyi algoritma sonucunu Karar Ağacı 0,88 doğruluk ile elde etmişlerdir. Rastgele Orman algoritmasında 0,84 olarak bulmuşlardır. Veri Seti II için Pal ve Chaurasia (2014) yaptığı çalışmada, alkol tüketen öğrencilerin performansını en çok etkileyen en önemli özelliklerin öğrencilerin dönemlere ait notları olduğunu tespit etmiştir. Diğer faktörler ise devamsızlık sayısı, anne ve babanın mesleği, okul ve ailenin eğitim desteğini, ekstra ücretli ders, yüksek eğitim almak istemesi, okulu seçme sebebi olarak tespit etmiştir. Veri Seti II için Özkan ve ark. (2018) yapmış oldukları çalışmada, G3 özelliğini iki sınıfa indirgemişlerdir. Yaptıkları çalışmada en iyi algoritma sonucunu C5.0 Karar Ağacı kullanarak 0,86 doğruluk ile elde etmişlerdir. Rastgele Orman algoritmasında 0,84; Destek Vektör Makinalarında ise 0,78 olarak bulmuşlardır. Veri Seti II için Aghalarova ve Keser (2021) yapmış oldukları çalışmada, Rastgele Orman üzerinde 0,87 doğruluk bulmuşlardır. Veri Seti II için Senthil ve Lin (2017) yapmış oldukları çalışmada, G3 özelliğini iki sınıfa indirgemişlerdir. Yaptıkları çalışmada en iyi algoritma sonucunu Basit Lojistik Regresyon 0,9322 doğruluk ile elde etmişlerdir. Rastgele Orman

algoritmasında 0,9306 olarak bulmuşlardır. Bu çalışma ile kıyaslandığında yaş, baba mesleği, anne eğitim seviyesi, okulu seçme sebebi, alkol tüketimi, devamsızlık sayısı önemi ile benzerlik göstermektedir.

İlk soru olan “Öğrencilerin akademik başarısını etkileyen önemli faktörler nelerdir?” sorusunun cevabına Veri Seti I ve Veri Seti II için önemli faktörleri en iyi algoritma sonuçları ile bir araya getirildiğinde;

- Önceki Yıllardaki Başarı Ortalamaları,
- Önceki Dönemden Dersi,
- Sınav Dili,
- Anne Mesleği ve Anne Eğitim Seviyesi,
- Aile Genişliği,
- Aile Aylık Geliri ve Ailenin Eğitim Desteği,
- Bir Önceki Okulu ve Tipi
- Okuduğu Okul ve Okulun Desteği,
- Etkinliklere Katılma,
- Evde İnternet Erişimi,
- Alkol Tüketimi,
- Yüksek Eğitim Almak İstemesi,

olarak tespit edilmiştir.

Çırak (2012)’ın başka bir veri seti kullanarak yaptıkları çalışma ile bu tez çalışması kıyaslandığında, bir önceki okulu ve tipi, önceki yıllardaki başarı ortalamaları önemi ile tespitleri benzerlik göstermektedir. Tavşancıl (1989)’ın yapmış olduğu nitel araştırma ile bu tez çalışması kıyaslandığında, önceki yıllardaki başarı ortalamaları önemi ile tespitleri benzerlik göstermektedir. Khan ve ark. (2015) başka bir veri seti kullanarak yaptıkları çalışma ile bu tez çalışması kıyaslandığında, anne ve babanın mesleğinin önemi ile benzerlik göstermektedir. Yıldız ve Börekçi (2020)’nin başka bir veri seti kullanarak yaptıkları çalışma ile bu tez çalışması kıyaslandığında, anne eğitim seviyesi ve aile aylık geliri ile benzerlik göstermektedir. Civan ve Coşkun (2016)’un yapmış olduğu nitel araştırma ile bu tez çalışması kıyaslandığında anadili olmayan dilde eğitim alan ve sınava giren öğrencilerin akademik başarısını olumsuz etkilemesi ile benzerlik göstermektedir. Yolcu (2020)’nun başka bir veri seti kullanarak yaptığı çalışma ile bu tez çalışması kıyaslandığında, evde internet erişimi önemi ile benzerlik göstermektedir.

4.1.6. Konuşlandırma

Konuşlandırma adımında değerlendirme adımında en iyi sonucu veren algoritma modeli kaydedildi. Aynı veri setine yeni veriler eklendikçe aynı model çağrılıp tekrar denenebilir. Değerlendirme adımlarında “Medeni Hal Durumu” özelliğinin önemli olmadığı ortaya çıkmıştır. Düşük öneme sahip faktörler çıkarılıp model tekrar çalıştırıldı ve çalıştırılmış ve yakın sonuçlar elde edilmiştir.



SONUÇLAR VE ÖNERİLER

Bu tez çalışması ile makine öğrenmesi algoritmaları kullanarak öğrencilerin akademik başarısını etkileyen faktörlerin tespit edilmesinde Rastgele Orman, Aşırı Gradyan Güçlendirmesi ve Destek Vektör Makinaları yöntemlerinin karşılaştırılması ve önemli faktörlerin tespitinin belirlenmesi amaçlanmıştır. Çalışmada ilk veri seti olarak Dahan, Hussain, Ba-Alwib ve Ribata'nın yapmış oldukları Hindistan'ın Assam eyaletinde gerçekleştirilen bir araştırmadan elde edilen veriler kullanılmıştır. İkinci veri seti olarak Cortez ve Silva'nın yapmış oldukları Portekiz'in Alentejo bölgesinden 2005 yılında iki devlet okulundan anket ile toplamış oldukları veriler kullanılmıştır. Yapılan çalışma ile öğrencilerle yapılacak şeffaf bir anket sonucu elde edilen bilgiler ile onların başarısına nelerin etkili olabileceği görülmüştür. Rastgele Orman, Aşırı Gradyan Güçlendirmesi ve Destek Vektör Makinaları ile sınıflandırma modelleri oluşturarak öğrencilerin akademik başarısına etki eden önemli faktörler incelenmiştir. Rastgele Orman sınıflandırmasına göre Veri Seti I için başarı oranı %93, Veri Seti II için başarı oranı %89 olarak bulunmuştur. Aşırı Gradyan Güçlendirmesi sınıflandırmasına göre Veri Seti I için başarı oranı %97, Veri Seti II için başarı oranı %97 olarak bulunmuştur. Destek Vektör Makinaları sınıflandırmasına göre Veri Seti I için başarı oranı %71, Veri Seti II için başarı oranı %49 olarak bulunmuştur. Bu tez çalışmasında makine öğrenmesi algoritmaları kullanarak öğrencilerin akademik başarısını etkileyen faktörlerin tespit edilmesinde Aşırı Gradyan Güçlendirmesi sınıflandırmasının Rastgele Orman ve Destek Vektör Makinalarına göre daha başarılı sonuçlar verdiği tespit edilmiştir. Yapılan çalışma çoklu sınıflandırma içerdiğinden dolayı Aşırı Gradyan Güçlendirmesi ve Rastgele Orman üzerinde rahatça uygulanabildiği, Destek Vektör Makineleri'nde başarısız olduğu ortaya çıkmıştır. Çok sınıflı Destek Vektör Makineleri'nde optimizasyon problemi olduğundan dolayı, Destek Vektör Makineleri uygulanacağına sınıf sayısının düşürülmesi gerekir (Franc ve Hlaváč, 2002). Çoklu hedef sınıf içeren veri setlerinde boyut indirgemesi yapılarak iki sınıfa indirgenerek kullanılabilir. Ancak verilerin normal sınıf dağılıma sahip olmadığı veri setlerinde, tahmin sınıfları arasında dengesiz dağılıma yol açabilir. Bundan dolayı veri setinde hedef sınıf dağılımları incelendikten sonra sınıf indirgeme ve model seçimi yapılmalıdır.

Algoritmalar kıyaslanırken K-Kat Çapraz doğrulama sonuçları kullanılmamıştır. Çoklu sınıflandırma çalışması yapıldığından dolayı sınıflar arasında dengesizlikler mevcuttur. “Ne yazık ki, K-Kat Çapraz Doğrulama, dengesiz sınıflandırıcıları

değerlendirmek için uygun değildir. Özellikle makine öğreniminde en yaygın olarak kullanılan hata tahmin yöntemi olan 10 Katlı Çapraz Doğrulama, çarpıklık önceden düşünülenenden daha az aşırı olsa bile sınıf dengesizlikleri durumunda kolayca bozulabilir.” (He ve Ma, 2013). F1 Skorlar üzerinden yorumlamalar yapılmıştır. K-Kat Çapraz doğrulama sadece algoritmalar için hiperparametre seçimi sırasında kullanılmıştır. Veri setinde sınıf indirgeme ile sınıflar arasında dengesiz veri seti ortaya çıkar ise, Random Under-Sampling, Random Over-Sampling, Smote, Adasyn ve Nearmiss yöntemlerinden biri kullanılabilir. Bu araştırma çalışmasında öğrencilerin akademik başarısına etki eden faktörler Çizelge 4.25 ve Çizelge 4.60 yer almaktadır. Yapılacak diğer çalışmalarda bu çizelgelerde yer alan faktörler dışında başka faktörler dahil edilerek yeni bir anket çalışması yaparak kullanılabilir. Veri setinde daha az değişken kullanarak yüksek başarıya sahip modeller elde edilebilir. Bu alanda yapılacak yeni çalışmalarda yüksek korelasyona sahip olan özellikler yerine orta düzeyde korelasyona sahip değişkenler seçilerek, aynı bölümde, aynı dersleri alan, farkı üniversitelerden öğrenciler üzerinde anket çalışması yapılabilir. Bu tez çalışmasında elde edilen sonuçlara göre, öğrencilerin daha önceki değerlendirmelerin önemli etkili olduğu ortaya çıkmıştır. Ancak sosyal faktörlerin tespiti için daha kapsamlı bir çalışma yapılarak, bunlar içerisinde eleme yapılarak yapılan bir anket çalışması ile yapılmalıdır. Ancak buradaki en önemli nokta yapılan anket çalışmasının ne kadar doğru olduğudur. Öğrenciler sorulan kişisel sorulara cevap vermeyebilir. Birden çok sınıflandırma algoritması ile modeller geliştirilirken tüm modellere aynı veri setini uygulamak önemlidir. Bu sayede şeffaf bir karşılaştırma yapılabilir. Geleneksel yöntemler ile başarıya etki eden faktörlerin tespiti zordur. Çünkü teknolojik gelişmeler arttıkça, yeni keşifler ortaya çıktıkça bunların etki edip etmeyeceği bir anda tespit edilemez. Bu çalışmada CRISP-DM metodolojisi ile uçtan uca bir makine öğrenmesi çalışması gerçekleştirilmiştir.

KAYNAKÇA

- Aghalarova, S. ve Keser, S. B., 2021, Önerilen Yapay Sinir Ağı Algoritması ile Ortaokul Öğrencilerin Akademik Performansının Tahmini, *Veri Bilim Dergisi*, 4(2), 19-32.
- Agrawal, S., Vishwakarma, S. K. ve Sharma, A. K., 2017, Using Data Mining Classifier for Predicting Student's Performance in UG Level, *International Journal of Computer Application*, 172(8).
- Ahmed, N. S. ve Sadiq, M. H., 2018, Clarify Of The Random Forest Algorithm in An Educational Field. 2018 International Conference On Advanced Science And Engineering (ICOASE), IEEE: 179-184.
- Ali, M., Khattak, A. M., Ali, Z., Hayat, B., Idrees, M., Pervez, Z., Rizwan, K., Sung, T.-E. ve Kim, K. I., 2021, Estimation And Interpretation Of Machine Learning Models With Customized Surrogate Model, *Electronics*, 10 (23), 3045.
- Amit, Y. ve Geman, D., 1997, Shape Quantization And Recognition With Randomized Trees, *Neural Computation*, 9 (7), 1545-1588.
- Anonymous, 2021, Tuning The Hyper-Parameters Of An Estimator [online], https://scikit-learn.org/stable/modules/grid_search.html: [Ziyaret Tarihi: 19 Ocak 2022].
- Asyraf, A. S., Abdul-Rahman, S. ve Mutalib, S., 2017, Mining Textual Terms For Stock Market Prediction Analysis Using Financial News, *Communications in Computer and Information Science*, 788, 293–305.
- Aydın, Z. E., 2021, Balina optimizasyonu algoritması ve rastgele alt uzaylar temelli eksik veri tamamlama yöntemleri.
- Aytekin, H. T., 2021, Makine Öğreniminin Araştırmacıların Veri Analizi Bağlamında Potansiyel Önemi, 10 (19), 85-106.
- Barrak, A., Mashael, A. ve Muna, A.-R., 2016, Predicting Students Final GPA Using Decision Trees, 6.7 (528).
- Breiman, L., 2001, Random Forests, 45 (1), 5-32.
- Brownlee, J., 2016, XGBoost with Python, Machine Learning Mastery.
- Calkins, K. G., 2005, Correlation Coefficients, <https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm>: [Ziyaret Tarihi: 19 Ocak 2022].
- Chambers, E. A. ve Schreiber, J. B., 2004, Girls' Academic Achievement: Varying Associations Of Extracurricular Activities, 16(3), 327-346.
- Cheewaparakobkit, P., 2013, Predicting Student Academic Achievement by Using The Decision Tree and Neural Network Techniques, *International Multi Conference of Engineers and Computer Scientists*, 12 (34-43).
- Christian, T. M. ve Ayub, M., 2014, Exploration of Classification Using NBTree for Predicting Students' Performance.
- Civan, A. ve Coşkun, A., 2016, The Effect of the Medium of Instruction Language on the Academic Success of University Students, *Educational Sciences: Theory and Practice*, 16(6), 1981-2004.
- Cortez, P. ve Silva, A., 2008, Using Data Mining To Predict Secondary School Student Performance.
- Çemrek, F. ve Filiz, Z., 2007, Üniversite Öğrencilerinin Barınma Uygunluk Analizi İle İncelenmesi, 8 (2).
- Çırak, G., 2012, Yükseköğretimde Öğrenci Başarılarının Sınıflandırılmasında Yapay Sinir Ağları Ve Lojistik Regresyon Yöntemlerinin Kullanılması, Yüksek Lisans, *Ankara Üniversitesi Eğitim Bilimleri Enstitüsü*.

- Dam, H., 2008, Öğrencinin Okul Başarısında Aile Faktörü, *Hitit Üniversitesi İlahiyat Fakültesi Dergisi*, 2 (14).
- El-Refae, G. A. ve Al-Shayea, Q. K., 2010, Predicting Students' Academic Performance Using Artificial Neural Networks: A Case Study., *International Journal of Computer Science and Information Security*, 5, 97-100.
- Farid, D. M., Maruf, G. M. ve Rahman, C. M., 2013, A New Approach Of Boosting Using Decision Tree Classifier For Classifying Noisy Data, *International Conference on Informatics, Electronics and Vision (ICIEV)*, 1-4.
- Franc, V. ve Hlaváč, V., 2002, Multi-Class Support Vector Machine, 236-239.
- Gislason, P. O., Benediktsson, J. A. ve Sveinsson, J., 2006, Random Forests For Land Cover Classification, 27 (4), 294-300.
- Guo, B., Rui, Z., Li, Y., Guang, X. ve Chuangming, S., 2015, Predicting Students Performance in Educational Data Mining, *International Symposium on Educational Technology*, 125-128.
- Güldal, H., 2018, Karar Ağacı Algoritmalarının Eğitsel Veriler Üzerindeki Performanslarının İncelenmesi, *International Balkan Education and Science Congress*.
- Güneş, F., 2020, Eğitimde Ezber Tartışmaları, *Eğitim Kuram ve Uygulama Araştırmaları Dergisi*, 6 (3).
- Harvey, J. L. ve Kumar, S. A., 2019, A Practical Model For Educators To Predict Student Performance in K-12 Education Using Machine Learning, *IEEE Symposium Series on Computational Intelligence (SSCI)*, 3004-3011.
- He, H. ve Ma, Y., 2013, Imbalanced Learning: Foundations, Algorithms, And Applications.
- Ho, T. K., 1998, The Random Subspace Method For Constructing Decision Forests, 20(8), 832-844.
- Hussain, S., Dahan, N. A., Ba-alwib, F. M. ve Ribata, N., 2018, Educational Data Mining and Analysis of Students' Academic Performance, 9 (447-459).
- Ibrahim, Z. ve Rusli, D., 2007, Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression, *21st Annual SAS Malaysia Forum*, Kuala Lumpur.
- Kaur, P., Singh, M. ve Josan, G. S., 2015, Classification And Prediction Based Data Mining Algorithms To Predict Slow Learners In Education Sector, 57 (500-508).
- Kavak, Y., 2011, Türkiye'de Yükseköğretimin Görünümü ve Geleceğe Bakış, *Yükseköğretim ve Bilim Dergisi*, 1(2).
- Khan, B., Khattak, M. D. ve Khiyal, M. S. H., 2015, Final Grade Prediction Ofsecondary School Student Using Decision Tree, 115 (21).
- Koç, N., 1981, Liselerde Öğrencilerin Akademik Başarılarının Değerlendirilmesi Uygulamalarının Etkinliğine İlişkin Bir Araştırma, *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Yayınları*, 104.
- Kulkarni, S. R. ve Harman, G., 2011, Statistical Learning Theory : A Tutorial, *Wiley Interdisciplinary Reviews*, 3, 543-556.
- Kumar, S. A. ve Vijayalakshmi, M. N., 2011, Efficiency Of Decision Trees İn Predicting Student's Academic Performance. *International Conference on Computer Science, Engineering and Applications (CCSEA)*: 335-343.
- Márquez-Vera, C., Ventura, S., Cano, A. ve Romero, C., 2013, Predicting Student Failure At School Using Genetic Programming And Different Data Mining Approaches With High Dimensional And Imbalanced Data, *Applied Intelligence*, 1, 1-16.

- Marzano, R. J., 2003, What Works in Schools: Translating Research Into Action, ASCD.
- Marzano, R. J., Pickering, D. J. ve Pollock, J. E., 2008, Öğrenci Başarısını Artıran Öğretim Stratejileri, *Sev Yayınevi*.
- Mesarić, J. ve Šebalj, D., 2016, Decision Trees Forpredicting The Academic Success Of Students, *Croatian Operational Research Review*, 7.2, 367-388.
- Mitchell, T. ve McGraw, H., 1997, Machine Learning. New York, McGraw-Hill Higher Education.
- Morde, V., 2019, XGBoost Algorithm: Long May She Reign, <https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d>: [Ziyaret Tarihi: 19 Ocak 2022].
- Ng, H., bin Mohd Azha, A. A., Yap, T. T. V. ve Goh, V. T., 2021, A Machine Learning Approach to Predictive Modelling of Student Performance, *F1000Research*, 10, 1144.
- Nurafifah, M., Suhaimi Shuzlina, A.-R., Sofianita, M., Nurzeatul Hamimah, A. H. ve Ariff Md, A. M., 2019, Review on Predicting Students' Graduation Machine Learning Algorithms, 7 (1-13).
- Ogwoka, T. M., Cheruiyot, W. ve Okeyo, G., 2015, A Model For Predicting Students 'Academic Performance Using A Hybrid K-Means And Decision Tree Algorithms, *International Journal of Computer Applications Technology and Research*, 4 (9), 693-697.
- Ong, E. J. ve Bowden, R., 2004, A Boosted Classifier Tree For Hand Shape Detection, *Sixth IEEE Int. Conf. Autom. Face Gesture Recognit*, 889–894.
- Oshiro, T. M., Perez, P. S. ve Baranauskas, J. A., 2012, How Many Trees in a Random Forest?, *Including Subseries Lecture Notesin Artificial Intelligence and Lecture Notes in Bioinformatics*, 7376, 154-168.
- Osmanbegović, E. ve Suljić, M., 2012, Data Miningapproach for Predicting Student Performance, *Economic Review*, 10.1 (3-12).
- Özkan, Y., Önay, K. F. ve Selçukcan, E. Ç., 2018, Öğrenci Performansının Yapay Zeka Derin Öğrenme Algoritması İle Öngörülmesi, *7th International Conference on "Innovations in Learning for the Future": Digital Transformation in*, İstanbul, 136-141.
- Özoğlu, S. Ç. ve Koç, N., 1995, Çağdaş Üniversitede Öğrencinin Başarısının Ölçülmesi ve Değerlendirilmesi, *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Yayınları*.
- Pal, M., 2005, Random Forest Classifier For Remote Sensing Classification., *International Journal Of Remote*, 61, 217-222.
- Pal, S. ve Chaurasia, V., 2014, Performance Analysis of Students Consuming Alcohol Using Data Mining Techniques, *International Journal of Advance Research in Science and Engineering*, 6 (2).
- Peker, R., 2003, Öğrencilerinin Genel Akademik Başarılarının Bazı Değişkenlere Göre İncelenmesi, *Uludağ Üniversitesi Eğitim Fakültesi Dergisi*, 17 (16).
- Qi, Y., 2012, Random Forest For Bioinformatics, *Ensemble Machine Learning*, 307-323.
- Ragab, M., Abdel Aal, A. M., Jifri, A. O. ve Omran, N. F., 2021, Enhancement of Predicting Students Performance Model Using Ensemble Approaches and Educational Data Mining Techniques, *Wireless Communications and Mobile Computing*, 2021.
- Ramesh, V. ve Ramar, K., 2013, Predicting Student Performance A Statistical and Datamining Approach, *International Journal of Computer Applications*, 63 (8).

- Ruby, J. ve David, K., 2014, Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 2 (9), 173-180.
- Samuel, A. L., 1959, Some Studies In Machine Learning Using The Game Of Checkers, *IBM Journal Of Research And Development*, 3, 210-229.
- Santra, A. ve Christy, C., 2012, Genetic Algorithm And Confusion Matrix For Document Clustering, *International Journal of Computer Science Issues*, 9(1).
- Senthil, D. S. ve Lin, W. M., 2017, Applying Classification Techniques To Predict Students' Academic Results, *IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, Bengaluru.
- Srinivasan, A. V., 2019, Why Exclude Highly Correlated Features When Building Regression Model ?? [online], <https://towardsdatascience.com/why-exclude-highly-correlated-features-when-building-regression-model-34d77a90ea8e>: [Ziyaret Tarihi: 19 Ocak 2022].
- Stanley, A. J. ve Glock, M. D., 1967, Evaluating Pupil Growth, *Boston*, Allyn and Bacon, Inc, p.
- Tavşancıl, E., 1989, Lise Tür Ve Kolunun Yükseköğretimdeki Akademik Başarıya Etkisi, Doktora, *Ankara Üniversitesi*.
- Thomas, R. M., 1998, Conducting Educational Research: A Comparative View, Greenwood Publishing Group.
- Turgut, M. F. ve Baykul, Y., 2012, Eğitimde Ölçme Ve Değerlendirme.
- Vijayalakshi, K. ve Venkatachalapathy, V., 2019, Comparison of Predicting Student's Performance using Machine Learning Algorithms, 12 (34-45).
- Wook, M., Yahaya, Y. H., Wahab, N., Isa, M. R. M., Awang, N. F. ve Seong, H. Y., 2009, Predicting Ndundum Student's Academic Performance Using Data Mining Techniques, *International Conference on Computer and Electrical Engineering*, 357-361.
- Yıldız, M. B. ve Börekçi, C., 2020, Predicting Academic Achievement with Machine Learning Algorithms, *Journal of Educational Technology & Online Learning*, 3(3), 372-392.
- Yolcu, H. H., 2020, Koronavirüs (Covid-19) Pandemi Sürecinde Sınıf Öğretmeni Adaylarının Uzaktan Eğitim Deneyimleri, 6 (4).
- Yök, 2020, Yüksek Öğrenim Bilgi Sistemi 2020 Yılı Ulusal Verileri, Yükseköğretim Bilgi Yönetim Sistemi.