
GAZİ ÜNİVERSİTESİ
Gazi
1926

AÇIKÖĞRETİM ÖĞRENCİLERİNİN OKUL TERKİNE ETKİ EDEN FAKTÖRLERİN VERİ MADENCİLİĞİ İLE KEŞFİ

Selma Tosun

DOKTORA TEZİ

EĞİTİM BİLİMLERİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

KASIM, 2024

TELİF HAKKI VE TEZ FOTOKOPİ İZİN FORMU

Bu tezin tüm hakları saklıdır. Kaynak göstermek koşuluyla tezin teslim tarihinden itibaren 12 (on iki) ay sonra tezden fotokopi çekilebilir.

YAZARIN

Adı : Selma

Soyadı : Tosun

Bölümü : Eğitimde Ölçme ve Değerlendirme

İmza :

Teslim tarihi :

TEZİN

Türkçe Adı : Açıköğretim Öğrencilerinin Okul Terkine Etki Eden Faktörlerin Veri Madenciliği ile Keşfi

İngilizce Adı: Predicting The Factors Affecting School Dropout in Open and Distance Education Students by Using Data Mining

ETİK İLKELERE UYGUNLUK BEYANI

Tez yazma sürecinde bilimsel ve etik ilkelere uyduğumu, yararlandığım tüm kaynakları kaynak gösterme ilkelerine uygun olarak kaynakçada belirttiğimi ve bu bölümler dışındaki tüm ifadelerin şahsıma ait olduğunu beyan ederim.

Yazar Adı Soyadı: Selma Tosun

İmza :

JÜRİ ONAY SAYFASI

Selma Tosun tarafından hazırlanan "Açıköğretim Öğrencilerinin Okul Terkine Etki Eden Faktörlerin Veri Madenciliği ile Keşfi" adlı tez çalışması aşağıdaki jüri tarafından oy birliği ile Gazi Üniversitesi Eğitim Bilimleri Ana BilimDalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı'nda Doktora Tezi olarak kabul edilmiştir.

Danışman: Prof. Dr. Dilara Bakan Kalaycıoğlu …….
(Eğitimde Ölçme ve Değerlendirme, Gazi Üniversitesi)

Başkan: Doç. Dr. Kübra Atalay Kabasakal …….
(Eğitimde Ölçme ve Değerlendirme, Hacettepe Üniversitesi)

Üye: Prof. Dr. İsmail Karakaya …….
(Eğitimde Ölçme ve Değerlendirme, Gazi Üniversitesi)

Üye: Prof. Dr. Melek Gülşah Şahin …….
(Eğitimde Ölçme ve Değerlendirme, Gazi Üniversitesi)

Üye: Doç. Dr. Önder Sünbül …….
(Eğitimde Ölçme ve Değerlendirme, Mersin Üniversitesi)

Tez Savunma Tarihi: 18/10/2024

Bu tezin Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı'nda Doktora tezi olması için şartları yerine getirdiğini onaylıyorum.

Prof. Dr. Şaban Çetin …….
Eğitim Bilimleri Enstitüsü Müdürü

Canım aileme...

TEŞEKKÜR

Öncelikle bu araştırmanın başından sonuna kadar her aşamasına verdiği destekle beni daima canlı tutan ve motive eden, akademik bilgisi ile yoluma ışık olan, yolda düşmeme asla izin vermeyen sevgili tez danışmanım, değerli hocam Prof. Dr. Dilara Bakan Kalaycıoğlu'na ne kadar teşekkür etsem az olacaktır.

Doktora sürecimin değişik aşamalarında gelişimime özveriyle katkı sunan Prof. Dr. İsmail Karakaya'ya, Prof. Dr. Bayram Çetin'e, Prof. Dr. Nilüfer Kahraman'a, Prof. Dr. Hakan Yavuz Atar'a, Prof. Dr. Mehtap Çakan'a ve Doç. Dr. Emine Önen'e minnettarım. Tez izleme sürecimde sabırla beni dinleyen, değerli bilgilerini paylaşan ve ihtiyaç duyduğum her anda yanımda olan Prof. Dr. Melek Gülşah Şahin ve Doç. Dr. Önder Sünbül'e çok teşekkür ederim. Ayrıca eğitim sürecimde yardımlarını esirgemeyen Eğitimde Ölçme Değerlendirme Bilim dalındaki araştırma görevlilerine; tez sürecimde desteklerini hep hissettiğim İÜ AUZEF yönetimine ve mesai arkadaşlarıma özellikle teşekkür ederim.

Son olarak, desteklerini her alanda hissettiren değerli ailem; annem, babam ve kardeşlerim, size hep minnettar kalacağım. Çoğu zaman aile olarak geçireceğimiz vakitlerden eksilterek yürüttüğüm bu çalışmanın en yorgun günlerinde, en umutsuz anında bile yüzümü güldüren, enerji kaynağım sevgili oğullarım Hasan Alp ve Ozan Kaan iyi ki varsınız. Ve hayatımın her aşamasına olduğu gibi akademik çalışmalarıma da sevgiyle emek veren, güldüğüm, üzüldüğüm, çalıştığım, yorulduğum nice anların şahidi sevgili evdeşim Mehmet Engin Tosun, birlikte yürüdüğümüz bu yolda beni her daim cesaretlendirdin, iyi ki varsın.

AÇIKÖĞRETİM ÖĞRENCİLERİNİN OKUL TERKİNE ETKİ EDEN FAKTÖRLERİN VERİ MADENCİLİĞİ İLE KEŞFİ

(Doktora Tezi)

Selma Tosun

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Kasım 2024

ÖZ

Bu çalışmanın amacı, açık ve uzaktan eğitimde okuyan öğrencilerin sistemler üzerinden elde edilebilecek verileri ile okul terki ve akademik başarı durumlarını tahmin etmek ve sınıflandırma performansı yüksek veri madenciliği modellerini belirleyerek bunlara göre akademik başarı ve okul terkinde etkili olan değişkenleri ortaya çıkartmaktır. Araştırmada, yükseköğretimdeki açık ve uzaktan eğitim öğrencilerinin demografik ve kişisel verileri ile ağırlıklı genel not ortalaması puanı değişken olarak alınarak okul terki riski taşıyan öğrencileri erken tespit edebilmek için veri madenciliği algoritmaları ile bir model geliştirilmiştir. Çalışmada veri madenciliği yaklaşımının eğitim alanında kullanılması ile bu algoritmalara dayalı olarak elde edilen sonuçların güvenirlik ve geçerlik değerlerinin ne düzeyde olduğu belirlenmiştir. Bu araştırma amacı bakımından nicel araştırma yöntemlerinden olan ilişkisel araştırma modelindedir. Araştırmanın türü, çalışma grubu, veri toplama araçları ve verilerin analizine yönelik işlemler, veri madenciliği için Alanlararası Standart Prosedürler (CRISP-DM) sürecinde yer alan aşamalar ile ilişkilendirilerek ele alınmıştır. Araştırma, İstanbul Üniversitesi Açık ve Uzaktan Eğitim Fakültesi'nde (İÜ AUZEF) 2021-2022 bahar dönemine kadar en az bir defa bir programa kayıt olan tüm

öğrencilerin İstanbul Üniversitesi öğrenci bilgi sistemi veri tabanından elde edilen verileri ile yürütülmüştür. Analiz sürecinde okul terki ve akademik başarıya göre öğrencilerin sınıflandırılmasında Rastgele Ağaçlar, C&RT, C5.0, CHAID, QUEST gibi Karar Ağaçları İle Naive Bayes (NB), Lojistik Regresyon (LR), Yapay Sinir Ağları (YSA) ve Destek Vektör Makinesi (DVM) gibi çeşitli veri madenciliği algoritmaları kullanılmış ve bunların performansları doğruluk, duyarlılık, hassasiyet, kesinlik, negatif öngörü değeri, F1 puanı, MCC, AUC kriterlerine göre değerlendirilmiştir. Araştırmanın sonucunda okul terkine göre öğrencilerin sınıflandırılması için oluşturulan modellerin genel doğru sınıflandırma oranlarının 0,754 ile 0,790 arasında değiştiği görülmüştür. Tüm veri setinde genel sınıflandırma doğruluğu en yüksek model C5.0'dır. Ayrıca C5.0 modeli diğerlerine göre en yüksek duyarlılık, F1 puanı, kesinlik, negatif öngörü değeri ve MCC değerine de sahip olduğundan okul terkinin tahmininde modelin, diğerlerine göre performans değerlerinin daha iyi olduğu sonucuna varılmıştır. En yüksek hassasiyet C&RT modeli; en yüksek AUC değeri ise yapay sinir ağları modeli ile elde edildiği için bu üç modelin okul terkinin belirlenmesinde kullanılabileceği sonucuna ulaşılmış ve terkte etkili olan faktörler bunlar aracılığı ile ele alınmıştır. Modellerin okul terki için sınıflandırma yaparken ilk iki sırada ağırlıklı genel not ortalamasını ve öğrenci kayıt yılı faktörünü kullandığı; daha sonra ise modellere göre önem sırası farklılaşsa da program türü, yaş, kayıt/geliş şekli, öğrenim tipi, program düzeyi, ikamet ili ve cinsiyet değişkenlerinin sınıflandırmada etkili olduğu sonucuna ulaşılmıştır. Araştırma sonucunun okul terkinin azaltılmasına yönelik geliştirilecek stratejilere katkı sunması beklenmektedir.

Anahtar Kelimeler: açık ve uzaktan öğretimde okul terki, eğitimde veri madenciliği, C&RT, Rastgele Ağaçlar, C5.0, YSA
Sayfa Adedi : xx + 180
Danışman : Prof. Dr. Dilara Bakan Kalaycıoğlu

PREDICTING THE FACTORS AFFECTING SCHOOL DROPOUT IN OPEN AND DISTANCE EDUCATION STUDENTS BY USING DATA MINING

(Ph. D Thesis)

Selma Tosun

GAZI UNIVERSITY

GRADUATE SCHOOL OF EDUCATIONAL SCIENCES

November 2024

ABSTRACT

The aim of this study is to predict the dropout and academic success of students studying in open and distance education with the data that can be obtained through the systems and to determine the variables that are effective in academic success and dropout by determining data mining models with high classification performance. In the study, a model was developed with data mining algorithms to identify students at risk of dropout early by taking demographic and personal data of open and distance education students in higher education and weighted GPA score as variables. In the study, the use of data mining approach in the field of education and the level of reliability and validity values of the results obtained based on these algorithms were determined. This research is in the relational research model, which is one of the quantitative research methods in terms of its purpose. The type of the research, the study group, the data collection tools and the procedures for the analysis of the data were handled in relation to the stages in the process of the Cross-Regional Standard Procedures for Data Mining (CRISP-DM). The research was conducted with the data obtained from the Istanbul University student information system database of all students enrolled in a programme at least once until the 2021-2022 spring semester at Istanbul University Open

and Distance Education Faculty (IU AUZEF). In the analysis process, decision trees such as Random Trees, C&RT, C5.0, CHAID, QUEST and various VM algorithms such as Naive Bayes (NB), Logistic Regression (LR), Artificial Neural Networks (ANN) and Support Vector Machine (SVM) were used to classify students according to dropout and academic achievement, and their performances were evaluated according to accuracy, sensitivity, precision, precision, negative predictive value, F1 score, MCC, AUC criteria. As a result of the study, it was observed that the overall correct classification rates of the models created for the classification of students according to school dropout ranged between 0.754 and 0.790. The model with the highest overall classification accuracy in the whole data set is C5.0. In addition, since the C5.0 model has the highest sensitivity, F1 score, precision, negative predictive value and MCC value compared to the others, it is concluded that the model has sufficient validity in predicting school dropout. Since the highest sensitivity was obtained with the C&RT model and the highest AUC value was obtained with the artificial neural networks model, it was concluded that these three models can be used in the determination of school dropout and the factors affecting dropout were addressed through them. It was concluded that the models used the weighted grade point average and the year of enrolment as the first two factors in classification for school dropout, and then the variables of programme type, age, enrolment/arrival type, education type, programme level, province of residence and gender were found to be effective in classification, although the order of importance differed according to the models. The results of the research are expected to contribute to the strategies to be developed to reduce school dropout.

Key Words : dropout in open and distance education, educational data mining, NeuralNet, C&RT, Random Trees, C5.0
Page Number : xx + 180
Supervisor **:** Prof. Dr. Dilara Bakan Kalaycıoğlu

İÇİNDEKİLER

TELİF HAKKI VE TEZ FOTOKOPİ İZİN FORMU i
ETİK İLKELERE UYGUNLUK BEYANI ii
JÜRİ ONAY SAYFASI iii
TEŞEKKÜR v
ÖZ vi
ABSTRACT viii
İÇİNDEKİLER x
TABLOLAR LİSTESİ xiv
ŞEKİLLER LİSTESİ xvii
KISALTMALAR xix
BÖLÜM I 1
GİRİŞ 1
1.1. Problem Durumu 1
1.2. Araştırmanın Amacı ve Önemi 8
1.3. Sayıltılar 10
1.4. Sınırlılıklar 10
1.5. Tanımlar 11
BÖLÜM II 12
KAVRAMSAL ÇERÇEVE 12

2.1. Açıköğretimde Okul Terki ve Devamlılık 12

2.1.1. Okul Terki 17

2.2. Eğitimde Veri Madenciliği 23

2.2.1. Veri Madenciliği Süreçleri ve Algoritmalar 26

2.2.1.1. Sınıflandırma Yöntemleri 31

2.2.1.1.1. Karar Ağaçları 31

2.2.1.1.2. Rastgele Ağaçlar 36

2.2.1.1.3. CHAID 38

2.2.1.1.4. C5.0 39

2.2.1.1.5. C&RT 41

2.2.1.1.6. QUEST 42

2.2.1.1.7. BAYES Algoritması 43

2.2.1.1.8. Lojistik Regresyon -LR 46

2.2.1.1.9. Yapay Sinir Ağları -NeuralNet 48

2.2.1.1.10. Destek Vektör Makineleri (DVM) 50

2.2.1.2. Kümeleme (Clustering) 52

2.2.1.3. Birliktelik Kuralı- İlişki Analizi (Association Rule Learning) 52

2.2.1.4. Zaman Serisi Analizi (Time Series Analysis) 52

2.3. İlgili Araştırmalar 52

2.3.1. Açık ve Uzaktan Öğrenmede Veri Madenciliği ile Yapılan Okul Terki veya Devamlılık Araştırmaları 53

2.3.2. Eğitimde Veri Madenciliği ile Yapılan Okul Terki veya Devamlılık Araştırmaları 59

2.3.3. Açık ve Uzaktan Öğrenmede (Yükseköğrenimde) Okul Terki ve Devamlılık Kavramına Odaklı Araştırmalar 70

2.3.4. Eğitimde Veri Madenciliği Çalışmaları; Öğrenci Başarı ve Performans Sınıflandırmaları 74

BÖLÜM III 78

YÖNTEM 78

3.1. Araştırma Modeli 78

3.2. Çalışma Grubu ve Prosedür 79

3.3. Verilerin Toplanması, Hazırlanması ve İşlenmesi 80

3.4. Verilerin Analizi 88

3.5. Sınıflandırma Performansının Değerlendirilmesi ve Geçerlik 89

BÖLÜM IV 95

BULGULAR 95

4.1. Verinin Keşfi ve Tanımlayıcı İstatistikler 95

4.2. Yükseköğretimde Açık ve Uzaktan Öğrenmedeki Öğrencilerin Okul Terkinin Tahmininde Veri Madenciliği Sınıflandırma Algoritmalarının Performansları Nasıl Değişmektedir? 110

4.3. Yükseköğretimde Açık ve Uzaktan Öğrenmedeki Öğrencilerin Okul Terki ile İlişkili Demografik Faktörler (Yaş, Cinsiyet, Medeni Durum, İkametgâh İli vb.) ve Diğer Nitelikler Nelerdir? 119

4.3.1. Tüm Değişkenlerin Yer Aldığı Veri Setinde Veri Madenciliği ile Öğrencilerin Okul Terkinin Tahmininde Hangi Faktörler Rol Oynamaktadır? 120

4.3.1.1. C5.0 120

4.3.1.2. C&RT 122

4.3.1.3. YSA 124

4.3.2. AGNO>0 Veri Setinde Yer Alan Öğrencilerin Veri Madenciliği ile Okul Terkinin Tahmininde Hangi Faktörler Rol Oynamaktadır? 127

4.3.2.1. Rastgele Ağaçlar.....127

4.3.2.2. CHAID.....130

4.3.2.3. C5.0.....132

4.3.2.4. YSA.....134

4.4. AGNO'ya göre Öğrencilerin Sınıflandırılmasında VM Modellerinin Performansları Nasıl Değişmektedir ve Öğrenci Başarısı ile İlişkili Faktörler Nelerdir? 136

4.4.1. Rastgele Ağaçlar.....139

4.4.2. C5.0.....141

4.4.3. C&RT.....142

BÖLÜM V 145

SONUÇ, TARTIŞMA ve ÖNERİLER.....145

5.1. Sonuç ve Tartışma.....145

5.2. Öneriler.....156

KAYNAKLAR.....159

EKLER.....175

EK 1. Tüm veri için C5.0 algoritmasının ilk 4 düzeyi.....176

EK 2. Okul terki öngörüsü için rastgele ağaçlar modelinde en iyi (TOP RULES) karar verme kuralları.....177

EK 3. AGNO> veri seti için C5.0 modelinin ağaç yapısı (3. düzey) ile 2. düzeyde düğümler ve değişkenlerin önem katsayıları.....178

EK 4. 'AGNO_2' için rastgele ağaçlar modelinde en iyi (TOP RULES) karar verme kuralları.....179

TABLOLAR LİSTESİ

Tablo 1. *Okul Terki ile İlgili Tanımlar* 20
Tablo 2. *2006-2016 Yılları Arasında Konusuna Göre EVM Çalışmaları* 24
Tablo 3. *EDM Yöntemleri Kullanılarak Öğrenci Terkinin Tahmin Edilmesi* 25
Tablo 4. *Karar Ağacı Algoritmaları Gelişim Süreci* 34
Tablo 5. *C&RT ve CHAID Karar Ağaçlarının Karşılaştırılması* 41
Tablo 6. *EDM Yöntemleri Kullanılarak Öğrenci Terkinin Tahmin Edilmesi* 82
Tablo 7. *Öğrenci Statüsü Verisinin Yeniden Kodlanması* 83
Tablo 8. *Öğretim Programları ve Programın Öğrenci Kabulüne Başladığı Yıl* 86
Tablo 9. *Araştırmada Yer Alan Değişkenler ve Olası Değerleri* 87
Tablo 10. *İkili Sınıflandırmalar İçin Karışıklık-Confusion Matrisi* 91
Tablo 11. *Sınıflandırma Performansının Hesaplanmasında Kullanılan Formüller* 92
Tablo 12. *AUC Değerlendirme Kriterleri (Hosmer)* 94
Tablo 13. *Veri Setleri ve Toplam Veri Miktarı* 96
Tablo 14. *Öğrenim Tipine Göre Öğrenci Sayıları* 96
Tablo 15. *Analizdeki Tüm Veriye Göre Öğretim Programlarındaki Öğrenci Sayıları* 97
Tablo 16. *Program Düzeyine Göre Öğrenci Sayıları* 100
Tablo 17. *Giriş Yılı Değişkenine Göre Öğrenci Sayıları* 100
Tablo 18. *Açık ve Uzaktan Öğrenme Programlarındaki Öğrencilerin Geliş/Kayıt Şekli* 101
Tablo 19. *Uyruk Değişkenine Göre Öğrenci Sayıları* 102
Tablo 20. *Cinsiyet Değişkenine Göre Öğrenci Sayıları* 102
Tablo 21. *Yaş Kategorilerindeki Öğrenci Sayıları (Tüm Veri)* 104
Tablo 22. *İl/İkametgâh Değişkenine Göre Öğrenci Sayıları* 105

Tablo 23. *Tüm Veri İçinde AGNO1 Değişkeni Kategorilerindeki Öğrenci Sayıları* 108
Tablo 24. *Veri Madenciliği Sınıflandırma Algoritmalarının Okul Terki Tahmininde Tüm Veriye Ait Sınıflandırma Performansları*.....110
Tablo 25. *Terk Tahmini: AGNO>0 için Sınıflandırma Algoritmalarının Okul Terki Sınıflandırma Performansları*.....114
Tablo 26. *Tüm Veri ve AGNO>0 Verisi ile Yapılan Analizlerde Model Performanslarının Karşılaştırılması* 117
Tablo 27. *Tüm Veri Seti İçin C5.0 Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri* 120
Tablo 28. *Tüm Veri Seti İçin C&RT Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri* 122
Tablo 29. *Tüm Veri Seti İçin YSA Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri* 126
Tablo 30. *Okul Terkine Etki Eden Faktörlerin Modellerde Karşılaştırması (Tüm Veri)* ..126
Tablo 31. *AGNO>0 Veri Seti İçin CHAID Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri*.....130
Tablo 32. *AGNO>0 Veri Seti İçin C5.0 Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri*.....133
Tablo 33. *AGNO>0 Veri Seti İçin YSA Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri*.....135
Tablo 34. *AGNO>0 Veri Seti için Okul Terkine Etki Eden Faktörlerin Karşılaştırılması*135
Tablo 35. *Öğrenci Başarısının (AGNO) Tahmini İçin VM Algoritmalarının Performanslarının Karşılaştırılması*.....137
Tablo 36. *Öğrenci Başarısının Tahmininde C5.0 Modelindeki Tahmin Değişkenleri ve Önemi*.....141
Tablo 37. *Öğrenci Başarısının Tahmininde C&RT Modelindeki Değişkenler ve Önem Dereceleri* 142

Tablo 38. *Öğrenci Başarısı ile İlişkili Faktörlerin Karşılaştırılması* 144

ŞEKİLLER LİSTESİ

Şekil 1. Bean ve Metzner'in okulu bırakma modeli....14
Şekil 2. Berge ve Huang'ın kavramsal okul terki modeli....15
Şekil 3. Kember'in uzaktan eğitimde okulu bırakma modeli....16
Şekil 4. Park'ın yenilenmiş bileşik devamlılık modeli....17
Şekil 5. Birinci dönemden üçüncü döneme kadar yeni öğrenciler-akış şeması....19
Şekil 6. Veri Madenciliği metodolojisinin tipik döngüsü....27
Şekil 7. CRISP-DM Modeli ve aşamaları....28
Şekil 8. Veri madenciliği yöntemleri....28
Şekil 9. Veri Madenciliği ve ilişkili disiplinler.....29
Şekil 10. Veri madenciliği yöntem ve algoritmaları....30
Şekil 11. Sınıflandırma için genel bir karar ağacı örneği....32
Şekil 12. Bayes ağı; Lauritzen ve Spegelhalter'in Asya ağı örneği....45
Şekil 13. Basit bir Naive Bayes ağacının yapısı (TAN)....46
Şekil 14. Sinir ağı düğüm yapısı....49
Şekil 15. Araştırma prosedürü....79
Şekil 16. Tüm veri seti için değişkenlerin önemlilik testi....84
Şekil 17. Özellik seçim fonksiyonuyla elde edilen değişkenler....85
Şekil 18. İyi bir model için (sol) ve bilgi sağlamayan bir model için (sağ) ROC eğrisi....93
Şekil 19. Programlardaki öğrenci sayıları (Tüm veri)....98
Şekil 20. Programlardaki öğrenci sayıları (AGNO>0)....99

Şekil 21. Giriş yılı değişkenine göre öğrenci devam/terk durumlarının dağılımı 101
Şekil 22. Cinsiyet değişkenine göre öğrencilerin devam/terk durumları 103
Şekil 23. Tüm veri için öğrenci yaş dağılımı 103
Şekil 24. Yaş kategorilerine göre öğrencilerin terk durumları 104
Şekil 25. Yaş değişkeninin veri gruplarında dağılımı 105
Şekil 26. İl/ikametgâh değişkenine göre öğrenci terk durumları 106
Şekil 27.Tüm veri setinde AGNO dağılımı 107
Şekil 28. AGNO2 değişkeninde başarılı ve başarısız öğrencilerin dağılımı 107
Şekil 29. Tüm veri ve AGNO>0 verisinde not kategorilerinin dağılımı 108
Şekil 30. Ağırlıklı genel not ortalaması değişkenine göre öğrenci terk oranları 109
Şekil 31. Öğrenci statü değişkeninin dağılımı; devam ve terk oranları 109
Şekil 32. Sınıflandırma algoritmalarının tüm veri seti için okul terki sınıflandırma performansının karşılaştırılması 111
Şekil 33. Tüm veri seti için VM modellerinin ROC eğrileri 113
Şekil 34. Terk tahmini; AGNO>0 veri seti için sınıflandırma modellerinin ROC eğrileri 116
Şekil 35. Tüm Veri ve AGNO>0 verisi için elde edilen VM modellerinin karşılaştırmalı performans grafikleri 118
Şekil 36. Tüm veri seti için C&RT modelinin ağaç yapısı 123
Şekil 37. Tüm veri seti için YSA modelinde sinir ağları şeması 125
Şekil 38. AGNO>0 veri seti C5.0 modelinde tahmin değişkenleri ve önemi 128
Şekil 39. AGNO>0 veri seti için YSA modelinde sinir ağları şeması 134
Şekil 40. Öğrenci başarısı sınıflandırmasında modellerin ROC eğrileri 138
Şekil 41. Öğrenci başarısının tahmininde RA modelinin tahmin değişkenleri ve önemi .. 139
Şekil 42. Öğrenci başarısı sınıflandırmasında C5.0 modelinin ağaç yapısı (2. düzey) 141
Şekil 43. Öğrenci başarısı sınıflandırmasında C&RT modelinin ağaç yapısı (3. düzey) ... 143

KISALTMALAR

AUZEF	Açık ve Uzaktan Eğitim Fakültesi
OULAD	(Open University Learning Analytics Dataset) Açık Üniversite Öğrenme Analitiği Veri Seti
İÜ	İstanbul Üniversitesi
VM	Veri Madenciliği
EVM	Eğitimde Veri Madenciliği
MOOC/KAÇK	(Massive Open Online Course) Kitlesel Açık Çevrimiçi Kurs
OECD	(Organisation for Economic Co-operation and Development) Ekonomik İşbirliği ve Kalkınma Teşkilatı
ÖSYM	Ölçme Seçme ve Yerleştirme Merkezi
AGNO	Ağırlıklı Genel Not Ortalaması
MERNİS	Merkezi Nüfus İdaresi Sistemi
İÜ AKSİS	İstanbul Üniversitesi Akademik Kayıt ve Öğrenci Bilgi Sistemi
DGS	Dikey Geçiş Sınavı
ÖSS	Öğrenci Seçme Sınavı-

YKS	Yükseköğretim Kurumları Sınavı
YOS	Yabancı Öğrenci Sınavı
YSA	Yapay Sinir Ağları
RA	Rastgele Ağaçlar
RO	Rastgele Orman
C&RT	Sınıflandırma ve Regresyon Ağaçları
LR	Lojistik Regresyon
BAYES	Bayes Sınıflandırıcısı
CHAID	(Chi Squared Automatic Interaction Detector) Ki Kare Otomatik Etkileşim Detektörü
QUEST	(Quick, Unbiased, Efficient, Statistical Tree) Hızlı, Yansız, Etkin, İstatistiksel Ağaç
DVM	Destek Vektör Makinesi

BÖLÜM I

GİRİŞ

Bu bölümde problem durumu ortaya konulmuş, araştırmanın amacı ve önemi hakkında bilgilere yer verilmiştir.

1.1. Problem Durumu

Açık ve uzaktan öğrenme özellikle sanayi devrimi ile birlikte eğitime erişimi artırmanın bir yolu olarak görülmüş; 21. yy'da ise gelişen teknoloji ve eğitimin toplum hayatında artan önemine binaen yaşam boyu öğrenmenin yöntemlerinden biri olmuştur. Sınırlı kaynaklar ile kitleleri eğitmenin yolu olan uzaktan öğrenme araçları, teknolojinin gelişimi ile birlikte farklı safhalardan geçmiştir. Mektup, gazete, TV ve radyo, video içerikler ve nihayetinde internet tabanlı öğrenme materyalleri sınıf ortamında eğitime katılamayacak olanlara erişebilmek için kullanılmıştır. Benzer şekilde teknoloji gelişimi ile uzaktan öğrenme kavramı da "açıköğretim, dijital eğitim, çevrimiçi öğrenme, e-öğrenme" vb. ile birlikte kullanılmaya başlanmıştır.

Uzaktan öğrenmenin ilk planlı uygulamasının 20 Mart 1728 tarihli Boston gazetesinde ilan edilen haftalık uzaktan "Steno dersi" öğreticiliği ile başladığı kabul edilir. 1856 yılında Charles Toussaint ve Gustaf Langenscheidt tarafından Almanya'da İngilizce ve ardından Almanca dil eğitiminin yazışmalı (eğitim alan ile eğitmenin karşılıklı yazışması) olarak yapıldığı tespit edilmiştir (Noffsinger, 1926). Amerika Birleşik Devletleri'nde (ABD) uzaktan eğitimin anası olarak bilinen Anna Eliot Tickner, kadınların evden eğitimlerini sürdürebilmelerini sağlamak için 1873 yılında mektupla öğretim yöntemini başlatmıştır (Mathieson, 1971). Avustralya'da Queensland Üniversitesi de 1911 yılında okula başlayamamış olan ilk ve ortaokul çağındaki çocuklar için sistematik ve kitlesel olarak mektupla eğitimi başlatmıştır (Holmberg, 2008).

Türkiye'de uzaktan eğitim faaliyetleri hem özel sektör hem de resmi kurumlar seviyesinde 1950 yıllarında başlamıştır. Limasollu Naci ve FONO, 1953 yıllında dil öğretimi için yeni bir uygulama olan mektupla öğretim tekniğini kullanarak abonelerine üniteler halinde konuları posta yoluyla iletmiştir (Geray, 2007). Bundan sonra Milli Eğitim Bakanlığı bünyesinde, Öğretici Filmler Merkezi (1951); Mektupla Öğretim Merkezi (1960); Yaygın Yükseköğretim Kurumu (1975) ve 1992 yılında lise seviyesinde uzaktan eğitim hizmetleri başlatılarak Açık Öğretim Lisesi kurulmuştur.

Üniversitelerde ise ilk 1982 yılında sistematik olarak Anadolu Üniversitesi açık öğretime başlamış ve hâlâ faaliyetlerine devam etmektedir. 2010 yılında İstanbul Üniversitesi ve Erzurum Atatürk Üniversitesi de açık öğretim kapsamında öğrenci almaya başlamıştır. 2021 yılında ise Ankara Üniversitesi'nde açık öğretim fakültesi kurulmuştur. Yıllar içinde, kurulan fakültelerdeki program sayısı ve çeşidi de artmıştır.

Türkiye'de yükseköğrenimde hem açık ve uzaktan öğretim bölümleri hem de bu öğretimi tercih eden öğrenci sayısı her geçen yıl artmaktadır. Bununla birlikte programlara yerleşim şartlarında yıllar içinde birtakım değişiklikler olmuştur. Türkiye'de daha önce örgün veya açıköğretim herhangi bir yükseköğretim programına yerleşmenin ön koşulu merkezi sınavdaki puan barajını geçmek iken; 2012 yılındaki mevzuat değişikliğiyle ikinci bir üniversite diploması almak isteyenler için sınava girmeden açıköğretim programlarına kayıt olanağı tanınmış; böylece açık öğretime olan talep daha da artmıştır. 2020 Yükseköğretim Kurulu (YÖK, 2021) verilerine göre lisans ve ön lisans düzeyinde yükseköğretimde toplam 7.791.280 öğrenci bulunmaktadır. Bu öğrencilerin yarısından fazlası (%56) öğrenimine açık veya uzaktan öğretim ile devam etmektedir. Türkiye'de öğrenci kabul eden üç açıköğretim üniversitesindeki programlara 2019'da 3.889.506; 2020'de 4.126.350; 2021'de 4.369.521 öğrencinin ön lisans ve lisans düzeyinde kayıtlı olduğu görülmektedir (YÖK, 2021).

Ülkemizde 2020-2021 eğitim öğretim yılında açık ve uzaktan öğretime kayıt olan 737.439 öğrenci bulunurken; mezun olan öğrenci sayısı ise 407.962'dir (YÖKSİS, 2023) . Hem dünya yükseköğretiminde hem de Türkiye'de alarm seviyesinde takip dilen konulardan biri uzaktan eğitimde öğrenimine devam etmeyen öğrencilerdir. Avrupa 2020 stratejisi hedeflerinden biri 30-34 yaş grubundaki gençlerin en az %40'ının yükseköğretimi tamamlamasıdır (Akbaş & Apar, 2010) . Bu hedefe ulaşmak için yükseköğretimde okul bırakma oranlarının azaltılması, ileri yaş grubunda öğrenimin desteklenmesi ve tamamlama oranlarının arttırılması kilit stratejilerden biri olarak tanımlanmıştır (Vossensteyn, Kottmann, Jongbloed vd., 2015). Öte

yandan Avrupa Uzaktan Eğitim Üniversiteleri Birliği (EADTU) açık ve uzaktan eğitim üniversitelerinde azalan kayıtlar ve yükselen okulu bırakma oranlarına ilişkin endişelerini çeşitli vesilelerle paylaşmaktadır (Marr, Gribble & Ubachs, 2024). Kore Ulusal Açık Üniversitesi de özellikle ilk dönemde okulu bırakan öğrencilerin yüksekliğinden endişelidir (Zawacki-Richter & Qayyum, 2019). Benzer şekilde Türkiye'de son yıllardaki YÖK raporlarında da açıköğretim programlarındaki milyonlar ile ifade edilen pasif, terk eden, donduran, eğitim faaliyetlerine katılmayan vb. öğrenci sayıları ile düşük mezuniyet oranları dikkat çekmektedir. 2019 verilerine göre Anadolu Üniversitesi açıköğretim programlarında yaklaşık 890 bin aktif öğrenci bulunurken 2,4 milyondan fazla pasif öğrenci yer almaktadır (Aguş, 2019). Açıköğretimde, sınıf öğretiminden farklı olarak sabit bir mekânda ve zamanda öğrenci varlığı kayıt altına alınmadığı için bu durum eğitim araştırmacılarında öğrenenlerin eğitime bağlılığının ve devamında uzaktan öğrenmenin başarısının da tartışılmasına yol açmaktadır.

Resmî Gazete 'de yayımlanan 17 Nisan (2020) tarih ve 31102 sayılı, Yükseköğretim Kanunu ile Bazı Kanunlarda Değişiklik Yapılmasına Dair Kanunla birlikte mevzuata "Açık öğretim sisteminde üst üste dört dönem kayıt yenileme koşullarını yerine getirmeyen öğrencinin ilgili programdan ilişiği kesilir." hükmü eklenmiştir. Bu durumda 4 eğitim öğretim dönemi pasif olan öğrencilerin programa daha sonraki yıllarda da devam etmeyeceği; dolayısıyla okulu terk ettiği varsayılmaktadır. Ayrıca, 2020'deki 31102 sayılı mevzuatta yer alan şekliyle üst üste dönemlerde kayıt yenilemeyen dolayısı ile pasif olarak tanımlanan öğrencilerin 2021-2022 eğitim öğretim yılı bahar döneminden itibaren programları ile ilişikleri kesilmiştir. 4 dönem kriterinin getirilmiş olması özellikle ara verme eğiliminde olan öğrencilerin devamını sağlayabilmek için tedbirlerin bir an önce alınmasını gerektirmektedir. Bu sebeple erken dönemde terk etme olasılığı yüksek olan öğrencilerin belirlenmesi önem kazanmaktadır. Ancak bu zamana dek hem Türkiye hem de dünyadaki örneklerinde açık öğretim sistemi içindeki pasif öğrencileri veya programları terk eden öğrencileri belirlemeye yönelik (Grau-Valldosera & Minguillón, 2014; Lee & Choi, 2011; Marr vd., 2024; Thomas & Unger, 2015) bir karmaşa mevcuttur. Bu karmaşanın temelde iki nedenini olduğu görülmektedir; ilkinde açık ve uzaktan öğrenmenin doğasından kaynaklı olarak terk eden öğrencinin fonksiyonel tanımı üzerinde görüş ve uygulama birliğinin olmaması ve dolayısı ile niteliklerinin ortaya çıkartılamaması; bir diğer husus da terk edecek öğrencilerin nasıl öngörülebileceğidir. Grau-Valldosera ve Minguillón'a (2014) göre ise okulu bırakmanın uygun bir tanımını elde etmek, nedenlerini açıklamak kadar zordur. Özellikle devam şartı olmayan uzaktan eğitim programlarında ardışık

yarıyıllarda ara veren öğrencilerin programa devam etmeme olasılığı da bulunduğundan okul terki, açık ve uzaktan öğrenmede tanımlaması oldukça zor bir kavramdır. Márquez ve diğerlerine (2016) göre okulu bırakma geç ve erken dönem okul terki olarak literatürde ayrılmaktadır.

ABD'deki örgün eğitim programlarında yükseköğretimi bırakmanın temel özelliklerini araştıran bir rapora göre (CollegeAtlas, 2014), birinci sınıf öğrencilerinin %30'u üniversitedeki ilk yıllarından sonra okulu bırakmaktadır. Bir başka eğitim raporunda (OECD, 2010), başladıkları yükseköğretim programından mezun olmayan öğrencilerin oranı ortalama %31 (ABD'de %50'den fazla, Japonya'da %10'a kadar) olup bu oranlar, okulu bırakmanın (veya tamamlamanın) nasıl ele alındığına göre değişmektedir. İngiltere Açık üniversitesinde yapılan çalışmalar, öğrencilerin yaklaşık %30'unun bir açıköğretim programına kaydolduktan bir dönem sonra programı terk ettiğini ve sonraki dönemlerde tekrar aktif olmadığını göstermektedir. Uzaktan eğitimde öğrenci tutmayı araştıran Simpson (2013), birçok uzaktan eğitim kurumunda mezuniyet oranlarının geleneksel öğrenimin dörtte birinden daha az olduğunu; bunun ise bir uzaktan eğitim açığı olduğunu öne sürmektedir. Valldosera'nın (2019) çalışmasını yürüttüğü İspanya örneğinde görüleceği üzere, üniversitede uzaktan eğitimde okulu bırakma oranı yüz yüze eğitime göre oldukça yüksektir: %60,5'e karşı %24.

Reed ve diğerlerine (2013) göre geleneksel ve çevrimiçi uzaktan eğitim metodolojisinde ilgili öğrenci profilleri arasındaki farklar (çevrimiçi ortamda, genellikle eğitime ek olarak iş ve aile yükümlülükleri olan yetişkinler) göz önüne alındığında, çevrimiçi uzaktan eğitimde okulu bırakmanın yüz yüze muadilinden daha sık ve farklı niteliklere sahip olduğu görülmektedir. Grau-Valldosera ve diğerleri (2018) öğrencilerin yaklaşık %33'ünün ikinci yarıyılda eğitimlerine ara verdiği görüldüğünden ilk yarıyılın stratejik öneme sahip olduğunu ve bu yarıyılın öğrencilerin programdaki kalıcılığını etkilediğini söylemektedir. İkinci yarıyılda eğitimlerine ara veren öğrencilerin sadece %6,2'si üçüncü yarıyılda döndüğü için bir programın "uzun mesafe yarışı" olduğunu belirtmek önemlidir.

Yükseköğretim içinde örgün eğitimde okul terkine yönelik; ABD'de bırakmaya etki eden faktörlerin araştırıldığı (CollegeAtlas, 2014) çalışmalar ile OECD'nin pek çok ülkede mezun olamayan öğrencilerin oranına yönelik raporları (OECD, 2010) mevcuttur. Okul terkine yönelik açık öğretimde yapılan araştırmalarda ise; öğrencilerin kayıt oldukları programdan beklentileri (Aguş, 2019) teknolojiye erişimde yaşanan sorunlar, öğrencide oluşan motivasyon eksikliği ve diğer bireysel değişimler (Aydın, Öztürk, Büyükköse vd., 2019), öğrenim için yeterli zamana

sahip olmama, sınavların zorluğu, maddi yetersizlikler, yaş, cinsiyet (Aulck, Velagapudi, Blumenstock vd., 2017), çocuk sahibi olma, kırsal bölgelerde yaşama gibi demografik nedenler (Bayer, Bydzovská, Géryk vd., 2012) ile çalışma durumu, eğitim düzeyi gibi farklı pek çok faktörün etkili olduğuna dair sonuçlar elde edilmiştir.

Ayrıca Cho ve Heron (2015), Reed vd. (2013) ile Grau-Valldosera vd. (2018), üniversitelerde açık öğretimdeki terk oranlarını ve yüksek terk oranlarının risklerini belirlemeye yönelik çalışmalar yapmıştır. Aguş (2019) ise 302 lisans ve ön lisans öğrencisi ile yürüttüğü çalışmada Anadolu Üniversitesi açıköğretim sistemi öğrencilerinin okulu bırakma nedenleri ve bu nedenlerin demografik gruplar arasındaki farklılıklarını ortaya çıkarmayı amaçlamıştır. Araştırma sonucunda; öğrencilerin okulu bırakmalarında en çok etkisi olan faktörün yeterli zaman bulamaması olduğu; bu durumun cinsiyete göre de farklılaştığı görülmüştür. Ayrıca öğrencilerde oluşan motivasyon eksikliğinin, maddi yetersizliklerin, çalışma durumunun da diğer önemli etkenlerden olduğu ortaya çıkmıştır. Kayhan (2019) Amasya Üniversitesi uzaktan eğitim öğrencilerinin mezuniyet durumlarını veri madenciliği yöntemleri ile tahminine yönelik yaptığı çalışmada programlara kayıt yaptıran ön lisans öğrencilerinin zamanında mezun olabilme durumlarına ilişkin çıkarımlarda bulunmuştur.

Yükseköğretimdeki okul terki ile ilgili araştırmalar ilk ve ortaöğretime göre daha azdır. Türkiye'de var olan akademik çalışmaların çoğunluğu ise kısıtlı bir örneklemden elde edilen ankete dayalı veriler ile gerçekleştirilmiştir. Ayrıca Türkiye'de yükseköğretimdeki açık ve uzaktan öğrenme programlarındaki okul terki ile ilgili çalışmalar oldukça sınırlıdır (Aydın vd., 2019; Qiu, Liu & Liu, 2018). Genel olarak okul terkiyle ilgili araştırmalar daha çok örgün eğitim üzerine odaklanmaktadır. Özellikle kitle eğitiminin yaygın olduğu ve program kabul şartlarının daha esnek tanımlandığı açık öğretim programlarında programlara kaydolan, devam eden ve terk eden öğrencilerin araştırılması başka pek çok açıdan önemlidir. Açık öğretimde okul terk oranları ve terk eden öğrenci nitelikleri dikkat çekici şekilde farklılaşmaktadır. Her yıl binlerce öğrenci açık öğretim öğrencisi olur iken bunların ancak bir kısmının aralıksız öğrenimine devam ettiği ve pek azının mezun olduğu görülmektedir. Okul terkini azaltıcı önlemler geliştirerek verimli bir öğrenme sürecinin oluşturulabilmesi için terk riski taşıyan öğrencilere yönelik etkili araştırmalar yapılması ve terke sebep olacak faktörlerin erken dönemde belirlenmesi, eğitimin ekonomik değeri ve verimliliği açısından hem eğitim planlamacıları hem de üniversite yönetimleri için önemli görünmektedir. Delen'in (2011) de belirttiği gibi üniversite sıralamalarını, okul itibarını ve finansal refahı etkileyen "öğrenciyi elde tutma", yükseköğretim kurumları için en önemli başarı ölçütlerinden biri haline gelmiştir.

Belirli bir ücret ile yükseköğretime kaydolan ancak pek çok etken ile okulu terk eden öğrencilerin, terkine etki eden faktörlerin keşfi özellikle genç nüfusun yoğun olduğu ve büyük bir kesimin eğitim sürecinde yer aldığı ülkelerde ekonomik gelişim açısından da dikkat edilmesi gereken bir değişkendir.

Öte yandan akademik başarı, akademik kalıcılık ve okul terkini araştıran çalışmalar yöntem olarak daha çok nitel veya karma desenler kullanmakta (Altınkurt, 2008; Bayhan & Dalgıç, 2012; Gökşen vd., 2008; Şimşek, 2011; Şirin & Uysal, 2007; Özer vd., 2011; Özdemir & Taylı, 2008); son dönemlerde ise veri madenciliği yöntemleri ile de yapılan çalışmalar dikkat çekmektedir (Agrusti vd., 2019; Aulck vd., 2017; Gil vd., 2020; Kotsainatis vd., 2003). Eğitimde veri madenciliği, verileri stratejik bir varlık olarak kullanarak eğitim kalitesini artırma ve öğrenme süreçlerini optimize etme potansiyeline sahiptir. Bu yönde bir takım çalışmalar yapılmış olsa da yine de dikkate alınması gereken birçok parametre vardır (Bonde & Kirange, 2018). Araştırmacılardan bazıları, öğrencilerin davranışlarını hesaplamanın yolunu bulmak için model geliştirmiş diğer bir kısmı ise öğrencilerin memnuniyetini ve ilerleme düzeylerini belirlemiştir. Ceyhan (2020) PISA verileri ile birtakım VM sınıflandırma yöntemlerinin değişken sayısı ve örnek büyüklüğü açısından performanslarını karşılaştırmıştır. Bununla birlikte akademik başarı veya okul terkini tahmin etmeye yönelik doğal veriler kullanıldığında, veri madenciliği sınıflandırma algoritmalarının performans değerlerinin nasıl değiştiğini ve geçerliliğini gösteren bir çalışmaya rastlanmamıştır. Dolayısı ile bu çalışmada özellikle okul terkini öngörebilecek geçerli bir sınıflandırma methodolojisinin bulunması da önemlidir. Özellikle büyük veriler ile açıköğretimde okul terkini tahmine yönelik, veri madenciliği çalışmaları bulunmamaktadır.

Okul terkinde etkili olan faktörler (kişisel, sosyal, çevresel, psikolojik, kurumsal vb). üzerinde araştırmalar yoğunlaşsa da açıköğretimin doğası gereği özellikle sisteme kaydolan ancak öğrenim faaliyetinde bulunmayan, terk veya devam etme potansiyeli olan öğrencilerin erken belirlenebilmesi de önemli görülmektedir. Öğrenci kaybına yönelik yakın zamana kadar yapılan araştırmaların ekseriyeti bu fenomonolojiyi anlamaya odaklanırken davranışsal ekolde, nitel veya ankete dayalı bu çalışmalar, çeşitli teorilerin ortaya konulmasında etkili olsalar da okullarda öğrenci kaybını doğru şekilde tahmin etmek, olası iyileştirmeler için yol gösterici olmak ve ihtiyaç duyulan kurumsal araçları sağlamaktan uzaktırlar (Delen, 2010; Veenstra, 2009). Bu noktada özellikle kurumların var olan veri yapıları üzerinde terk etme riski bulunan öğrencilerin tespiti önem kazanmaktadır.

Buradan hareketle terk olgusunun sebep ve sonuçlarını bulmak yerine bu çalışmada, mümkün olan en kısa sürede ve öğrencilerden ilk aşamada doğal olarak elde edilebilecek değişkenler aracılığıyla okul terkinin uygun veri madenciliği modelleri ile belirlenmesine odaklanılmıştır. Bu çalışmada bir programa ücret ödeyerek kaydolmuş öğrencinin, birinci dönemin sonunda kayıt yenilemediği/ders seçimi yapmadığı, kayıt sildirdiği, kayıt dondurduğu veya "devamlı öğrenci" statüsünde yer almadığı pasif tüm durumlar okul terki olarak ele alınacaktır.

Bu araştırma kapsamında üniversitede açık veya uzaktan eğitim programlarına kayıt yaptıran öğrencilerin demografik özellikleri ve diğer kişisel bilgilerinden yola çıkarak erken dönemde okul terki ve akademik başarı durumlarını tahmin edecek etkili veri madenciliği algoritmalarının belirlenmesi ile sınıflandırma performansı yüksek veri madenciliği modellerine göre akademik başarı ve okul terki ile ilişkili değişkenlerin ortaya çıkartılması amaçlanmıştır.

Buna göre;

1. Yükseköğretimde açık ve uzaktan öğrenmedeki öğrencilerin okul terkinin tahmininde veri madenciliği sınıflandırma algoritmalarının performansları nasıl değişmektedir?
    i. Tüm değişkenlerin yer aldığı veri setinde VM sınıflandırma algoritmalarının performansları nasıl değişmektedir?
    ii. En az 1 dersten başarılı (AGNO>0) olan öğrencilerin okul terkinin tahmininde VM sınıflandırma algoritmalarının performansları nasıl değişmektedir?
2. Yükseköğretimde açık ve uzaktan öğrenmedeki öğrencilerin okul terki ile ilişkili demografik faktörler (yaş, cinsiyet, medeni durum, ikametgâh ili vb.) ve diğer nitelikler nelerdir?
    i. Tüm değişkenlerin yer aldığı veri setinde veri madenciliği ile öğrencilerin okul terkinin tahmininde hangi faktörler rol oynamaktadır?
    ii. AGNO>0 veri setinde yer alan öğrencilerin veri madenciliği ile okul terkinin tahmininde hangi faktörler rol oynamaktadır?
3. Yükseköğretimde açık ve uzaktan öğrenmedeki öğrencilerin Ağırlıklı Genel Not Ortalamasının (AGNO) tahmininde;
    i. veri madenciliği sınıflandırma algoritmalarının performansları nasıl değişmektedir?
    ii. Veri madenciliği modellerine göre AGNO ile ilişkili faktörler nelerdir?

problemlerine bu araştırma sürecinde cevap aranmaktadır.

1.2. Araştırmanın Amacı ve Önemi

Eğitimin sosyo-ekonomik değişkenler ve gelecek beklentisi üzerinde yadsınamaz bir etkisi vardır. Genç nüfusun fazla olduğu ülkelerde eğitim özellikle toplumun gelişimi için bir kaldıraç rolü üstlenmektedir. Öğrenciler her yıl pek çok hayal ile üniversite sınavlarına girmektedir. İyi bir mesleğin ilk koşulu olarak nitelikli bir eğitim hakkı isteyen genç nüfusun bir bölümü daha sonra tercih ettikleri programları ya hiç tamamlamamakta ya da geç bitirerek hayallerini ertelemekte veya bundan vazgeçmektedir. Üniversitenin bitirilmesinin ötelenmesi veya terk edilmesi, devamında bireyin toplum için de var olmasının ertelenmesi anlamına gelmektedir. Türkiye'de 12 yıllık zorunlu eğitim ve devamında sınavlara hazırlık süreçleri de düşünülürse en erken 23 yaşında bir yükseköğretim lisans programından mezun olacak birey, okul terki ile ya istediği mesleğe hiç başlayamamakta ya da iş hayatına geç adım atmaktadır. Öte yandan zorunlu eğitimin hemen ardından çalışmaya başlayan veya bir yükseköğretim programında hem çalışan hem de eğitimine devam eden öğrenciler de ya eğitimlerini çok uzun sürelerde tamamlamakta ya da bitirememektedir.

OECD istatistiklerinde toplumun gelişmişlik düzeyinin belirlenmesinde ülkelerin yükseköğrenimi tamamlayan öğrenci oranları önemli bir değişkendir. Gelişmekte olan ülkeler için nitelikli insan gücü anlamına gelen en az bir yükseköğretim programını bitiren birey, toplumların gelişmesine ön ayak olmaktadır. Açıköğretim programlarının geniş kitleleri hızlıca eğiterek sanayi devrimi için nitelikli insan gücü oluşturma misyonu, 21. yy ile birlikte hayat boyu öğrenme ve değişen şartlar için insanlara yeni meslekler kazandırmaya doğru evrilmiştir. Buna paralel üniversiteye yönelik talep artarken, üniversite terki de şaşırtıcı şekilde artmaktadır.

Amerika Birleşik Devletleri'nde yükseköğretim için federal finansmanın sağlanmasında ve Yüksek Öğrenim Komisyonunun uzaktan eğitimin değerlendirilmesi kriterleri arasında öğrenci tutma oranlarının da olduğu bildirilmektedir (Lykourentzou, Giannoukos, Nikolopoulos vd., 2009). Benzer şekilde Avrupa 2020 stratejisindeki hedeflerden biri, yetişkinlerin (30-34 yaşındaki) en az %40'ının yükseköğrenimi tamamlamasıdır (Akbaş & Apar, 2010). Bu nedenle hem Avrupa'da üretkenliği ve sosyal adaleti teşvik etmeye yönelik yararlı olan üst düzey becerileri oluşturmak, hem de çok önemli kabul edilen yükseköğrenim tamamlama hedefine ulaşmak için, yükseköğretimde okulu bırakmanın azaltılmasına ve akademik kalıcılığın iyileştirilmesine artan bir ilgi vardır (Vossensteyn vd., 2015). Bu ilgi özellikle yükseköğretimde okul terki üzerine hazırlanmış komisyon raporlarında görülmektedir (Thomas & Unger, 2015).

Şu anda, akademik kurumların zorluklarından biri, okulu bırakan öğrenci sorunlarıdır. Hangi öğrencilerin okulu terk etme riski altında olduğunu ve okul terki altında yatan faktörlerin neler olduğunu anlamak çok önemlidir (Gil, Jhone, Delima vd., 2020). Okulu terk eden öğrencilerin incelenerek okul terkinde etkili olan faktörlerin belirlenmesiyle yöneticiler ve öğretmenler risk altındaki öğrencilere gerekli desteği vererek öğrencilerin eğitimlerini sürdürmelerine destek olabilecektir. Yapılan araştırmalar ile öğrenci performansına ilişkin bulgular sonucu üniversiteler, öğrencilerin performansını etkileyen faktörleri belirleyebilir ve bir öğrencinin davranışını tahmin edebilirse, yöneticiler tarafından uygun önlemler alınabilir ve sonuç olarak bu, üniversitenin kalitesini artırmaya yardımcı olur (Bonde & Kirange, 2018).

Üniversite seviyesinde uzaktan eğitimdeki terk oranları ise geleneksel üniversitelerdekinden çok daha yüksektir. Terk, mesleki, akademik, sağlık, ailevi ve kişisel nedenlerden kaynaklanabilir. Uzaktan eğitim kurumları, öğrencileri ile aralarındaki mesafe ne olursa olsun, onları sürekli olarak desteklemek ve eğitimde tutmak zorundadır (Kotsiantis, Pierrakeas & Pintelas, 2003). Bu nedenle, okulu terk olasılığı yüksek olan öğrencileri (okulu bırakmaya eğilimli öğrenciler) öngörebilen yöntemler bu konuda önlem alınması ve dolayısıyla öğrencilerin okul terkini azaltması için önemlidir.

Okul terkinin, kişinin eğitimine bireysel olarak olumsuz etkisinin yanında kurumsal ve finansal boyutu da önemli bir araştırma sebebidir. Kurumlar ilk kayıt sayılarına göre yatırım ve eğitim geliştirme planları yaparken pasif konumda olan veya terk eden öğrenci sayılarının fazlalığı özellikle uzaktan öğrenmede eğitimi sekteye uğratacak riskler barındırmakta ve yükseköğretimde mali dengeleri bozmaktadır.

2012-2015 yılları arasında Anadolu Üniversitesi Açıköğretim Sistemindeki terk etme nedenlerini araştıran Aguş (2019) yüz yüze görüşme ve çevrimiçi anket uygulamasıyla yürüttüğü çalışmasında, öğrencilerin terk nedenlerini kişisel, sistemle ilgili ve çevresel olarak üç kategoriye ayırmıştır. Çalışmada okulu bırakma nedenlerinin ve bu nedenlerin demografik gruplar arasındaki farklılıklarının ortaya çıkarılması amaçlanmış; bu nedenlerin okula başlama amacı, cinsiyet, yaş, eğitim düzeyi, medeni durum, program türü, çalışma durumu ve ikamet yerine göre de farklılaştığı gösterilmiştir. Bunun yanında Türkiye'de açıköğretim öğrencilerinin terk durumlarını büyük veri grubu üzerinde tahmin etmeye yönelik bir çalışmaya rastlanmamıştır.

Tüm bunlardan yola çıkarak bu araştırmanın amacı açık ve uzaktan eğitimde okuyan öğrencilerin sistemler üzerinden elde edilebilecek verileri ile okul terki ve akademik başarı

durumlarını tahmin etmek ve sınıflandırma performansı yüksek veri madenciliği modellerine göre akademik başarı ve okul terki ile ilişkili olan değişkenleri ortaya çıkartmaktır. Ayrıca kurulan veri madenciliği modelleri ile riskli öğrencileri önceden belirleyen bir erken uyarı modeli geliştirerek öğrenenlerin, ailelerin, eğitim kurumlarının ve yükseköğretim yöneticilerinin uzun vadeli okul terkini önlemeye yönelik planlarına katkı sunmak amaçlanmaktadır.

1.3. Sayıltılar

Araştırmada kullanılacak veriler İÜ AUZEF öğrenci otomasyon sisteminden elde edilmiştir. Bu sebeple veri tabanına kayıtlı öğrenci verilerinin doğru bilgiler içerdiği varsayılmıştır.

Bu çalışmada öğrenci devam statülerinde kayıt dondurmuş, kayıt sildirmiş, kayıt yenilememiş öğrencilerin tamamı devam etmeyen öğrenci statüsünde değerlendirilmiştir. Açık öğretimde okuldan atılma, öğrenim süresi ve öğrencilik hakları, kayıt dondurma, kayıt sildirme, süresiz devamsızlık, öğrenim statüleri ve mevzuata göre tanımlanmakta ve bu tanımlara çalışma kapsamında riayet edilmektedir.

Üniversiteye giriş sınavı, YÖS (Yurtdışından veya Yabancı Uyruklu Öğrenci Sınavı), Kurumlar arası Geçiş ya da Yatay Geçiş ile yerleşen öğrenciler sınav ile yerleşmiş olarak varsayılmış ve sınavla yerleşenler kategorisi altında birleştirilerek incelenmiştir.

1.4. Sınırlılıklar

Türkiye'de 2010 yılı itibarıyla açıköğretim programlarına öğrenci kabul eden üç devlet üniversitesi bulunmakla birlikte araştırma İÜ AUZEF'in 2010-2021 yılları arasındaki öğrenci verileriyle sınırlıdır. Ayrıca eğitimde veri madenciliği alanında pek çok algoritma kullanılmakta iken bu araştırma, sınıflandırma algoritmaları ile sınırlandırılmıştır.

2011 yılında yapılan bir mevzuat değişimi, daha önceki yılların aksine kayıt yenilemeyen öğrencinin okuldan atılmasının önüne geçmiş dolayısı ile tekrar mevzuatın değiştiği 2020 yılına kadar öğrenciler kayıt yenilemeden pasif olarak sistem içinde kalmışlardır. Öğrencilerin gerçekten okulu bıraktıkları ya da ara verdiklerinden emin olunamadığı için bu yeni durum okulu bırakma oranları üzerine yapılan araştırmaları ve sonuçların farklı ülkelerde bulunan üniversitelerdeki araştırma sonuçları ile karşılaştırılmasını oldukça zor kılmaktadır. Bu

araştırmanın sınırlılıklarından birisi de budur ve araştırmada yer alan veriler her iki mevzuat değişiminden de etkilenmiştir.

Eğitimde veri madenciliği yöntemlerinden destek vektör makinesi (DVM) algoritmasının hem okul terki hem de akademik başarının sınıflandırılmasında kullanılması planlanmıştı ancak tüm veri seti ile yapılan okul terki modelinde donanım yetersizliği (172 saat aralıksız çalışmasına rağmen bilgisayarın CPU kaynağının yetersizliği ile işlemin tamamlanamaması) sebebi ile model oluşturulamamıştır. Bu da araştırmanın bir diğer sınırlılığıdır.

1.5. Tanımlar

Açık Öğretim: Simonson'a (2003) göre öğrenme gruplarının birbirinden ayrıldığı; öğrenen, öğretici ve öğrenme materyallerinin etkileşimli iletişim ortamında bir araya geldiği; kurumsal resmi eğitim sistemidir.

Uzaktan Eğitim: YÖK (2016) mevzuatına göre yükseköğretim kurumlarında öğretim faaliyetlerinin bizzat öğretim elemanı tarafından aynı mekânda bulunma zorunluluğu olmaksızın eşzamanlı biçimde verildiği öğretim türü olarak tanımlanmıştır.

Uzaktan eğitim veya açık öğretim kavramları zaman içinde, teknolojik gelişmeler ışığında, farklılaşmaktadır (M. Tosun, 2016). Öğretici ile öğrenenin birbirinden ayrı mekânlarda olduğu eğitim süreci olarak tanımlanan uzaktan öğretim kavramı yerine, aynı anlama gelecek şekilde çevrimiçi (online) öğrenme, çevrimiçi eğitim, e-öğrenme, uzaktan eğitim, açık öğretim ve açık öğrenme gibi farklı kavramlar kullanılmaktadır (Bozkurt, 2015). Bu çalışmada kavramların tarihsel gelişimine uygun olarak Türkiye mevzuatında yer alan tanımlardaki teferruata da dikkat edilerek açık ve uzaktan öğrenme kavramları çoğu kez birbirinin yerine kullanılmıştır.

Veri Madenciliği: Veri madenciliği büyük hacimli verilerin işlenmesi için geliştirilmiş algoritmalar ile geleneksel veri analiz yöntemlerinin karması olan bir teknolojidir (Taşdemir, 2012). Veri madenciliği modelleri, gördükleri işlevlere göre; 1. Sınıflama ve Regresyon, 2. Kümeleme, 3. Birliktelik Kuralları olmak üzere temel olarak üç grupta toplanır (Savaş, Topaloğlu & Yılmaz, 2012). Bunlardan sınıflama ve regresyon modelleri tahmin edici; kümeleme ve birliktelik kuralları modelleri ise tanımlayıcı modeller olarak literatürde ayrılmaktadır.

BÖLÜM II

KAVRAMSAL ÇERÇEVE

Bu bölümde açık ve uzaktan öğrenmede okul terkinin kuramsal yapısı ve bunun üzerine yapılan araştırmalar ile bunlarla ilgili eğitimde veri madenciliği metodolojisine dayanan çalışmalara yer verilecektir.

2.1. Açıköğretimde Okul Terki ve Devamlılık

Eğitimde okul terki ve devamlılık üzerine (ki bu iki kavramı zıt olgular olarak ele alan veya koşut olarak tanımlayan çalışmalar mevcut) pek çok araştırma yapılmıştır. Özellikle yükseköğretimde yapılan okul terki çalışmaları kuramsal çerçeve (Bean ve Metzner,1985; Berge ve Huang, 2009; Kember, 1995; Park, 2007; Rovai, 2003; Tinto, 1975), olgusal tanım (Grau-Valldosera, 2019; Hagedorn, 2005; Lee ve Choi, 2011; Pinheiro vd., 2023) ile terk veya devam kararında etkili faktörler (Cheung ve Kan, 2002; Lottering vd., 2020; Moore vd., 2003) üzerinde toplanmaktadır. Son yıllarda ise özellikle terk etme eğilimi/ihtimali olan öğrencilerin kestirimine yönelik (Kotsiantis vd., 2003) çalışmalar bulunmaktadır.

Konuyu kuramsal olarak ele alan Tinto (1975) okul terki ve devam üzerine yaptığı çalışmaların sentezi sonucu öğrenci entegrasyon modelini oluşturmuştur. Model, geleneksel kurumlar için tasarlanmış olmasına rağmen, okulu bırakma kararlarının ardında yatan durumları ortaya koymayı amaçlayan sonraki çevrimiçi öğrenme terki modelleri için de bir referans olmuştur (Grau-Valldosera, 2019). Tinto'ya göre okul terki kararı, temel olarak hem akademik hem de sosyal düzeyde öğrenci entegrasyonuyla ilgili unsurların etkisiyle oluşmaktadır. Bağrıaçık ve Karataş'da (2022) açık ve uzaktan öğrenme terk modellerinin genellikle yüz yüze eğitim modellerine dayandığına işaret etmektedir. Bu bağlamda, Tinto'nun geleneksel yüz yüze

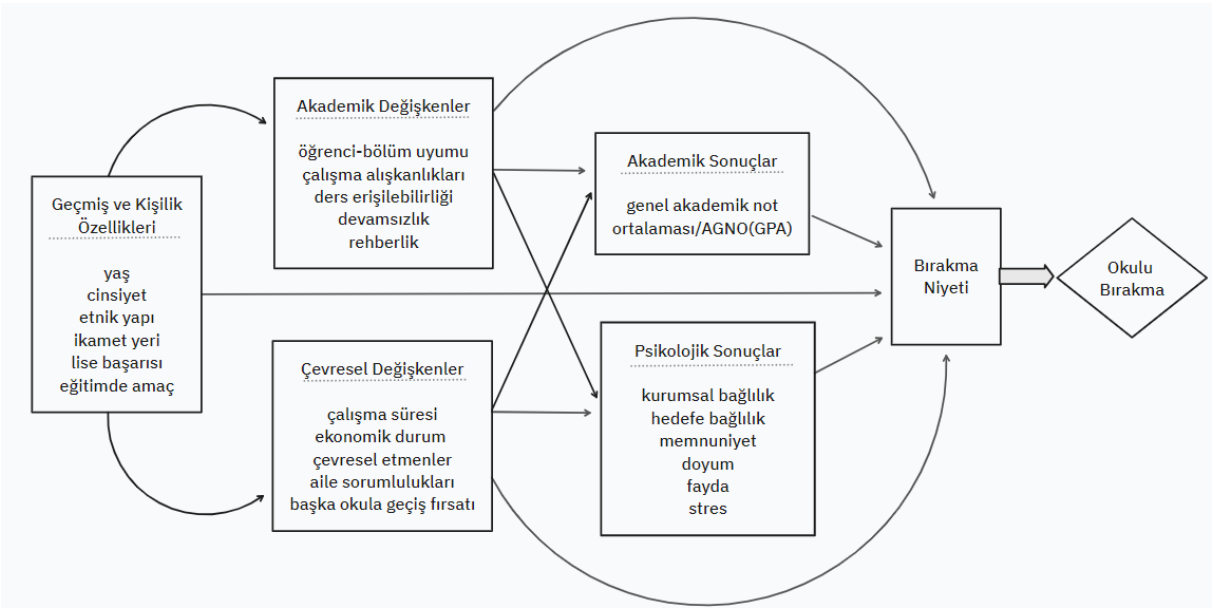
üniversitelere dayanan modeli, okul terki faktörlerini incelemek için yaygın olarak kullanılmaktadır. Burdan hareketle açık ve uzaktan öğrenmede okulu bırakma ve devam kararı ile ilgili modeller özetle 7 başlık altında toplanabilir: (Aguş, 2019; Grau-Valldosera, 2019; Yılmaz, 2020)

1. Tinto'nun Öğrenci Entegrasyonu Modeli
2. Astin'in Katılım Kuramı
3. Geleneksel Olmayan Öğrencilerde Bean ve Metzner'in Okulu Bırakma Modeli
4. Berge ve Huang'ın Kavramsal Okul Terki Modeli
5. Kember'ın Açık ve Uzaktan Eğitimde Okulu Bırakma Modeli
6. Rovai'nin Bileşik Devamlılık Modeli
7. Uzaktan Eğitimi Bırakma Üzerine Park'ın Yenilenmiş Bileşik Devamlılık Modeli

Tinto (1975), yükseköğretim terkini, öğrencinin kurumla etkileşimi sonucu ortaya çıkan zaman içinde bir süreç olarak anlamaktadır. Buna göre 1975'te olmasına rağmen okul terkini açıklamak için öğrenci geçmişi, motivasyon ve beklentiler, finansal bağlılık gibi çok çeşitli boyutlar önermiştir. Tinto'ya göre okulu bırakma kararı; öğrencinin giriş öncesi özellikleri, hedefleri, kurum ile ilişkileri ve deneyimi, akademik ve sosyal uyumunun bir sonucudur. Öğrenci akademik ve sosyal olarak kuruma ne kadar entegre olursa, kalma olasılığı da o kadar artar. Bu da terk veya kalıcılığın bütüncül ve uzun vadeli vizyon ile ele alınmasını gerektirir (Grau-Valldosera, 2019). Buna ek olarak, Tinto kalıcılığı okulu bırakma analizinden ayırarak, bir şekilde aynı madalyonun iki yüzü olsalar da, bunun %100 karşıt bir ilişki olmayabileceğini ve ayrı ayrı analiz edilebilecekleri fikrini ileri sürmektedir. Benzer şekilde kuramsal modellerin bir kısmı okul terkini, diğer bir kısmı da devam veya kararlılık olgularını öncelikle ele almışlardır.

Astin'in (1984) Katılım kuramına göre öğrenci akademik faaliyetlere ne kadar zaman ayırır ve dahil olursa derste veya kursta kalma olasılığı o kadar artar. Astin'e göre öğrenci katılımı öğrencinin akademik deneyime ayırdığı fiziksel ve psikomantıksal enerji miktarını ifade eder. Astin'in tanımı açık ve uzaktan eğitim bağlamında ele alınırsa, açık ve uzaktan eğitim öğrencisinin, kayıtlı olduğu programla ilgili WhatsApp, Facebook gibi sosyal medya hesaplarında vakit geçirmesi, canlı derslere katılması veya canlı ders kayıtlarını sonradan izlemesi, kurum tarafından düzenlenen etkinliklere dahil olması onun eğitim-öğretim sürecine katılım sağladığının göstergesi sayılabilir (Yılmaz, 2020). Buna göre etkileşimin takibi ve attırılması öğrenci tutmayı hedefleyen kurumlar için önemli olmalıdır.

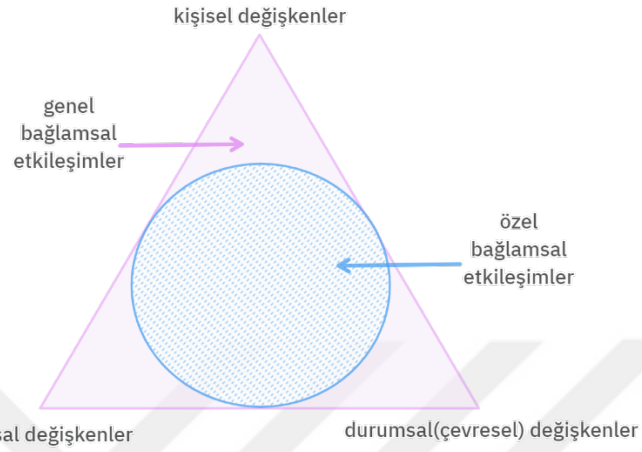
Bean ve Metzner'in (1985) okulu bırakma modeli, okul terkini öğrencinin sosyal ilişkileri ve etkileşimi temelinde ele alan Tinto ve Astin'nin aksine geleneksel eğitimin içinde olmayan öğrencilerin okul içi etkileşiminin okul dışı etkileşiminden daha az olduğundan yola çıkarak uzaktan öğrenenlerin niteliklerinin çeşitliğine vurgu yapmış ve temelde 3 kategori altındaki değişkenlerin önemini belirtmiştir. Bunlar; akademik değişkenler, öğrenci geçmişi ve kişilik özellikleri ile çevresel değişkenler ile bunların sonucundaki akademik ve psikolojik sonuçlardır (Şekil 1). Modelde öğrencinin giriş özelliklerine (yaş, cinsiyet, etnik yapı, ikamet vb.) yer verilse de J. P. Bean (1990) esasında geleneksel olmayan lisans öğrencilerinin yıpranmasında çevresel faktörlerin etkisine dikkat çekmektedir.



Şekil 1. Bean ve Metzner'in okulu bırakma modeli

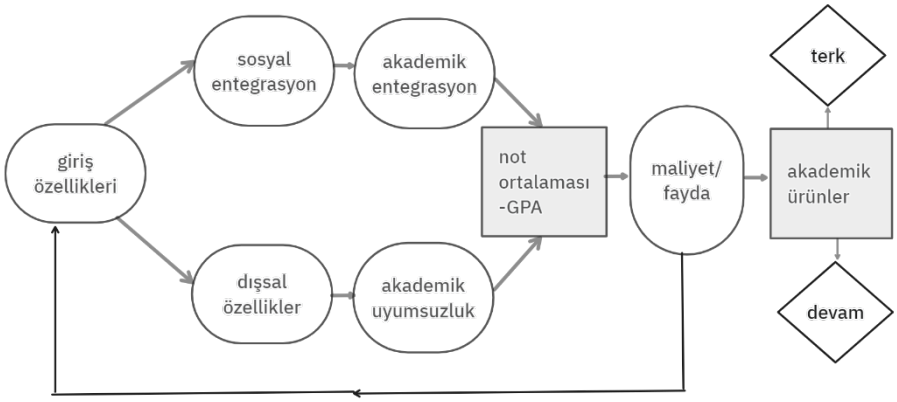
Öte yandan bırakma veya devam kararının karmaşıklığı ve çok yönlülüğü; konuyu nedensel bir modelden ziyade kavramsal bir model olarak ele alan Berge ve Huang (2004) Kavramsal Okul Terki Modeli'ni ortaya çıkarmıştır. Şekil 2'de gösterilen model üçgeninin üç ana alanı, kişisel değişkenleri, kurumsal değişkenleri ve durumsal (çevresel) değişkenleri ve bunlar arasındaki etkileşimleri içeren bağlamı temsil etmektedir. Kişisel değişkenler, öğrencilerin demografik özellikleri gibi bir dizi öğrenci faktörünü kapsamaktadır. Modellerinin kapsayıcı ve bağlama duyarlı olarak kalıcılığı artırmaya yönelik ikili bir yaklaşım sergilediğini ifade eden Berge ve Huang (2004) bu açıdan yenilikçi bir yaklaşım sunduklarını ifade etmektedir. Buna göre devam etme ya da okulu bırakma konusunda gönüllü bir karar verildiğinde, bu karar öğrencinin kişisel

koşullarından etkilenerek kendisi tarafından verilir. Sonuç, eğitim bağlamında ise algılanan fırsat, uygunluk, stres, sorumluluk ve memnuniyetten kaynaklananlar ile tüm faktörlerin öğrencinin sürekli maliyet/fayda analizine dayanır. Bunun için Berge ve Huang bir bakıma, her bağlam için "uyarlanabilir" bir model önermektedir (Grau-Valldosera, 2019).



Şekil 2. Berge ve Huang'ın kavramsal okul terki modeli. (Grau-Valldosera, J. (2019). *A dropout definition for continuance intention and effective re-enrolment models in online distance learning.* Doktora Tezi, Universitat Oberta de Catalunya, Katolanya.

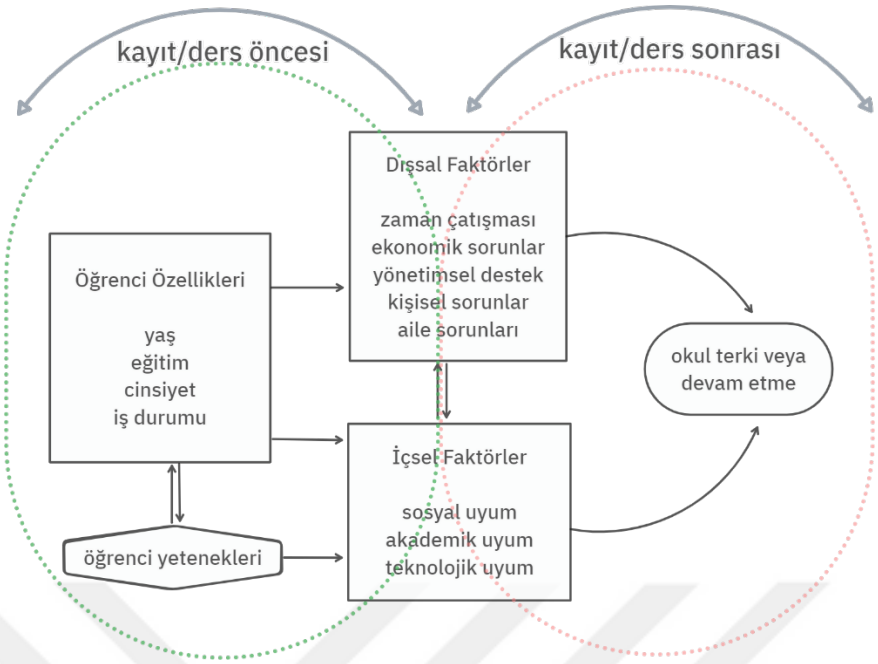
Alanda başka bir araştırmacı olan Kember'in (1995) uzaktan eğitimde okul terki modeli, çevrimiçi eğitim kurslarında kalıcılığı anlamak için faydalı bir çerçeve sunmaktadır. Kember'in modeline göre, öğrenciler iki farklı öğrenme yolundan geçmektedir: (1) sosyal entegrasyondan akademik entegrasyona ve (2) dışsal atıflardan akademik uyumsuzluğa (bkz. Şekil 3). Yetişkin bir öğrenci bu iki yoldan birini seçer ve sonunda öğrenmenin sonucu olan not ortalamasıyla yüzleşir. Öğrenmeye devam etmenin maliyeti ile not ortalamasını zihninde tartan ve karşılaştıran öğrenciler okulu bırakma ya da devam etme kararını kendileri verirler. Kember'in modeli, Tinto'nun geleneksel öğrencilerle yaptığı ve yaygın olarak atıfta bulunulan öğrenci entegrasyon modeline (Tinto, 1975) dayanmaktadır. Uzaktan eğitim alan yetişkin öğrencilere odaklanan Kember başarılı öğrencilerin hem sosyal (yani aile, iş vb. ile) hem de akademik (eğitim kurumuyla) olarak entegre olabilen öğrenciler olduğunu öne sürmüştür. Tüm bunların yanında önceki modellerde de görüleceği üzere bireysel özellikler, demografik değişkenler birey için alabileceği devam ya da terk kararının bileşenlerinden biridir. Okur vd. (2019) ise öğrenci değerlendirme modeli olarak ele aldıkları bu yaklaşımı demografik özellikleri, öğrenci motivasyonu, akademik yetenekler ve sosyal faktörler gibi bileşenler ile açıklamıştır.



Şekil 3. Kember'in uzaktan eğitimde okulu bırakma modeli

Rovai (2003) de "(devam) kararının karmaşıklığına" atıfta bulunarak öğrencinin devamlılığını sağlayan basit bir formülün olmadığını, çevrimiçi bir programda yetişkinlerin devamlılığının birden fazla soruna karşın karmaşık bir yanıt olduğunu söylemektedir. Buna göre "Öğrenci kaybını tek bir öğrenci, kurs veya okul özelliğine bağlamak inandırıcı değildir". Rovai (2003), program düzeyinde devam etme kararına dahil olan faktörleri gösteren bir "bileşik model" ortaya koymaktadır. Modeli oluşturan ana parçalar arasında temel referans olarak Tinto (1975, 1982) ve Bean ve Metzner'in (1985) geleneksel olmayan lisans öğrenci yıpranması üzerine yaptığı kavramsallaştırma yer alır. Ayrıca, Rovai'nin modeli çevrimiçi öğrencilerin ihtiyaç duyduğu beceriler, uzaktan eğitim öğrencilerinin benzersiz ihtiyaçları ve öğrenme ve öğretme stillerinin uyumlaştırılması gerekliliği hakkında bilgiler içermektedir. Model eğitim devamlılığını giriş öncesi ve giriş sonrası bölümlerden oluşan iki ana bölümde incelemektedir: bir tarafta, kabulden önce öğrenci özellikleri ve becerileri, diğer tarafta ise kabulden sonra öğrencileri etkileyen dış ve iç faktörler.

Rovai'nin (2003) birleşik devamlılık modelinden yola çıkan Park (2007) ise modeli Şekil 5'teki gibi yeniden şematize etmiştir. Park (2007) öğrenci özelliklerini, yeteneklerini, dışsal ve içsel faktörleri okula devam etme veya terk için etkili değişenler olarak almış; aynı zamanda bu özellikleri kurs öncesinde ve kursa kayıt olduktan sonra/ders esnasında olarak iki aşamada değerlendirmiştir.



Şekil 4. Park'ın yenilenmiş bileşik devamlılık modeli

Rovai, dışsal faktörleri tamamen kayıt sonrası evrede ele alırken, Park, bazı dışsal faktörlerin kayıt öncesinde de öğrenci kararını etkilediğini ifade ederek bu bileşeni kayıt öncesi ve sonrası evrelerin ortasına almıştır. Ayrıca modelde, fazla deneysel destek bulunamadığı için "öğrenci becerileri" kutusu farklı şekilde gösterilmiştir (Yılmaz, 2020). Bu modelde, içsel ve dışsal faktörler arasında bir etkileşim olduğu da vurgulanmıştır. Şöyle ki iş yükü çok yüksek ve çalışmak için çok az zamanı olan bir öğrenci, öğretim elemanı ile zor iletişime geçiyor ve dönüt alamıyorsa, öğretim elemanı ile kolay iletişim kurabilen bir öğrenciye göre bırakma ihtimali daha yüksek olacaktır (Park, 2007). Park, her ne kadar deneysel kanıtlar bulunmadığı için Rovai'nin modelinin sadeleştirilmesi gerektiğini ifade etse de atıf sayılarından yola çıkarak alan yazında Rovai'nin modelinin daha çok kabul gördüğü söylenebilir.

2.1.1. Okul Terki

Okul terki veya devam durumunu sosyolojik, ekonomik ve diğer pek çok boyutu ile ele alarak oluşturulan modellerin yanında özellikle son dönemde uzaktan eğitimde terk kavramının tam olarak neyi ifade ettiğine odaklı, daha çok fonksiyonel tanımını yapmaya çalışan araştırmalar da artmıştır. Alan çalışmaları incelendiğinde okul terki bir dizi ilgili kavram olan terk, ayrılma,

yıpranma, geri çekilme, tamamlanmama (dropout, attrition, withdrawal, non-completion) ile ve zıt anlamlı olarak kalıcılık, sebat, devamlılık, tamamlama ve başarı (retention, persistence, continuance, completion, ve success) kavramları ile birlikte kullanılmaktadır (Xavier & Meneses, 2020). Okul terki, genel olarak öğrencinin birbirini takip eden belirli sayıda dönem boyunca kayıt yaptırmaması olarak tanımlanabilse de literatürde okul terkinin zamansal bir anlayışa bağlı olarak birçok farklı tanımı vardır ve konu tartışmalıdır (Grau-Valldosera & Minguillón, 2014). Ayrıca okul terki ve akademik kalıcılık kavramlarının bir madalyonun iki yüzü gibi ele alındığı görülmektedir. Akademik kalıcılık, öğrencinin üniversitenin eğitim sürecine sonuna kadar sürekli katılımı olarak tanımlanmaktadır (Thomas & Unger, 2015). Kalıcılık aynı zamanda bir öğrencinin üniversiteye kayıtlı kaldığı süredir ve üniversite başarısının ön koşulu yani yeterli olmasa bile gerekli koşulu olarak kabul edilebilir. Öğrenciler, amaçlarına ulaşmadan üniversiteden ayrıldıklarında, okulu bırakan öğrenciler olarak etiketlenirler (Thomas & Unger, 2015). Üniversite terki; "bir üniversite diploma programından tamamlanmadan çekilme" olarak da tanımlanmaktadır (Agrusti, Bonavolonta & Mezzini, 2019).

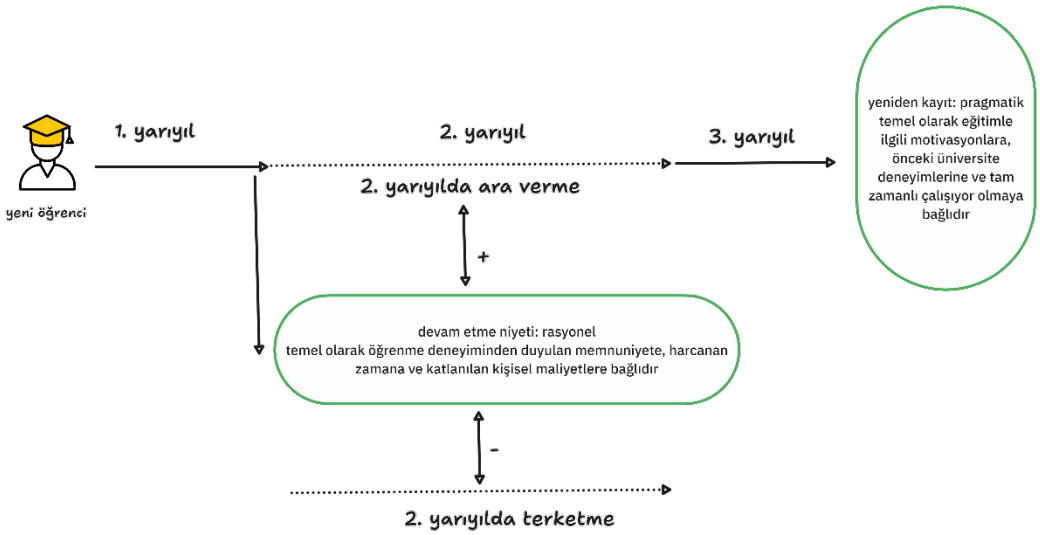
Grau-Valldosera (2019) Katolanya Açık Üniversitesinde yürüttüğü araştırmada öğrencilerin programlara devam etme durumlarına ve daha spesifik olarak, uzaktan öğrenme programlarında ikinci yarıyılda ara veren öğrencilerin yeniden kayıtları ile e-öğrenmede okul terkinin tanımlanmasına odaklanmaktadır. Çalışmada;

1. Kayıt davranışına ilişkin kanıtları ve e-öğrenme öğrencilerinin "bağlamını" (yani ara verme) hesaba katarak okul terki nasıl tanımlanabilir?
2. Bir sonraki dönemde ve aynı derece veya programa yeniden kaydolma niyetinin arkasında hangi değişkenler veya etkenler vardır?
3. Yeniden kaydolma veya arayı uzatma konusundaki nihai kararın arkasında hangi değişkenler veya etkenler vardır?
4. Devam etme niyeti ve etkili yeniden kayıt için tespit ettiğimiz farklılıklar ve benzerlikler nelerdir?

sorularına yanıt aranmış; e-öğrenmede okul terki, devam etme niyeti ve yeniden kayıt kavramlarına yeni bir tanımı ve model önerisi getirilmiştir.

Okul terkinin bahsedilen karmaşıklığının üstesinden gelmenin en iyi yolu olarak tanımını kanıtlara, özellikle de öğrencilerin kayıt ve ayrılma davranışlarının nesnel bir analizine dayandıran Grau-Valldosera (2019) uzaktan eğitimde öğrencinin ara verme sayısına (N) göre

her program için bir terk olasılığı hesaplanabileceğini ortaya koymuştur. Grau-Valldosera (2019) geliştirdiği hesaplama yönteminin özellikle öğrencilerin daha fazla iş ve aile zamanı kısıtlamalarına sahip olduğu ve bu nedenle daha yüksek bir ara verme olasılığı ile daha üretken bir kayıt davranışı sergilediği uzaktan çevrimiçi eğitim kurumları için geçerli olduğunu iddia etmektedir.



Şekil 5. Birinci dönemden üçüncü döneme kadar yeni öğrenciler-akış şeması

Grau-Valldosera (2019) ara verme durumunu uzaktan eğitimde öğrenci ihtiyacı olarak ele almış (Şekil 5) ve bunun tam bir okul terki tanımı olamayacağını iddia etmiş ancak Aulck vd. (2017) birinci veya ikinci dönem sonunda ara veren öğrencilerin yaklaşık %46'sının tekrar okula dönmediğini belirtmektedir. Dolayısı ile ara verme davranışının bazı durumlarda okul terkinin ön bir tanımı olabileceği göz önüne alınmalıdır.

Bir öğrencinin özellikle açık ve uzaktan öğrenmede öğrenmeyi ne zaman bırakmış sayılacağı araştırmacılar tarafından farklı şekilde ele alınmış; hala açık alanlardan birisidir. Pek çok alan uzmanının araştırmalarında tam olarak neyi terk olarak ele aldıklarını tanımlamaları gerekebilir. Lee ve Choi (2011) terkin tanımı ile ilgi karmaşayı ele aldıkları araştırmalarında terk tanımlarının literatürde Tablo 1'deki gibi belirlendiğini ifade etmişlerdir.

Tablo 1
Okul Terki ile İlgili Tanımlar

Yazar	Yıl	Terk Tanımı
Castles	2004	Terk: Resmen ayrılan, üniversiteyi bilgilendirmeden ayrılan veya bir dönem içinde bir kursu tamamlamayan öğrenciler
Cheung ve Kan	2002	Terk: Başarısız olan ya da yeniden sınava girme hakkı kazanan öğrenciler
Dupin-Bryant	2004	Terk: Bir dönem içinde bir kursu tamamlamayan öğrenci
Finnegan vd.	2009	Çekilme: (1) Resmi olarak kurstan çekilenler; (2) başarılı tamamlayanlar - A, B veya C notu alanlar; (3) başarısız tamamlayanlar - D veya F notu alanlar veya eksik not alanlar
Frydenberg	2007	Terk: Sınıf başlamadan önce kayıt olan ancak sınıf başlamadan, yönlendirme haftası sırasında veya sonrasında bırakan öğrenciler
Ivankova ve Stick	2007	Terk: Programdan çekilen veya programdan çıkarılan öğrenciler
Kemp	2002	Tamamlanamama: Kursa başlamayan, kurstan çekilen veya akademik olarak başarısız olan öğrenciler
Levy	2007	Terk: e-öğrenmeden mali cezalar alarak gönüllü olarak ayrılan öğrenciler
Moore vd.	2003	Tamamlanamama: F notu alan veya resmen kurstan çekilen öğrenciler
Morgan ve Tam	1999	Tamamlanamama: Takip eden dönemde kayıt yaptırmayan öğrenciler
Morris vd.	2005	Çekilme: Resmi çekilme işlemi tamamlanmış öğrenciler. Başarısız tamamlayanlar: D, F veya eksik not alan öğrenciler
Perry	2008	Çekilme: merkezi çekilme (öğrenci yılda iki kursu tamamlama program gereksinimini yerine getiremez); akademik çekilme (programda iki kursu başarısız olan öğrenciler) ve öğrenci çekilmesi (merkezi veya akademik gereksinimlerle ilgili olmayan nedenlerle ayrılan öğrenciler)
Pierrakeas vd.	2004	Terk: En az bir modüle kaydolan ancak bir proje teslim etmeyen; bazı ya da tüm görevlerini tamamlamayan ancak çalışmalarına devam edeceğini belirten; gelecekte yeniden kayıt yaptırmayacaklarını belirten öğrenciler
Pigliapoco ve Bogliolo	2008	Terk: İlk yılın sonunda kaydını yenilemeyen öğrenciler
Shin ve Kim	1999	Terk: Üç ardışık dönem kayıt yaptırmayan öğrenciler
Tello	2007	Devam etmeme: Son not verme döneminden önce Kayıt Ofisi'ne bir kursu bıraktıklarını bildiren öğrenciler
Willging ve Johnson	2004	Terk: Derece programına başladıktan sonra ilk kurslarından sonra programdan ayrılan öğrenciler

Kaynak: Lee, Y., & Choi, J. (2011). A review of online course dropout research: implications for practice and future research. *Educational Technology Research and Development*, *59*(5), 593-618. https://doi.org/10.1007/s11423-010-9177-y

Lee ve Choi'nin araştırmasından yola çıkarak oluşturulan okul terki tanımlarından;

- ilk dönemin/yılın sonunda kaydını yenilemeyen öğrenciler (Pigliapoco ve Bogliolo, 2008)
- devam eden dönemde kayıt yaptırmayan öğrenciler (Morgan ve Tam, 1999)
- resmen kurstan çekilme kararı alan ve kuruma bildiren öğrenciler (Castles, 2004)

gibi tanımlar sık kullanılmıştır ve bu araştırmaya da bunlar rehberlik etmektedir. Bununla birlikte terkin kullanılan tanımları arasında "F" notu alma, sınavda başarısız olmak gibi akademik başarısız olmakla ilgili yaygın bir durumun olduğu da görülmektedir (Cheung ve Kan, 2002; Finnegan vd., 2009; Moore vd., 2003; Morris vd., 2005).

Elibol ve Bozkurt (2023) ise şimdiye kadar okulu bırakan öğrencilerin, aşağıdakileri başlıklar altında karakterize edildiğini aktarmaktadır:

- Herhangi bir nedenle kurumlarından ayrılanlar
- Öğrenim ücreti ödemeleri tamamlandıktan ve ders bırakma/ekleme dönemi sona erdikten sonra bölümlerinden kendi istekleriyle ayrılanlar
- Üç dönem üst üste kayıt yaptırmamayı takiben kayıt yaptırmamış olanlar
- F notu almış veya kurstan resmen çekilmiş olanlar
- En az bir modüle kaydolmuş ancak tek bir proje sunmamış olanlar
- Bir dönem boyunca bir dersi dahi tamamlayamayanlar
- Resmi çekilme prosedürünü uygulayarak ayrılanlar
- Mali cezalara maruz kalarak e-öğrenmeden çekilmeyi tercih edenler
- Programın tamamından istekli çekilenler veya çıkarılanlar
- Programın yılda iki ders tamamlama şartını (veya kuruma özgü diğer özel şartlarını) karşılayamayanlar

Okul terkini nihai bir sonuç değil de süreç ve karar verme aşamaları olarak ele alan çalışmalar da mevcuttur. Buna göre bırakma niyeti okul terkinin olasılığı ve erken bir işareti olarak görülmektedir. Véliz Palomino ve Ortega'ya (2023) göre ise üniversite terki farklı aşamalardan oluşan bir karar verme süreci olarak ele alınabilir: uygunsuzluk algısı, bırakma/değiştirme düşüncesi, düşünme, bilgi arama ve nihai karar gibi süreçte bu olguyu etkileyen çok sayıda faktör vardır.

Okul terkine etki eden faktörleri ortaya koyan çalışmalar da mevcuttur. Lottering vd. (2020) demografik faktörler, sosyo-ekonomik faktörler ve aile faktörleri gibi birçok faktörün

öğrencilerin okulu bırakmasıyla ilişkili olduğunu; okulu bırakmanın ise sistem ve sınav sonuçlarıyla ilgili hayal kırıklığı anlamına geldiğini aktarmaktadır.

Kabakchieva (2013) öğrenci performansını tahmin etmeye yönelik çalışmasında: (i) öğrenci geçmişi faktörleri (cinsiyet, doğum yılı, doğum yeri ve ülkede yaşadığı yer), (ii) önceki akademik performans (önceki eğitim türü, profili ve önceki eğitim yeri ve önceki eğitimin toplam puanı) ve (iii) mevcut akademik faktörler (üniversiteye giriş yılı, giriş sınavı ve elde edilen puan, üniversite uzmanlığı/yönü, mevcut dönem ve toplam üniversite puanı) gibi bir veri kümesi faktörlerine yer vermiştir.

Çeşitli araştırmalar, öğrencilerin okulu bırakma oranının önemli faktörlerinden birinin, ortaokul notlarının yanı sıra üniversitede okunan bölüm olduğunu göstermektedir. Gerçekten de, okulu bırakma oranı mühendislik disiplinlerindeki öğrenciler arasında ve nispeten düşük düzeyde ön yeterliliklere sahip öğrenciler arasında daha yüksektir (Okewu & Daramola, 2017). Ayrıca okulu bırakma oranı daha yüksek olan öğrencilerin diğer fakültelere kıyasla Bilgisayar ve İletişim Teknolojileri Fakültesine kayıtlı öğrenciler olduğu bulunmuştur (Paura & Arhipova, 2014). Altıntaş ve ark. (2021) da yükseköğretimde akademik riskli öğrencilerin tahmini için yürüttükleri çalışmada fakülte ve programların etkili faktörlerden olduğunu ortaya koymuşlardır. Yukarda bahsi geçen çalışmalarda dikkat çeken bir diğer husus nitelikler, faktörler, özellikler ve değişkenler gibi kavramların birbirlerinin yerine kullanılmasıdır. Bu durum daha çok veri madenciliği çalışmalarında araştırmacıya sağlanan girdiği değişkenlerinin yapısındaki esnekliğe dayanır.

Yakın zamana kadarki çalışmaların çoğunda araştırmacılar okulu bırakmayı açıklamak için kullandıkları yaklaşımlarda, öğrencileri 'düzeltmeye' odaklanmaktadır. Sadece sektörün bazı kesimlerinde, üniversitelerin bugünün öğrencilerinin ve onların gelecekteki yaşamlarının ve dünyalarının ihtiyaçlarını karşılamak için kendilerinin de değişmesi gerekebileceğinin farkına varılmıştır (Marr vd., 2024). Buna bağlı olarak, 'okul terki' ve 'başarı' terimleri pek çok araştırmada kurumların, yasa koyucuların ve eğitim politikası yapıcılarının gereksinimleri nedeniyle tanımlanmaktadır. Oysaki açık ve uzaktan öğrenmenin doğasını da unutmadan öğrencileri tanımaya ve anlamaya yönelik çalışmalar ile kurum ve sistemleri iyileştirmeye yön verebilecek çalışmaların önemi artmaktadır.

2.2. Eğitimde Veri Madenciliği

Veri madenciliği büyük hacimli verilerin işlenmesi için geliştirilmiş algoritmalar ile geleneksel veri analiz yöntemlerinin karması olan bir metodolojidir. Dolayısı ile veri madenciliği büyük hacimlerde gözlenen verilerin analiz edilmesi diğer bir deyişle veriye en uygun hipotezlerin bulunması ile ilgilenir. Tahmin edici ve tanımlayıcı veri madenciliği görevlerinin başarılmasında istatistik disiplininden örnekleme, tahmin ve hipotez testlerinden faydalanırken yapay zekâ, makine öğrenmesi, örüntü tanımlama disiplinlerinden de arama algoritmaları, modelleme teknikleri ve öğrenme teorileri kullanılır (Taşdemir, 2012).

Veri Madenciliğinin eğitim bağlamında uygulanması ise Eğitim Veri Madenciliği (EVM) olarak adlandırılır ve Uluslararası Eğitim Veri Madenciliği Topluluğu tarafından "eğitim ortamlarından gelen benzersiz veri türlerini keşfetmek için yöntemler geliştirmek ve bu yöntemleri öğrencileri ve öğrendikleri ortamları daha iyi anlamak için kullanmakla ilgilenen gelişmekte olan bir disiplin" olarak tanımlanır (Guarín, Guzmán & González, 2015).

Eğitim alanında veri madenciliği çok farklı ve giderek artan şekilde uygulama ve teorik araştırmalarda kullanılmaktadır. Temel olarak bu araştırmalar öğrenci performansını artırmak, eğitim süreçlerini iyileştirmek ve öğrenme deneyimlerini kişiselleştirmek için veriden anlamlı bilgiler çıkarmayı amaçlar. Eğitimde veri madenciliği, geniş veri setlerini analiz ederek öğrenciler, öğretmenler ve eğitim yöneticileri için öngörüler sağlar. Eğitimde veri madenciliği uygulamaları:

- Öğrenci Performans Analizi: Öğrencilerin akademik performanslarını değerlendirmek ve potansiyel başarı veya başarısızlık faktörlerini belirlemek için kullanılır. Örneğin, sınav sonuçları, katılım, ödevler ve diğer etkinlikler üzerinden öğrencilerin performans trendleri analiz edilebilir.
- Eğilim Tespiti ve Öngörü: Geçmiş ve mevcut verilere dayanarak gelecekteki eğilimleri tahmin eder. Örneğin, öğrenci devamsızlık oranlarının zaman içindeki değişimini izleyerek, hangi dönemlerde veya koşullarda devamsızlığın artabileceğini öngörebilir.
- Düşük Performans Gösteren Öğrencileri Tespit Etme: Risk altındaki öğrencileri erken tespit etmek ve müdahalede bulunmak için veri madenciliği teknikleri kullanılır. Bu, öğrencilerin eğitim yolculuklarında gerekli destekleri almasını sağlar.
- Kişiselleştirilmiş Öğrenme Deneyimleri: Öğrencilerin öğrenme stilleri, ilgi alanları ve performans düzeyleri göz önünde bulundurularak kişiselleştirilmiş öğrenme yolları ve

materyalleri geliştirilir. Bu uygulama, öğrenme yönetim sistemleri üzerinden yürütülebilir.

- Kurumsal Raporlama ve Karar Destek Sistemleri: Eğitim kurumları, kaynak tahsisinden, ders programlarına, personel yönetimine kadar çeşitli karar alma süreçlerinde veri madenciliğinden faydalanır. Veri madenciliği, kurum içi verimliliği artırmak ve stratejik kararlar almak için kritik rol oynar.
- Etkileşim ve Katılım Analizi: Öğrenci ve öğretmen etkileşimlerini analiz ederek, öğrenme ortamının etkinliğini artıracak yöntemler geliştirilir. Bu, forum katılımları, çevrimiçi ders aktiviteleri ve diğer etkileşimli platformlar üzerinden yürütülebilir.
- Eğitimde Duygu/Durum Analizi: Öğrenci geri bildirimleri, forum mesajları ve sosyal medya yorumları gibi metin verileri analiz edilerek, öğrencilerin eğitim deneyimleri hakkındaki duygusal eğilimler tespit edilir.

Eğitsel veri madenciliği (EVM) ile ilgili 2006-2016 yılları arasında yapılan çalışmaları inceleyen Tekin ve Öztekin (2018) bu çalışmaların araştırma konularına göre dağılımını Tablo 2'deki gibi özetlemiştir.

Tablo 2
2006-2016 Yılları Arasında Konusuna Göre EVM Çalışmaları

Araştırma Konusu	f	%
Öğrenci Performansı	15	10,7
Öğrenme Yönetim Sistemleri	10	7,1
Okuldan Ayrılan Öğrenciler	9	6,4
Akademik Başarı/Öğrenme Süreci	32	22,9
Literatür Taraması (Veri madenciliği modelleri, Öğrenme analitikleri)	48	34,3
Diğer	26	18,6
Toplam	140	100

Kaynak: Tekin, A., & Öztekin, Z. (2018). Eğitsel veri madenciliği ile ilgili 2006-2016 yılları arasında yapılan çalışmaların incelenmesi. *Eğitim Teknolojisi Kuram ve Uygulama*, *8*(2), 108-124. https://doi.org/10.17943/etku.351473

Bu yönde hali hazırda yapılan veri madenciliği çalışmalarını derleyen başka bir araştırmacı Bonde ve Kirange (2018) ilgili 57 araştırmanın yöntem ve bulgularını özetleyen bir tablo sunmuştur. Bu tabloda yer alan çalışmalardan bazıları aşağıdaki gibidir:

Dekker (2009) kümeleme ve sınıflandırma yöntemleri ile okul terkinin tahmin edilmesini amaçlamış ve çalışmada maliyete duyarlı öğrenme algoritmaları ve yanlış sınıflandırma analizinin sonuçları ele alınmıştır. Asma, Nusari ve Al-Shargabi (2010) K-Mean, Apriori,

Karar Ağaçları yöntemlerini kullanarak yaptıkları çalışmada öğrenciler başarılı veya başarısız olarak ikili sınıflandırmaya odaklanmış; öğrencinin finansal davranışını kullanarak okulu bırakmasını öngörmüştür. Penchenizkiy ve Kabakchieva (2012) ise sınıflandırma yöntemlerini kullanıp öğrencilerin kişisel, üniversite öncesi ve üniversite verilerini analiz ederek öğrencileri başarılı veya başarısız olarak sınıflandırmak için bir model geliştirmiştir. Murray (2014) istatistiğe dayalı veri madenciliği yöntemleriyle yürüttüğü çalışmasında, gönüllü veya gönülsüz olarak okul terk eden iki tür öğrenciyi hedeflemiş ve bunların sosyoekonomik ve üniversite kaynaklı faktörlerini modeline dâhil etmiştir.

Albreiki vd. (2021) makine öğrenimi teknikleri kullanılarak öğrenci performans tahmini üzerine literatür taraması yapmış ve sonuç olarak araştırmaları EVM yöntemleri kullanılarak öğrenci terkinin tahmin edilmesi başlığı altında Tablo 3'teki gibi sınıflandırmıştır. Tablo 3'te araştırmalar amaç ve yaklaşımı, ele aldığı değişkenler, kullandıkları model ve algoritmalar ile bunların sıklığı başlıklarında sunulmuştur.

Tablo 3
EVM Yöntemleri Kullanılarak Öğrenci Terkinin Tahmin Edilmesi

Yaklaşım	Öznitelikler	Algoritmalar	Sıklık
Terkin tahmini için zamanı içeren öznitelikler	Öğrencilerin kişisel karakterleri ve akademik performansı	DT, LR, SVM, ARTMAP, RF, CART, ve NB	10
Müfredat bazlı ve öğrenci performansına dayalı öznitelikler	Öğrenci performansı, sınıf dengesizlik sorunları	K-NN, SMOTE	2
Devam oranı	Birinci sınıf öğrencileri	DT, Yapay Sinir Ağları (ANN)	2
Terk faktörleri	Faydalı zaman serisi modellerinin değerlendirilmesi (Gizli Markov Modeli (HMM))	LSTM ile birleştirilmiş RNN	3
Olası öğrenci terkininin erken aşamada tahmini	Üniversite öncesi giriş bilgileri, transkript bilgileri	ICRM2 ile SVM, NB, DT, ID3, DL, ve KNN, CART, ve Adaboosting Ağacı	4

EVM'yi "eğitim bağlamında ortaya çıkan verileri keşfetmek için yöntemlerin geliştirilmesiyle ilgilenen, gelişmekte olan disiplinler arası bir araştırma alanı" olarak tanımlayan Romero ve Ventura (2010) 1995 ile 2005 yılları arasında yayımlanmış ilgili araştırmaları inceleyerek bu alana artan eğilimi ortaya koymuştur. Romero ve Ventura çalışmasında eğitimde veri madenciliğini sadece kullanılan veri ve VM tekniklerinin türüne göre değil, aynı zamanda ve daha da önemlisi, çözdükleri eğitim görevinin türüne göre sınıflandırmıştır. EVM, e-öğrenme, uyarlanabilir hiperortam, akıllı ders sistemleri, web madenciliği, veri madenciliği vb. dâhil olmak üzere birçok köklü araştırma alanıyla ilgili yeni ve gelişmekte olan bir araştırma alanı olarak tanıtılmıştır. Daha sonra, eğitim ortamında veri madenciliği teknikleriyle çözülen en

tipik/yaygın görevleri listelemiş ve gelecekteki en umut verici araştırma alanlarından bazılarını tartışmıştır. Buradan yola çıkarak Romero ve Ventura EVM'nin artık ergenlik çağına yaklaştığını, yani artık ilk günlerinde olmadığını, ancak henüz olgun bir alan da olmadığını söyler.

2.2.1. Veri Madenciliği Süreçleri ve Algoritmalar

Neden veri madenciliği popüler hale gelmiştir? Bu sorunun cevabını alandaki araştırmacılar (Dutt, Ismail & Herawan, 2017; Nouri, Larsson & Saqr, 2019; Yulia, 2019) klasik istatistiksel yöntemlerin güncel ihtiyaçlara dinamik ve hızlık şekilde cevap verememesinde bulmuşlardır. Son dönemlerde özellikle akademik alanda belirgin bir şekilde iki problem ortaya çıkmıştır. Bunlardan biri, mevcut veri kümelerinin boyutu (hem birim sayısı hem de kaydedilen değişken sayısı), diğeri ise "kendiliğinden" oluşan büyük verilerin genellikle istatistiksel açıdan gerekli prosedürlere uygun davranılmadan toplanmış olmasıdır (Şevgin, 2020). Bu sorunlar karşısında klasik istatistiki yöntemler yetersiz kalmış ve zamanla bunların geliştirilmesi ile veri madenciliği metodolojileri ortaya çıkmıştır. Böylelikle büyük verilerin anlamlandırılması olarak ifade edilen veri madenciliği özellikle eğitim alanında öğrencilerin performanslarını iyileştirmek, eğitimcilerin eğitim verimliliğini artırmak ve düzenli aralıklarla öğrencilere uygun şekilde danışmanlık verebilmek için yararlı bir şekilde uygulayabilecekleri bilgiler sağlayacak teknikler geliştirmek amacıyla yaygın olarak kullanılmaya başlanmıştır (Shyamala, 2008).

Veri madenciliği, büyük veri kümelerinden yararlı bilgiler çıkarmak için yapılan istatistiksel analizler, yapay öğrenme ve desen tanıma tekniklerini içeren bir süreçtir. Bu süreç, veri tabanlarından elde edilen bilgilerin işlenmesi, anlamlandırılması ve örgütlenmesini kapsar. Veri madenciliği, genelde büyük veri setlerini analiz ederek önceden tanımlanmamış ilişkileri, desenleri ve bilgileri ortaya çıkarmayı amaçlar.

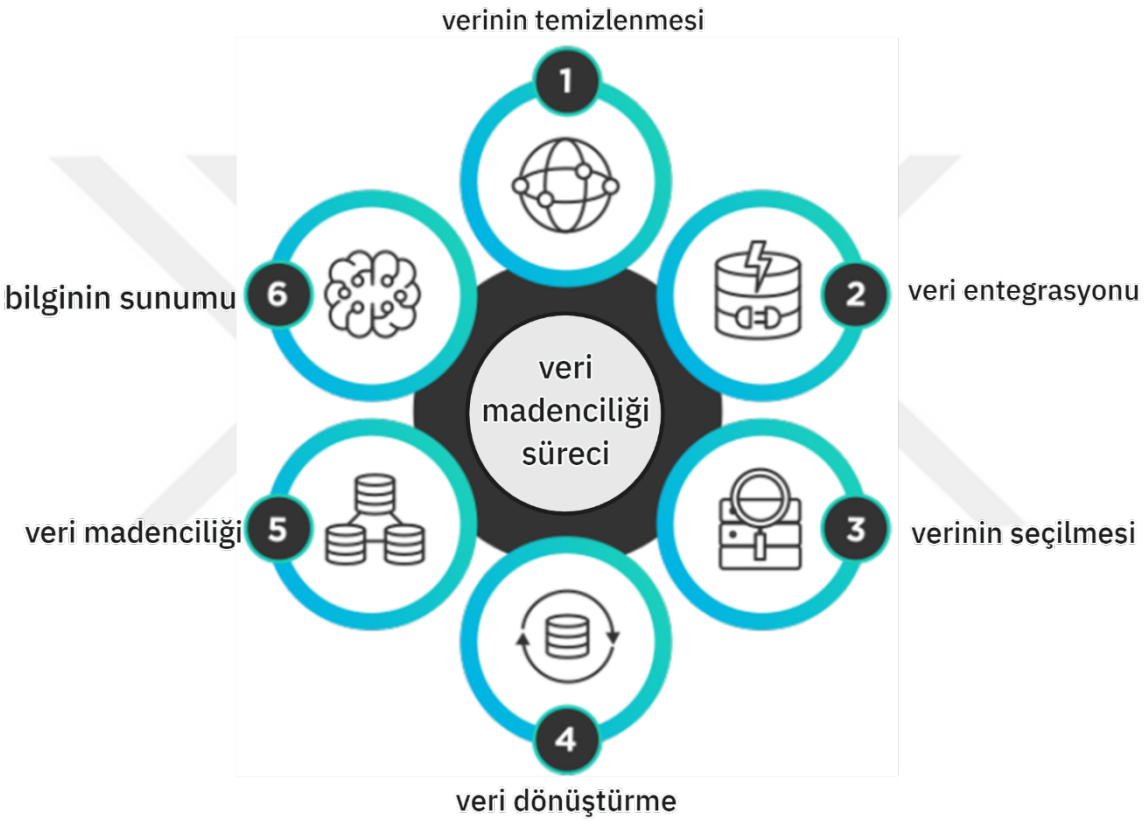
Veri madenciliği döngüsü, veri tabanlarından bilginin çekilerek elden geçirilmesi ve analiz sonucunda ortaya çıkan sonuçların karar vericinin yorumlanması ile tamamlanır (Ersöz, 2019). Veri madenciliği tekniği, veri ön işlemeden veri madenciliği sonuçlarının son işlenmesine kadar ham verileri faydalı bilgilere dönüştürme sürecinin bir parçasıdır (Tan vd., 2005).

Veri madenciliği sürecinde;

1. Problemin tanımlanması,
2. Verilerin hazırlanması,

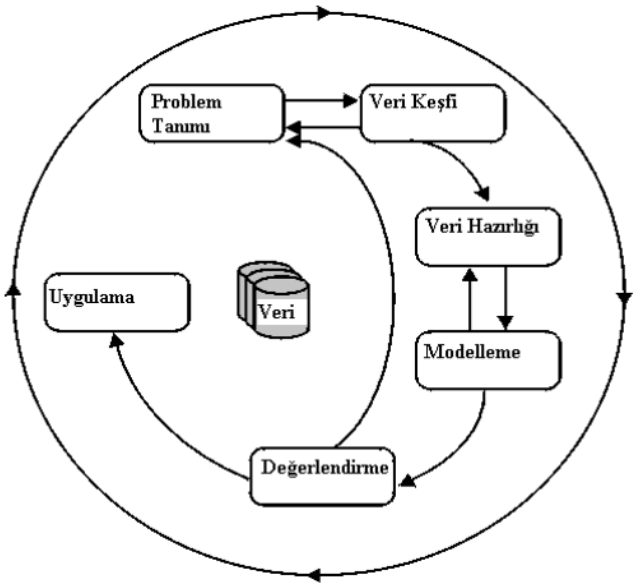
3. Modelin kurulması ve değerlendirilmesi,
4. Modelin kullanılması,
5. Modelin izlenmesi

adımları sırasıyla takip edilir (Shearer, 2000). Albreiki vd. (2021) ise tipik bir veri madenciliği döngüsünü; verinin temizlenmesi, aktarılması, entegrasyonu, veri seçilmesi, veri dönüşümleri, veri madenciliği modellemesi ve bilginin ortaya çıkartılması ve sunulması olarak Şekil 6'daki gibi şematize etmiştir.



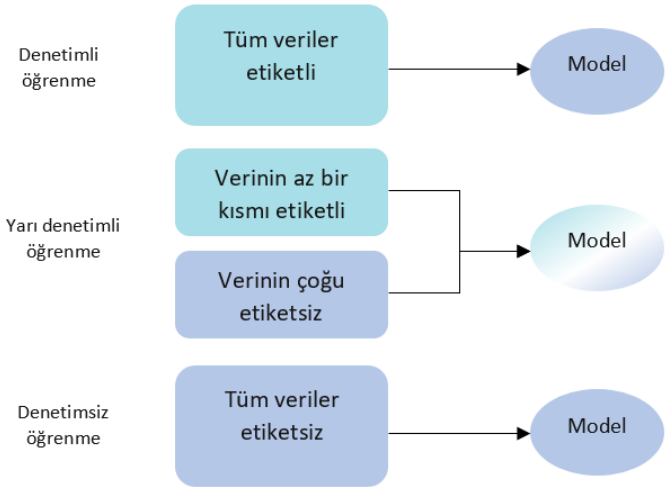
Şekil 6. Veri Madenciliği metodolojisinin tipik döngüsü

Veri madenciliğini süreç olarak tanımlayan sektör uygulayıcıları 1996 yılında CRISP-DM (Cross-Industry Standard Process for Data Mining) adıyla standart model altında bir yapı geliştirmiştir (Akküçük, 2011). Buna göre bir veri madenciliği süreci Şekil 7'deki gibi birbirine bağlı altı aşamadan oluşan bir döngü olarak özetlenebilir: problemin tanımlanması-veri keşfi-veri hazırlığı-modelleme-değerlendirme-uygulama.



Şekil 7. CRISP-DM Modeli ve aşamaları

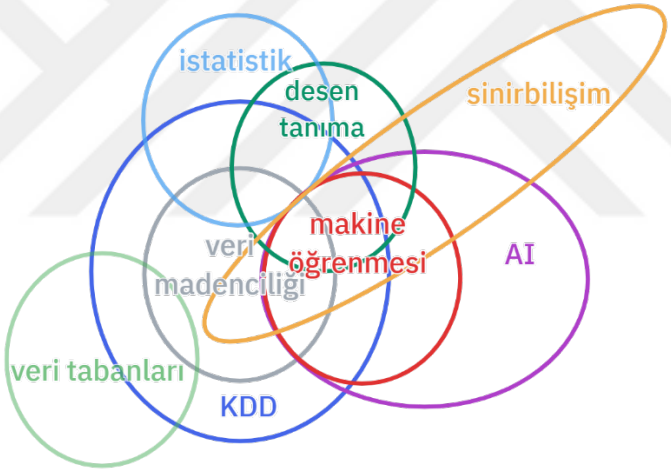
İstatistiki tekniklerin büyük veri ile geliştirilmesi sonucunda oluşan veri madenciliği algoritma ve öğrenme yöntemleri güçlü istatistiksel temelleri olan çoklu mantıksal sınama modellerine dayanmaktadır (Mease ve Wyner, 2008; akt: Aksu, 2018). Veri madenciliği yöntemlerinin genellikle denetimli (supervised), denetimsiz (unsupervised) ve yarı denetimli (semi-supervised) öğrenme olarak 3 türlü sınıflandırıldığı görülmektedir. (bkz: Şekil 8)



Şekil 8. Veri madenciliği yöntemleri

Veri madenciliği esasında pek çok alanla ilişkili görülmektedir (Şekil 9). Özellikle istatistik ve eğitim alanında veri madenciliği algoritmaları geliştirilirken bir yandan da yapay zeka ile

ilgilenen özellikle mühendislik alanlarında "makine öğrenmesi"nde çeşitli algoritmalar gelişmiştir. Günümüze gelindiğinde alan yazın incelenirse özellikle kullanılan yazılımlarda da veri madenciliği algoritmaları ile makine öğrenmesi algoritmalarının bir kısmının çakıştığı görülecektir (Akküçük, 2011; Akpınar, 2017; Ersöz, 2019). Şekil 9'da görüleceği gibi veri madenciliği ve makine öğreniminin, veri tabanlarında bilgi keşfi (Knowledge Discovery in Databases-KDD) gibi ortak bir noktası vardır. Özellikle sınıflayıcı teknikler çoğunlukla aynıdır ve hem veri madenciliğinde hem de makine öğreniminde kullanılmaktadır. Bunun yanında veri madenciliği çok boyutlu bir bakış açısı geliştirdiğinden, madenlenen verinin tipine (veritabanı verisi, işlemsel veri, uzaysal veri, zaman serileri, metin verisi, web verisi, multimedya verisi, sosyal ağ verisi), kullanılan veri madenciliği fonksiyonuna (sınıflandırma, kümeleme, birliktelik, aykırılık), kullanılan tekniklere (makine öğrenmesi, istatistik, OLAP, görselleştirme) ve uygulama alanına göre (perakende, iletişim, bankacılık, pazar-sepet analizi, metin madenciliği, web madenciliği) yöntem ve teknikler sınıflandırılabilir (Özcan, 2013).

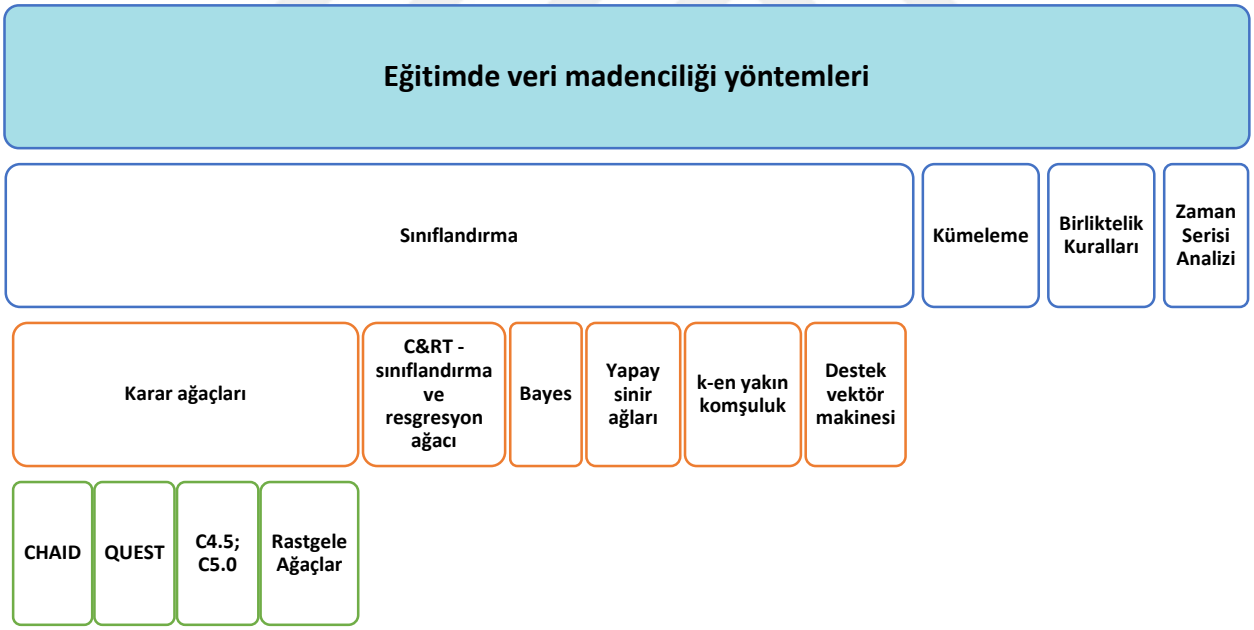


Şekil 9. Veri Madenciliği ve ilişkili disiplinler. Ersöz, F., & Çinar, Y. (2021). Veri madenciliği ve makine öğrenimi yaklaşımlarının karşılaştırılması: tekstil sektöründe bir uygulama. *European Journal of Science and Technology*. doi:10.31590/ejosat.1035124

Veri madenciliği alanında verinin niteliğine göre de pek çok algoritma geliştirilmiştir. Aksu'nun (2018) belirttiği gibi veri seti üzerinde kullanılan algoritma ve yöntemler duruma göre farklı şekillerde ortaya çıkabilmektedir. Örneğin bir veri setinde, tek bir özelliğin çok iyi çalıştığı diğerlerinin ilişkisiz veya gereksiz olduğu örnekler, özelliklerin birbirinden bağımsız ve çıktı değişkenine eşit bir şekilde katkı sağladığı durumlar, basit bir mantıksal yapıda tüm

çıktıların karar ağacı tarafından elde edildiği az özellikli durumlar, bağımsız kural kümesinin olduğu veya özelliklerin farklı alt kümeler arasında birbirine bağımlı olduğu durumlar olabilir. Ayıca veri setinde uygun ağırlıklandırma yönetimiyle belirlenmiş değerler ve aralarındaki doğrusal bağımlılık ya da örneklerin birbiri arasındaki mesafeyi esas alarak uzaydaki yerleşim konumlarını belirleyen durumlar olabilmektedir (Witten ve Frank, 2007). Tüm bunlar veri madenciliği algoritmalarının ihtiyaca göre yeniden geliştirilmesini sağlamıştır. Dolayısı ile bilginin keşfi için farklı veri setleri bünyesinde farklı algoritmaların kullanılması gerekebilir.

Eğitim bağlamında uygulanması ile EVM olarak isimlendirilen yöntemlere, Romero ve Ventura (2010) yine eğitim görevlerine dayalı bir taksonomi önermektedir: Analiz ve Görselleştirme, Geribildirim Sağlama, Tavsiye, Performans Tahmini, Öğrenci Modelleme, Davranış Tespiti, Öğrencileri Gruplandırma, Sosyal Ağ Analizi, Kavram Haritası Geliştirme, Planlama ve Çizelgeleme ve Eğitim Yazılımı Oluşturma. Bununla birlikte eğitim alanı uygun çıktı ve bazı özel uygulamalar gerektirse de, EVM bir Veri Madenciliği uygulaması olduğundan, görevleri de aynıdır: Sınıflandırma, Kümeleme ve Birliktelik Kurallarının analizi gibi (Guarín vd., 2015).



Şekil 10. Veri madenciliği yöntem ve algoritmaları

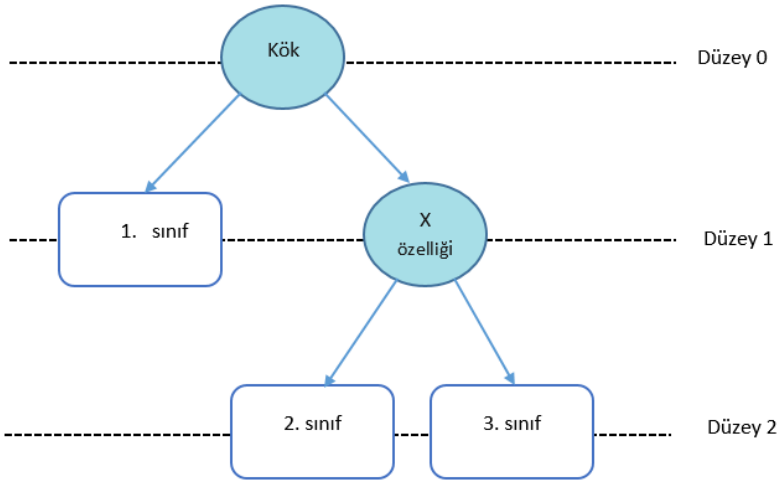
Buradan yola çıkarak Şekil 10'daki şemada yer alan yöntemlere ilişkin alanyazında aşağıdaki tanımlamalar yapılmıştır (Büyükatak, 2022; Dol & Jawandhiya, 2023; Kayri & Şevgin, 2023).

2.2.1.1. Sınıflandırma Yöntemleri

Denetimli öğrenme yöntemleri içinde yer alan sınıflandırma işlemi, verileri önceden tanımlanmış kategorilere ayırma işlemidir. Yeni gelen veriler, bu kategorilere göre sınıflandırılır. Sınıflandırma, denetimli öğrenme yöntemlerinin bir parçası olarak en sık kullanılan yöntemlerden biridir. Ulaşılmak istenen bilginin elde edilmesi için uygulanan sınıflandırma algoritmaları, içerdiği verinin ortak özelliğine göre veri setinin belirli sınıflara ayrılmasını sağlar. Yapılan işlem ile elde edilen modeli yeni bir veri seti üzerinde uygulayarak model ile belirlenmiş olan sınıfların veri seti içindeki benzerlerinin varlığı araştırılır (Demir, 2021). Sınıflandırma algoritması, denetimli öğrenme modeli olduğu için öğrenme ve test verilerini girdi ve çıktı olacak şekilde kullanır. Burada önemli olan eğitim veri setiyle öğrenen modelin, testi yapılarak çıkarımı sağlanan kuralların mümkün olan yüksek doğruluk oranına ulaşmasının sağlanmasıdır (Büyükatak, 2022). Büyük veriler için olası gizli kalmış örüntülerin keşfi için alanda yaygın olarak kullanılmaktadır. Literatürde birçok sınıflandırma yöntemi bulunmaktadır. Karar ağacı, sınıflandırma ve regresyon ağacı, k-en yakın komşuluk, destek vektör makinesi, Naive Bayes sınıflandırıcısı ve yapay sinir ağları temel sınıflandırma yöntemleri arasında en sık kullanılanlardandır (Demir, 2021).

2.2.1.1.1. Karar Ağaçları

Verilen bir kararın önceden belirlenmiş olasılıklar dâhilinde başka karar noktalarını tetiklediği ardışık karar verme süreçlerini yönetmek için karar ağaçları kullanılan en klasik yöntemlerde biridir(Akpınar, 2017). Dolayısıyla sınıflandırma amacıyla veri madenciliğinde de yaygın olarak kullanılmaktadır. Heterojen veri koleksiyonu ve veriyi açıklayan özelliklerin bir kümesi verildiğinde, karar ağaçları, bu özellik değerlerini, veri kümesini daha homojen küçük alt kümelere bölmede kullanmaktadır (Yakut, 2012). Tipik bir karar ağacı öğrenmesinde, arama uzayının bir bölümünde bir çözüm arayan yukarıdan aşağıya bir strateji kullanılır. Karar ağacı, özniteliklerin test edildiği düğümlerden (node) oluşur. Bir düğümün altında kalan dallar, düğümde test edilen tüm olası sonuçlara karşılık gelir. Karar ağaçları akış şemalarına benzeyen yapılardır. Her bir nitelik bir düğüm tarafından temsil edilir. Dallar ve yapraklar ağaç yapısının elemanlarıdır. En son yapı “yaprak”, en üst yapı “kök” ve bunların arasında kalan yapılar ise “dal” olarak isimlendirilir (Yakut, 2012). Şekil 11’de örnek bir karar ağacı yer almaktadır



Şekil 11. Sınıflandırma için genel bir karar ağacı örneği. Aksu, G. (2018). *PISA başarısını tahmin etmede kullanılan veri madenciliği yöntemlerinin incelenmesi.* Doktora Tezi, Hacettepe Üniversitesi, Ankara.

Karar ağacında bulunan düğümler ağaç içinde birbiriyle olan düzeyine göre isimlendirilir. Her aşamada gerçekleştirilen bölme işlemi bir öğrenme sürecidir ve buna göre oluşan yapı kökten yapraklara doğru ilerlediği için bir tümden gelim yaklaşımıdır (Akpınar, 2017). Ağaçta başlangıç noktasında olan kök düğümü "Düzey 0" da iken diğer dal ve yapraklar sırasıyla "Düzey 1", "Düzey 2" olarak devam eder. Ağaçta yer alan çemberler kök ve iç düğümleri, kareler ise yaprak düğümlerini belirtir. Şekil 11'deki özel örnekte, karar ağacı sınıflandırma için tasarlandığından yaprak düğümleri sınıf etiketleri ile isimlendirilmiştir. Düzey 1'de ağacın sol tarafındaki gibi tüm örnekler tek bir sınıfa gidiyorsa bu durumda ağacın alt dalları oluşmayacaktır. Eğer sağ tarafta görüldüğü gibi belirli bir özellik için iki farklı sınıf değeri varsa dallanma devam edecektir (Aksu, 2018).

Karar ağaçları öğreniminde iki ana hedef vardır; sınıflandırma ağaçları olarak isimlendirilen ilk kısımda verinin olabildiğince homojen şekilde alt sınıflara ayrılması; regresyon ağaçları olarak isimlendirilen ikinci grupta ise tahmin modellerinin oluşturulması hedeflenir (Akpınar, 2017).

Akpınar'ın (2017) belirttiği gibi karar ağacı algoritmaları aşağıdaki temel özelliklere göre birbirlerinden ayrılır:

- Kullanılan ölçü skalası (nominal, ordinal, kategorik veya sürekli...)
- Her düğümden sonra ortaya çıkan yeni düğümlerin sayısı (ikili, üçlü veya çoklu düğüm)
- Ağacın büyümesindeki durdurma kriteri: Ağacın büyümesi veya düğümlerin bölünmesi süreci, aşağıdaki kriterlerden birinin sağlanması durumunda sona erer:

i. Bulunulan düğümdeki tüm nesneler tek bir kategoriye aittir,
ii. Başlangıçta tanımlanan ağacın azami seviye sayısına erişilmiştir,
iii. Bulunulan düğümdeki nesnelerin sayısı, kök düğüm için önceden tanımlanan asgari sayıdan daha küçüktür,
iv. Bulunulan düğüm bölündüğünde, ortaya çıkacak bir veya daha fazla yeni düğümde nesnelerin sayısı, önceden tanımlanan asgari sayıdan daha küçüktür,
v. En iyi bölen kriteri, her bölme işlemi sonucunda homojenleşmenin belirli bir düzeyde artması gibi, önceden belirlenen eşik değerin altındadır.

- En iyi bölen özniteliğin seçilmesi (Kategorik değişkenler için Entropi Endeksi, Gini Endeksi, Sınıflandırma Hatası Endeksi, Twoing veya Ordered Twoing (sadece ordinal değişkenler için), sürekli değişkenler için ise En Küçük Kareler Sapması (least-squared deviation (LSD)) en sık kullanılan yöntemlerdendir.
- Ağacı budama süreci

Ağacı budama (pruning) süreci karar ağaçlarının farklılaşmasında önemli aşamalardan biridir ve bu işlemle karar ağaçları optimize edilir. Budama sürecinde ağaçtan uzaklaştırılan yaprak veya dallarda bulunan nesnelerin, hiyerarşik yapı içinde kendilerinden daha üstte yer alan düğümlere devredilmesi gerekir (Akpınar, 2017). Bazı durumlarda ağacın büyümesini sona erdirmek için önceden tanımlanan kriterler esnek/zayıf tutularak karar ağaçlarının olabildiğince büyümesi ve daha sonrasında optimize edilmesine yönelik bir strateji izlenir. Budama süreci ile bir taraftan ağacın küçültülerek derli toplu bir hale getirilmesi, diğer taraftan ezbere öğrenme (overfitting) sorununun giderilmesi amaçlanır (Akpınar, 2017) .

Karar ağacı öğrenimi, ekseriyetle kök düğümden başlayarak yapraklara doğru, sürekli daha saf/homojen düğümleri elde edecek şekilde, bütünü en iyi bölen özniteliği bularak kendini tekrarlayan (recursive) süreçlerdir.

1950'li yılların sonlarından başlayarak sınıflandırma veya tahmin amaçlı (bazen de her ikisi) karar ağaçları çok sık kullanılsa da bu öğrenmenin bazı avantaj ve dezavantajlarını olduğunu söyleyen Akpınar (2017) dezavantajlarını aşağıdaki gibi sıralamıştır:

- Karar ağaçlarında veri dizisinin bölünmesi öznitelik seçiminde kullanılan kriterlerin yetkinliğine bağlıdır. Heuristic özellikle sahip bu kriterlerin optimal bölme işlemi gerçekleştireceğinin bir garantisi olmadığı için, elde edilen sonuçların global optimum olacağını ifade edebilmek mümkün değildir. Süreç genellikle lokal optimada son

bulacaktır. Bu soruna çözüm bulabilmek amacı ile çifte enformasyon uzaklığı (dual information distance (DID)) gibi bazı yöntemler önerilmiştir.

- Karar ağacı algoritmaları çoğu zaman eğitim verisini genelleştiremediğinden, ezbere öğrenme yoğundur. Bu nedenle budama yöntemlerinin etkin kullanımı gerekmektedir.
- XOR, parity ve multiplexer karar ağaçlarında öğrenilmesi, dolayısı ile çözümü zor olan problemlerdir. Bu problemler de karar ağaçları aşırı büyüyebilir. Bu nedenle bu tip problemlerde, istatistiksel ilişkisel öğrenim (statistical relational learning) veya tümevarımsal mantık programlama (inductive logic programming) gibi öğrenme algoritmaları kullanılır.

Bunun yanında; modellerinin anlaşılırlığı ve sonuçları yorumlamanın son derece kolay olması, diğer yöntemler verinin normal dağılmamasına veya eksik değerlere karşı hassasken, karar ağacı modellerinde verinin fazla bir ön işleme gerek duyulmadan kullanılabilmesi, hem sürekli hem de kategorik değerleri bir arada veya ayrı ayrı kullanabilen algoritmalarının bulunması ve ölçeklenebilir özelliği sayesinde büyük veri dizilerinde de kabul edilebilir süreler içerisinde sonuç elde edilebilmesi karar ağacı öğreniminin avantajlarındandır.

Karar Ağacı Algoritmaları

Akpınar (2017) geliştirilen algoritmaları kendi içlerinde, kronolojik olarak bir öncekini eksikliklerinin tamamlandığı aile mantığı içinde gruplandırmıştır. İlk örneklerinden sonra CLS (Concept Learning System) ve AID (Automatic Interaction Detection) aileleri olarak isimlendirilen algoritmaların gelişmesi karar ağacı öğrenimini önemli ölçüde etkilemiştir. Bazı karar ağacı modelleri kronolojik olarak aşağıdaki gibidir.

Tablo 4
Karar Ağacı Algoritmaları Gelişim Süreci

1988 yılında Loh ve Vanichsetakul tarafından QUEST karar ağacı algoritmasının öncülü olarak sunulan FACT, her düğümde bölen özniteliğin seçiminde istatistik testler ve bölen noktanın (split point) bulunmasında diskriminant analizi kullanılır.
1991 yılında Salford Systems tarafından sunulan MARS (Multivariate Adaptive Regression Splines), melez lineer regresyonların (piecewise linear regressions) uyumlandırılması ile esnek modellerin kurulmasını sağlayan parametrik olmayan bir algoritmadır.
1993 yılında Holte tarafından sunulan tek bir öznitelik kullanarak nesneleri sınıflandıran tek seviyeli bir karar ağacı algoritması olan T1, hem sürekli hem de nominal değerlere sahip özniteliklerle çalışır ve eksik değerli işlemleri de destekler.

1994 yılında Müller ve Wysotzki tarafından sunulan CALC5 ve Murthy ve Salzberg tarafından sunulan OC1 (Oblique Classifier 1), özellikle sürekli değere sahip öznitelikler için tasarlanmış bir karar ağacı algoritmasıdır.
1995 yılında Brodley ve Utgoff tarafından oluşturulan LMDT, çok değişkenli testleri temel alan bir karar ağacı algoritmasıdır.
2000 yılında Rastogi ve Shim tarafından sunulan PUBLIC algortiması:, hesaplama karmaşasını azaltmak amacı ile minimum tanımlama uzunluğu (minimum description length) maliyetini kullanarak ağacın büyümesini ve budanmasını entegre eder.
2003 yılında Khoshgoftaar ve Seliya tarafından geliştirilen SPRINT (Scalable Parallelizable Induction of Decision Trees) ise özellikle büyük veri kümelerinin işlenmesi ve analiz edilmesi sırasında ortaya çıkan performans ve hız sorunlarını çözmeyi amaçlamaktadır.

Tablo 4'te verilenlerin dışında CLS ailesi olarak adlandırılan bir dizi algoritma birçok karar ağacı yönteminin temel taşıdır. 1959 yılında Hoveland ve Hunt tarafından yayımlanan algoritma sadece sınıflandırma yapan bir karar ağacı üretmemekte, aynı zamanda eksik sınıflandırma maliyetlerini de gözetmektedir. Buna göre ID3, C4.5, J48, TimeSleuth ve C5.0/See5 algoritmaları bu ailedendir.

Bir diğer önemli karar ağacı algoritmaları grubu ise AID Ailesi olarak isimlendirilir. AID 1971 yılında Sonquist, Baker ve Morgan tarafından sunulan ailenin ilk karar ağacı algoritmasıdır. Bu algoritmayı 1972 yılında Gillo tarafından geliştirilen MAID, 1973 yılında Morgen ve Messenger tarafından geliştirilen THAID (Theta Automatic Interaction Detection) takip etmiştir. CHAID ve Exhaustive CHAID de bu ailenin yakın zaman algoritmalarındandır.

"Hangi nitelik seçimi ölçütü en iyisidir?" literatür göstermektedir ki tüm ölçümlerin bazı önyargıları vardır. Karar ağacı tümevarımının zaman karmaşıklığının genellikle ağaç yüksekliği ile üstel olarak arttığı gösterilmiştir. Bu nedenle, daha sığ ağaçlar üretme eğiliminde olan (örneğin, ikili bölünmeler yerine çok yönlü ve daha dengeli bölünmeleri tercih eden) ölçütler tercih edilebilir. Ancak bazı çalışmalar, sığ ağaçların çok sayıda yaprağa ve daha yüksek hata oranına eğilimli olduğunu da göstermiştir. Birçok karşılaştırmalı çalışmaya rağmen, hiçbir öznitelik seçim ölçütünün diğerlerinden önemli ölçüde üstün olduğu bulunmamıştır (Han, Kamber & Pei, 2011).

Bu çalışma kapsamında SPSS Modeler 18 programı içinde tanımlı karar ağacı algoritmalarından "Random Trees, CHAID, C5.0, QUEST ile C&RT (sınıflandırma ve regresyon ağacı) kullanıldığı için bunların ve sınıflandırmada kullanılan diğer modellerin ayrıntılarına aşağıda sırasıyla yer verilmiştir.

2.2.1.1.2. Rastgele Ağaçlar

Rastgele orman veya rasgele karar ağaçları topluluğu (ensemble learning), sınıflandırma amaçlı kullanılan algoritmalardan biridir. İlk olarak Ho (1998) tarafından kullanılan algoritma rasgele alt uzay örneklemesiyle ve Eugene Kleinberg tarafından öne sürülen "stokastik ayrımcılık" yaklaşımıyla ele alınan sınıflandırmayı çözmeyi amaçlamıştır (Ho 1998; akt. Demir, 2021). Rastgele karar ağaçları, karar ağaçlarının aşırı uyum davranışını çalışma zamanında (training time) düzeltmektedir.

Rastgele ağaç (RandomTree) yönteminde her bir düğümde daha önce belirlenen sayıda rastgele özellik seçilerek bir test yapılırken ağacın dallarında budama yapılmamaktadır. Karar ağaçlarında her bir düğümde en iyi çalışan yani en çok bilgi veren özellik seçilirken rastgele ağaç yönteminde bu seçim tesadüfi olarak gerçekleşmektedir (Aksu, 2018). Rastgele karar ağacı algoritması tesadüfi öznitelik seçimi kullanan budanmamış sınıflandırma ve regresyon ağaçlarının bir topluluğudur (Akpınar, 2017). Yöntem, her adımda safsızlığı en aza indirerek eğitim kayıtlarını segmentlere ayırmak için özyinelemeli bölümleme kullanır (Breiman, 2001).

Breiman'a (2001) göre Rastgele Orman karar ağacı, veri setinde en iyi niteliklerden seçilen düğümleri dallara ayırmak yerine, her bir düğümde rastgele alınan niteliklerin en iyisini seçerek tüm düğümleri dallara ayırır. Her veri kümesi asıl veri setinden yer değiştirmeli olarak üretilir. Rastgele özellik seçimi kullanılarak ağaçlar geliştirilir ve budama işlemi yoktur. Rastgele orman algoritmasının diğer algoritmalara göre daha hızlı ve doğru olmasının sebebi de bu yöntemdir (Akçetin ve Çelik 2014; akt. Demir, 2021). Rastgele orman algoritmasında sonradan dâhil edilen veriye ilişkin tahmin yapılmasının yanı sıra, değişkenlere ait önem derecesi de hesaplanmaktadır. Veri seti bünyesinde çok sayıda değişken barındırıyorsa, değişkenlerin önem derecelerinin hesaplanmasıyla değişken sayısını azaltma yönünden etkilidir (Akman, 2010). Bu süreçte birden fazla karar ağacı oluşturarak doğru sınıflandırma oranını yükseltmede etkili bir algoritmadır. Rastgele olarak seçilen karar ağaçlarının birleşmesiyle karar ormanı oluşturulur. Veri setlerinde bulunan fazla sayıdaki değişken ve sınıf etiketine sahip kategorik değişken bulunduran veri setlerinde iyi sonuçlar verir (Aydın vd., 2019).

SPSS Modeler 18'de yer alan "random trees" yani "rastgele ağaçlar" algoritması, genellikle sınıflandırma ve regresyon görevleri için kullanılan bir topluluk öğrenme tekniğidir. Bu algoritma, birden fazla karar ağacı oluşturarak çalışır ve bu ağaçların her biri, veri setinin rastgele seçilmiş alt kümeleri üzerinde eğitilir (IBMCorporation, 2022). C&RT'de olduğu gibi,

tahmin yöntemi de eğitim kayıtlarını benzer çıktı alanı değerlerine sahip segmentlere ayırmak için özyinelemeli bölümleme kullanır. Düğüm, bölünmeden kaynaklanan bir safsızlık endeksindeki azalma ile ölçülen en iyi bölünmeyi bulmak için mevcut girdi alanlarını inceleyerek başlar. Bölme iki alt grup ile başlar, bunların her biri daha sonra iki alt gruba daha bölünür ve durdurma ölçütlerinden biri tetiklenene kadar bu böyle devam eder (SPSSModeler, 2020). Dolayısı ile tüm bölünmeler ikilidir. Rastgele ağaçlarla SPSS Modelerda model oluştururken veriler sürekli (sayısal aralık) veya kategorik olabilir. "Both" veya "None" olarak belirtilen alanlar analizde yok sayılır. Modelde kullanılan alanların türleri tam olarak örneklendirilmiş olmalıdır ve modelde kullanılan tüm sıralı (sıralı küme) alanlar sayısal depolamaya sahip olmalıdır (dize değil). Gerekirse, bunları dönüştürmek için Reclassify düğümü kullanılmalıdır.

Rastgele ağaçlar da C&RT benzeri ağaçlar kullanır. Bu tür ağaçlar ikili olduğundan, bölme işleminde her alan iki dalla sonuçlanır. Birden fazla kategoriye sahip değişkende bu kategoriler, iç bölme kriterine göre iki gruba ayrılır. Her ağaç mümkün olan en geniş ölçüde büyür (budama yoktur). Puanlamada, Rastgele Ağaçlar bireysel ağaç puanlarını çoğunluk oylaması (sınıflandırma için) veya ortalama (regresyon için) ile birleştirir.

Rastgele Ağaçlar, C&RT'ye kıyasla orijinal veri kümesinden değiştirme ile örnekleme yaparak eğitim veri kümesinin kopyalarını oluşturur ve buna torbalama işlemi (bootstrap örneklemesi) denir. Yani orijinal veri kümesine eşit boyutta bootstrap örnekleri oluşturup örnek verilerle ağaç büyütülürken Rastgele Ağaçlar verileri tekrar örneklemez. Bunun yerine, tahmin edicilerin bir kısmını rastgele seçer ve bir ağaç düğümünü bölmek için bunlardan en iyisini kullanır. Bu işlem her ağaç düğümü bölünürken tekrarlanır ve böylece ağaç yapısı bu temel fikir üzerinde geliştirilir (IBMCorporation, 2022).

Rastgele Ağaçlar, C&RT'den farklı olarak belirli sayıda tahminciyi rastgele seçer ve her bir düğümü bölmek için bu seçimden en iyisini kullanır. Ayrıca ağacın her bölünmesinde, safsızlık ölçümü için yalnızca girdi alanlarının bir örneklemini dikkate alır. Buna karşılık, C&RT tüm tahmin ediciler arasından en iyisini bulur. Rastgele Ağaçlardaki her ağaç, her yaprak düğümü tipik olarak tek bir kayıt içerene kadar tamamen büyür. Dolayısıyla ağaç derinliği çok büyük olabilir. Ancak standart C&RT, ağaç büyümesi için farklı durdurma kuralları kullanır ve bu da genellikle çok daha sığ bir ağacı ortaya çıkartır.

Rastgele Ağaçlar modelleri, büyük veri setleri ve alan sayılarıyla uğraşırken sağlamdır. Torbalama ve alan örneklemesi kullanımı nedeniyle, aşırı uyuma çok daha az eğilimlidirler.

Yüksek doğruluk oranları ve robust yapısı nedeniyle birçok farklı alanda tercih edilen güçlü bir makine öğrenimi modelidir. Rastgele ağaçlar algoritması, hangi özelliklerin çıktıyı en çok etkilediğini belirleme konusunda da bilgi sağlar. Bu da modelin açıklanabilirliğini artırır ve hangi değişkenlerin tahminler üzerinde daha etkili olduğunun anlaşılmasını sağlar. SPSS Modeler bu algoritma ile kullanıcılara karmaşık veri setleri üzerinde etkili analizler yapma olanağı sunar (IBMCorporation, 2022). Ancak mevcut yapısı etkileşimli veri analizi için ideal değildir çünkü yüksek etkileşimli önbelleğe alma gereksinimlerini karşılayamazlar.

2.2.1.1.3. CHAID

1980 yılında Gordon Kass tarafından geliştirilen CHAID (Chi-Squared Automatic Interaction Detector) algoritması, tahmin değişkenleri içerisinden en iyi ayrımı yapabilen kriteri istatistiksel olarak analiz eder ve fark kalıplarını belirler (Akdemir, 2021). Tahmin değişkenleri hedef değişkene uyan bir çiftin içinde istatistiki olarak bir fark kalmayıncaya kadar birleştirilmesiyle en iyi sınıfın hesaplanması tekniğine dayanır en iyi sınıf için ki-kare istatistiği kullanır (Ersöz, 2019). CHAID, kategorik tüm değişken tipleriyle çalışabilen bir algoritmadır ve bu değişkenler arasında önemli farklılıkları belirlemek için de kullanılır. Ağaç yöntemleri içinde yer alan bu algoritma, özellikle C4.5 ve C&RT gibi ağaçları tetikler ancak onların aksine ikili olmayan ağaçlar da üretir. CHAID, kolay uygulanabilirliği ve analitik açıdan otomatik işlemleri nedeniyle tercih edilir.

Akdemir (2021) CHAID algoritmasında sürecin üç adımdan oluştuğunu belirtmektedir; birleştirme, dağıtma, durdurma. Buna göre SPSS Modeler 18'de yer alan CHAID algoritması; ilk olarak her bir girdi alanı ile sonuç arasındaki çapraz tabloları inceler ve ki-kare bağımsızlık testi kullanarak anlamlılığı test eder. Bu ilişkilerden birden fazlası istatistiksel olarak anlamlıysa, CHAID en anlamlı (en küçük p değeri) olan girdi alanını seçer. Bir girdinin ikiden fazla kategorisi varsa, bunlar karşılaştırılır. Sonuçlarda hiçbir farklılık göstermeyen kategoriler bir araya getirilir. Kategori birleştirme, en az anlamlı fark gösteren kategori çiftinin art arda birleştirilmesiyle yapılır (SPSSModeler, 2020).

Daha sonra dağıtma aşamasında, en uygun şekilde birleşmiş kategoriler için önem derecesi hesaplanır, en büyük değere sahip olan değişken diğerlerinden ayrılır. Eğer önem derecesi belirlenen değerden büyük olursa veri kümesi seçilen açıklayıcı değişkenin birleşmiş gruplarına göre bölünür. Veri setinde analiz edilmemiş her grup için birleştirmenin ilk adımına dönülür ve burada en az sayıda gözleme sahip olan gruplar göz ardı edilebilir (Elkonca, 2013). Bu kategori

birleştirme işlemi, kalan tüm kategoriler belirtilen test seviyesinde farklılık gösterdiğinde durur. Nominal girdi alanları için herhangi bir kategori birleştirilebilir; sıralı bir küme için yalnızca bitişik kategoriler birleştirilebilir.

CHAID modelinin nominal ve sürekli değişkenler ile çalışılabilmesi modelin avantajlarındandır. İkili olmayan ağaçlar ve sayısal aralık veya kategorik girdi alanları ile optimum bölünmeleri belirlemek için ki-kare istatistiklerini kullanarak karar ağaçları oluşturur. Kapsamlı (Exhaustive) CHAID, CHAID'in bir modifikasyonudur ve her bir tahmin edici için tüm olası bölünmeleri incelemeye yönelik daha kapsamlı bir iş yapar, ancak hesaplamaları daha uzun sürer (IBMCorporation, 2022).

CHAID analizinde, ilk basamakta yapılan aşamalı (stepwise) regresyon analizi sonucu; bağımlı değişken ile en çok ilişkili değişkenler düğümleri oluşturur. Dolayısı ile analiz sonucunda bağımlı değişken ile en çok ilişkili olan bağımsız değişkenlerin bir göstergesi ortaya çıkar (Kayri, 2007). Bu doğrultuda, CHAID analizindeki karar ağacının altında oluşan düğümler bir anlamda değişkenlerin önem sırasını da belirlemektedir. Pek çok alanda uygulama olanağı bulunan karar ağaçları içinde yer alan CHAID analizinin eğitim bilimleri ile ilgili araştırmalar için de verimli bir şekilde kullanılabileceğini belirten Kayri (2007), bu istatistiksel metot kullanılarak, özellikle heterojen veri setlerinde, homojen düğümlerin (node) oluşturulabileceğini ve elde edilecek olan bulguların gerçeği daha yansız gösterebileceğini söylemektedir.

2.2.1.1.4. C5.0

Ross Quinlan tarafından 1994 yılında piyasa sunulan algoritma yine karar ağacı algoritmalarından C4.5'in ve öncesinde yayınlanan ID3'ün bir uzantısı olarak kabul edilir. C5.0 ise C4.5'e göre gelişimsel iyileştirmelere sahiptir.

C5.0'ı daha iyi anlamak için öncülü olan C4.5 ve ID3 algoritmasından bahsetmek gerekir. Karar ağacı öğreniminde, ID3 (Iterative Dichotomiser 3) Ross Quinlan tarafından 1986 yılında kamuya açıklanan bir algoritmadır (Topuz, 2021). ID3, C4.5 algoritmasının öncülü olup, makine öğrenimi ve doğal dil işleme (natural language processing) alanlarında tipik olarak kullanılmıştır. ID3 algoritmasında bölen özniteliğin seçilmesinde entropi veya gini endeksi kullanılır. Algoritma sadece daha önce bölen olarak kullanılmamış öznitelikleri seçerek ağacın yapraklara doğru homojenleşmesine çalışır (Akpınar, 2017). Algoritmada yeni dallara ayrılma

işlemi; seçilecek tüm özniteliklerin tüketilmesi veya seçilen özniteliğin kategorisinde tanımlanan kritere uygun gözlemlerin tüketilmesi durumunda, (Örneğin yaş>100 kriterini gerçekleştirecek örnek kalmaması) sona erer.

C4.5 ise 1993 yılında Ross Quinlan tarafından yayımlanan ve ID3'ün uzantısı olarak kabul edilen karar ağacı algoritmasıdır. CLS ailesinin özelliklerini bünyesinde barındıran C4.5, en iyi bölen öznitelik seçiminde entropi ve enformasyon kazancı tekniklerini kullanır.

C4.5'e özgü bazı özellikler şunlardır:

- Tüm örnekler aynı sınıfa ait olduğunda, bir yaprak düğüm oluşturularak, bu düğüm bu sınıfın adı ile isimlendirilir.
- Özniteliklerin hiç birinin enformasyon kazancına sahip olmaması ve daha önce görülmemiş bir sınıfın ortaya çıkması durumunda, C4.5, sınıfın beklenen değerini kullanarak bir karar düğümü oluşturur.

C4.5, ID3'e göre hem kategorik, hem de sürekli öznitelik değerlerini işleyebilir. Sürekli öznitelik değerleri için veriyi ikiye bölecek bir eşik değeri kullanır. Veri dizisinde eksik değerlerin bulunması durumunda, bu değerler "?" ile işaretlenir ve bu eksik değerler entropi ve enformasyon kazancı hesaplamalarına dahil edilmez (Akpınar, 2017). C4.5 algoritması, J48 ismi ile açık kaynak kodlu olarak da kullanılabilir.

Quinlan C5.0 algoritmasını Unix / Linux için geliştirirken, Windows için ise See5 algoritmasını piyasaya sunmuştur. C5.0 da hem kategorik hem de sürekli verileri işleyebilir. Özellikle sürekli verileri işleyebilmek için bir eşik değer hesaplar ve değerleri eşiğin üstünde - altında olarak ikiye böler. C5.0 modeli, örneği maksimum bilgi kazancına sahip alana göre bölerek bir karar ağacı veya kural seti oluşturur. İkiden fazla alt gruba birden fazla bölünmeye izin verir ve her düğümden çıkan çoklu dallar ile karar ağacı oluşturur. Dalların sayısı tahmin edicinin kategori sayısına eşittir ve tek bir sınıflayıcıda birden çok karar ağacını birleştirebilir. Dalları ayırma işlemi için bilgi kazancı kullanılır. Budama işlemi ise her yapraktaki hata oranına dayanmaktadır (Ersöz, 2019). C5.0, öncüllerine göre bazı gelişmiş özelliklere sahiptir. C5.0'ın gözle görülür biçimde daha süratli olması, daha etkin bellek kullanımı, daha küçük karar ağaçları, boosting desteği ve muhtemel faydasız özniteliklerin elimine edilmesini sağlayan winnowing özelliği bunların başlıcalarıdır (Quinlan, 2013; akt. Akpınar, 2017).

2.2.1.1.5. C&RT

Breiman, Freidman, Olshen ve Stone tarafından 1984'te geliştirilmiş olan C&RT (Classification and Regression Trees) modeli, özyinelemeli bölümleme kullanarak ve safsızlığı en aza indirerek gelecekteki gözlemleri tahmin etmek için bir karar ağacı oluşturur (Şevgin, 2020). Vakaların %100'ü bir hedef alan kategorisine giriyorsa bir model "saf" olarak kabul edilir.

Algoritma, kök düğümden başlayarak her bir düğümde ikili dallandırma yaparak büyür. Bu ikili dallandırma sebebiyle ikili ağaç (binary tree) yapısına sahiptir. Yani her bir dallandırma işleminde, bağımsız değişken üzerinden, sorulara evet veya hayır şeklinde cevaplar verilir. Bu ikili dallandırma, bir sonraki seviyede daha homojen hale getirilmiş düğümler elde etmeyi sağlamaktadır. Hedef ve girdi alanları ikili bölünmelerle sayısal veya kategorik olabilir. Bağımlı değişkenin kategorik olması durumunda genellikle twoing kriteri ve gini ölçüsü; sürekli değişkenlerin bölen noktalarının bulunmasında ise en küçük kareler sapması (least squared deviation) kriteri kullanılmaktadır (Akpınar, 2017).

C&RT ve CHAID algoritmaları, veri madenciliğinde yaygın olarak kullanılan karar ağacı yöntemleridir. Her iki yöntem de kategorik ve sürekli değişkenlerle çalışabilir ve karar ağaçları oluşturarak veri setlerindeki yapıları keşfetmeye yöneliktir. Ancak, metodolojileri ve uygulama alanları açısından bazı önemli farklılıklar içerirler. Aşağıda bu iki algoritmanın benzer ve farklı yönlerini özetleyen bir tablo bulunmaktadır:

Tablo 5
C&RT ve CHAID Karar Ağaçlarının Karşılaştırılması

Özellik	C&RT	CHAID
Yöntem	İkili bölünmeler kullanır.	Çoklu bölünmeler kullanır.
Bölünme Kriteri	Gini veya entropi kullanarak en iyi bölünmeyi belirler.	Ki-kare istatistiği kullanarak kategoriler arasındaki bağımsızlık testi yapar.
Veri Türleri	Sürekli ve kategorik değişkenleri işleyebilir.	Genellikle kategorik değişkenlerle daha etkilidir.
Ağaç Budama	Ağacın büyümesini kontrol etmek için budama yapar.	Budama genellikle yapmaz, doğrudan en iyi bölünmeyi belirler.
Model Karmaşıklığı	Daha basit ve az parametreli modeller üretir.	Daha karmaşık modeller oluşturabilir.
Kullanım Alanları	Hem sınıflandırma hem de regresyon analizlerinde kullanılır.	Çoğunlukla sınıflandırma problemleri için kullanılır.
Aşırı Öğrenme Riski	Budama ile aşırı öğrenme riskini azaltır.	Budama yapmadığı için aşırı öğrenme riski daha yüksektir.

Tablo 5, C&RT ve CHAID algoritmalarının temel özelliklerini ve kullanım alanlarını karşılaştırmakta ve bu iki popüler karar ağacı algoritması arasındaki temel benzerlik ve farklılıkları vurgulamaktadır.

2.2.1.1.6. QUEST

1997'de Loh ve Shih tarafından sunulan QUEST (Quick, Unbiased, Efficient, Statistical Tree) çok kategorili bağımsız değişkenlerin olduğu durumlarda daha az hata ile hızlı hesaplamayı amaçlayan karar ağacı algoritmalarındandır (Kayri & Şevgin, 2023). QUEST modeli, kapsamlı C&RT analizleri için gereken işlem süresini azaltmak ve sınıflandırma ağacı yöntemlerinde bulunan daha fazla bölünmeye izin veren girdileri tercih etme eğilimini azaltmak için tasarlanmış karar ağaçlarında ikili bir sınıflandırma yöntemi sunar. Girdi alanları sayısal aralıklar (sürekli) olabilir, ancak hedef alan kategorik olmalıdır. Tüm bölünmeler ikilidir. QUEST modeli, karar ağaçları için ikili sınıflandırma sunarak işlem süresini de kısaltır ve ikili bölünmelerle kategorik girdileri tercih ederek C&RT analiz önyargısını azaltır (IBM Corporation, 2022). Ancak birim değişken ayracı tarafsız değişken seçimini tahmini olarak yapar ve ikili dal yapısına sahip olduğu için karar ağacı çok büyük olabilir. Genel olarak C&RT algoritması üzerine kurulu bir yapısı vardır fakat CHAID ve C&RT 'den farklı olarak değişken seçimi ve ayrım noktası seçimi işlemlerini ayrı ayrı ele almaktadır. QUEST Algoritması, C&RT'nin sağladığı bütün avantajlara sahip olmasına rağmen yine de ağaç hantal yapıdadır. QUEST, bağımlı değişkenler kategorik yapıda ise, yansız ağaç önemli ise, büyük ve karmaşık yapıda veri seti varsa, ağaç tahmininde etkili bir algoritmaya gerek varsa ve ağaç ikili bölünme ile sınırlanmışsa kullanılır (Ersöz, 2019). Ağaç bölünmelerinde bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi analiz eder. Her bir bölmede bağımsız değişken ile bağımlı değişken arasındaki ilişki ANOVA F-testi, Levene testi (Sıralı ve sürekli nitelikler için) veya Pearson Ki-Kare (Chi-square)(nominal nitelikler için) testi kullanılarak hesaplanır (Kayri & Şevgin, 2023).

Kayri ve Şevgin'in belirttiği gibi QUEST:

- Büyük veri setlerinde etkin çalışmak üzere tasarlanmıştır. Veriyi bölmek için kullanılan yöntemler, hesaplama karmaşıklığını azaltır ve algoritmanın hızını artırır.

- Çapraz doğrulama yaparak bölme işlemi için en iyi değişkeni ve kesme noktasını belirler. Bu, ağaçların genelleştirilebilirliğini artırır ve önyargıyı azaltır.
- C&RT'de olduğu gibi ön olasılıklar kullanmaktadır.
- CHAID'de olduğu gibi maliyet matrisi doğrudan ağaç büyütme sürecine katılmaz.
- Tek değişkeni veya doğrusal kombinasyonları bölen bir karar ağacı algoritması olduğu için değişken seçim yöntemi tutarlıdır.

Özetle, QUEST algoritması, sınıflandırma ve veri madenciliği uygulamalarında hızlı, önyargısız ve verimli sonuçlar elde etmek için geliştirilmiş bir karar ağacı algoritmasıdır. İstatistiksel testlerle bölme işlemleri yapar ve büyük veri setlerinde etkin ve hızlı çalışır.

2.2.1.1.7. BAYES Algoritması

Bayes sınıflandırıcısı, tahminler arasında bağımsızlık varsayımıyla matematikçi Thomas Bayes'in adını verdiği Bayes Teoremine dayanan bir sınıflandırma tekniğidir. Basit bir ifadeyle, bir sınıftaki belirli bir özelliğin varlığının diğer herhangi bir özelliğin varlığı ile ilgisiz olduğunu varsayar. Dolayısı ile Naive Bayes yönteminde sınıflandırma yapılırken özellik değerleri birbirinden bağımsız olarak hesaba katılmaktadır. Bu özellikler birbirine veya diğer özelliklerin varlığına bağlı olsa bile, bu özelliklerin tümü bağımsız olarak olasılıklara katkıda bulunur (Demir, 2021). Naive Bayes modelinin oluşturulması kolaydır ve özellikle çok büyük veri kümeleri için kullanışlıdır.

Veri setindeki örnekler üzerinden gerçekleştirilen olasılık hesapları yardımıyla veri dosyasında daha önce yer almayan yeni örneğin daha önce elde edilmiş olasılık değerlerine göre hangi sınıfta yer alacağı tespit edilmeye çalışılır (Qin, Xia, Wang ve Du, 2011; akt. Aksu, 2018).

Buna göre *x* gözleminde oluşan C sınıfının koşullu olasılık değeri;

$$P(C|x) = \frac{P(x|C).P(C)}{P(x)}$$

olarak hesaplanır.

Verilen eşitlikte $P(C)$ ifadesi C sınıfına ait önsel olasılık değerlerini, $P(x|C)$ ifadesi C sınıfından bir örneğin x olma olasılığını, $P(x)$ ifadesi veri dosyasındaki gözlemlerin dağılımını ve $P(C|x)$ ifadesi kendisi x olarak bilinen bir örneğin C sınıfına ait olmasına ilişkin sonsal olasılık değerini göstermektedir (Wolf, 2016; akt. Aksu, 2018). Naive varsayımına göre tüm özellikler istatistiksel olarak bağımsız ve her bir özelliğin önsel olasılık değerinin birbirine eşit olduğu

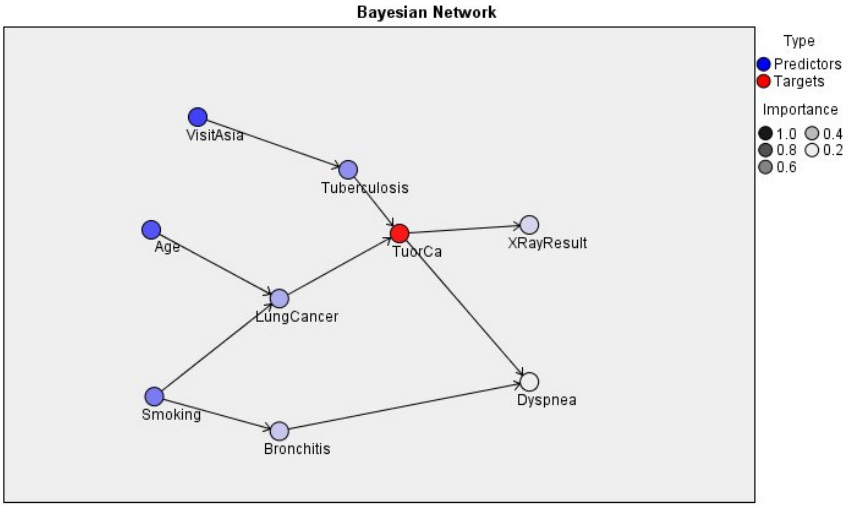
kabul edilir. Günlük hayatta verilerde sınıf değerleri için niteliklerin nadiren bağımsız olduğu dolayısı ile Naive Bayes yönteminin önemli sınırlılıklarının bulunduğu belirtilse de Ceci (2005) Naive Bayes yönteminin daha karmaşık öğrenme yöntemleri olarak görülen karar ağacı, kurallı öğrenme ve örnek temelli öğrenme algoritmalarından daha doğru kestirimler yaptığını söylemektedir (Aksu, 2018; Demir, 2021).

SPPS Modeler'da yer alan Bayesian Network düğümü de Bayes olasılığına dayalı olarak, görünüşte bağlantısız nitelikleri kullanarak olayların olasılığını belirlemek için gözlemlenen ve kaydedilen kanıtları "sağduyulu" gerçek dünya bilgisiyle birleştirerek bir olasılık modeli oluşturmaya olanak tanır. Programda bunun için Tree Augmented Naïve Bayes (TAN) ve Markov Blanket ağları kullanılır ve eksik bilgilerde bile tahminler yapılabilir.

Bayes ağı, bir veri kümesindeki değişkenleri (genellikle düğümler olarak adlandırılır) ve bunlar arasındaki olasılıksal veya koşullu bağımsızlıkları gösteren grafiksel bir modeldir. Düğümler arasındaki nedensel ilişkiler bir Bayes ağı ile temsil edilebilir; ancak ağdaki bağlantıların (yaylar olarak da bilinir) doğrudan neden ve sonucu temsil etmesi gerekmez. Ağ genel olarak, nedensel ilişkileri anlamaya, sonuçları tahmin etmeye ve aşırı uyumu önlemeye yardımcı olur.

Örneğin, bir Bayesian ağı, grafikte (Şekil 12) gösterildiği gibi semptomlar ve hastalık arasındaki olasılıksal bağımsızlık doğruysa, belirli semptomların ve diğer ilgili verilerin varlığı veya yokluğu göz önüne alındığında, bir hastanın belirli bir hastalığa sahip olma olasılığını hesaplamak için kullanılabilir. Ağlar, bilgilerin eksik olduğu durumlarda da mevcut bilgileri kullanarak mümkün olan en iyi tahmini yapar.

Bayesian ağının yaygın ve temel bir örneği Lauritzen ve Spiegelhalter (1988) tarafından oluşturulmuştur. Genellikle "Asya" modeli olarak adlandırılan Lauritzen ve Spiegelhalter modeli bir doktorun yeni hastalarını teşhis etmek için kullanılabilecek bir ağın basitleştirilmiş bir versiyonudur; bağlantıların yönü kabaca nedenselliğe karşılık gelir. Her bir düğüm, hastanın durumuyla ilgili olabilecek bir özelliği temsil etmektedir; örneğin, "Smoking" hastanın sigara içtiğini, "VisitAsia" ise yakın zamanda Asya'yı ziyaret edip etmediğini göstermektedir. Olasılık ilişkileri herhangi bir düğüm arasındaki bağlantılarla gösterilir; örneğin, sigara içmek hastanın hem bronşit hem de akciğer kanseri geliştirme olasılığını artırırken, yaş sadece akciğer kanseri geliştirme olasılığı ile ilişkili görünmektedir. Aynı şekilde, akciğer röntgenindeki anormalliklerin nedeni tüberküloz veya akciğer kanseri olabilirken, nefes darlığı (dispnea) çeken bir hastanın bronşit veya akciğer kanseri olma ihtimali de artmaktadır (SPSSModeler, 2020).



Şekil 12. Bayes ağı; Lauritzen ve Spegelhalter'in Asya ağı örneği

Bir Bayes ağı kullanmaya karar vermeniz için çeşitli nedenler vardır:

- Nedensel ilişkiler hakkında bilgi edinmenize yardımcı olur.
- Bu sayede bir sorun alanını anlamanızı ve herhangi bir müdahalenin sonuçlarını tahmin etmenizi sağlar.
- Ağ, verilerin aşırı uyumundan kaçınmak için etkili bir yaklaşım sağlar.
- İlgili ilişkilerin net bir görselleştirmesi kolayca gözlemlenir.

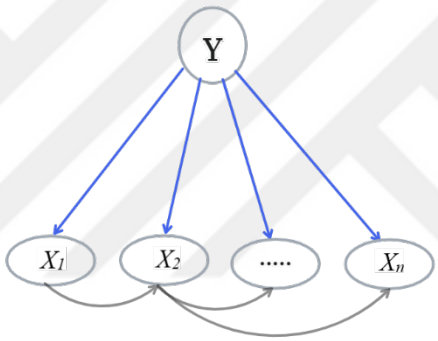
Bayes ağında hedef alanlar kategorik olmalıdır ve Nominal, Sıralı veya İşaret (Nominal, Ordinal veya Flag) ölçüm düzeyine sahip olabilir. Girdiler ise herhangi bir türde alanlar olabilir. Sürekli (sayısal aralık) girdi alanları otomatik olarak gruplandırılacaktır; ancak dağılım çarpıksa Bayesian Network düğümünden önce bir gruplandırma düğümü kullanılıp alanları manuel olarak gruplandırarak daha iyi sonuçlar elde edilebilir.

SPSS Modeler, Bayesian ağ modelleri oluşturmak için iki farklı yöntem sunar (IBMCorporation, 2022):

Markov Örtüsü tahmini (Markov Blanket estimation): Bir Bayes ağında hedef değişken için Markov örtüsü, hedefin ebeveynlerini, çocuklarını ve çocuklarının ebeveynlerini içeren düğüm kümesidir. Markov örtüsü, hedef değişkeni tahmin etmek için gerekli olan ağdaki tüm değişkenleri tanımlar. Bu daha karmaşık ağlar üretebilir ve üretilmesi daha uzun sürer. Veri işleme aşamasında özellik seçim fonksiyonunun kullanılması bu algoritmanın performansını önemli ölçüde artırabilir.

Tree Augmented Naive Bayes: Bu algoritma esas olarak sınıflandırma için kullanılır. Basit bir Bayes ağ modelini verimli bir şekilde oluşturur. Model, her bir tahmin edicinin hedef değişkene ek olarak başka bir tahmin ediciye bağlı olmasına izin verdiği için Naive Bayes modeline göre bir gelişmedir. Başlıca avantajları, genel Bayesian ağ modellerine kıyasla sınıflandırma doğruluğu ve çıkarım hızıdır. Dezavantajı da basitliğinden kaynaklanmaktadır: ağlar, değişkenler arasındaki bağımlılık yapısına çok fazla kısıtlama getirmektedir.

Son yıllarda, birçok çalışma Naive Bayes sınıflandırıcısını geliştirmeye odaklanmıştır. Önemli yöntemlerden biri bağımsızlık varsayımını (sınıf verildiğinde tahmin ediciler arasındaki koşullu bağımsızlık) gevşetmektir. SPSS Modeler basitçe Şekil 13'te modellenen bir ağaçla güçlendirilmiş Naive Bayes (TAN) sınıflandırıcı kullanarak (Friedman, Geiger ve Goldszmidt, 1997 akt: IBMCorporation, 2022) her tahmin edicinin hedefi bir ebeveyn olarak alması ve tahmin edicilerin bir başka tahmin ediciyi ebeveyn olarak alabilmesini sağlar.



Şekil 13. Basit bir Naive Bayes ağacının yapısı (TAN)

Bayes TAN modeli, bağımlılıkları dikkate alarak Naive Bayes modelinin doğruluğunu artıran ve çeşitli uygulama alanlarında kullanılabilen esnek ve etkili bir sınıflandırma yöntemidir. Genelde teorik olarak, Bayes sınıflandırıcıları diğer tüm sınıflandırıcılara kıyasla minimum hata oranına sahiptir. Ancak pratikte bu durum, sınıf-koşul bağımsızlığı gibi kullanımı için yapılan varsayımlardaki yanlışlıklar ve mevcut olasılık verilerinin eksikliği nedeniyle her zaman geçerli olmayabilir (Han vd., 2011).

2.2.1.1.8. Lojistik Regresyon -LR

Lojistik regresyon (LR), kategorik bir hedef alan kullanarak giriş değerlerine göre kayıtları sınıflandırmak için kullanılan istatistiksel bir tekniktir. Giriş alanı değerlerini çıkış alanı

olasılıklarıyla ilişkilendiren denklemler oluşturur. SPSS Modeler 18'de yer alan Lojistik Regresyon sınıflandırıcısı, özellikle ikili veya çoklu sınıflandırma problemleri için kullanılan güçlü bir istatistiksel modelleme aracıdır. Temel amacı, bir veya daha fazla bağımsız değişken arasındaki ilişkiyi kullanarak bir kategorik bağımlı değişkenin (genellikle ikili; evet/hayır, başarılı/başarısız gibi) sonuçlarını tahmin etmektir. Nominal regresyon olarak da bilinen lojistik regresyon, doğrusal regresyona benzer ancak sayısal bir alan yerine kategorik bir hedef alan alır. Hem binom modeller (iki ayrı kategoriye sahip hedefler için) hem de multinom modeller (ikiden fazla kategoriye sahip hedefler için) desteklenir (IBMCorporation, 2022). Ayırma analizine (Diskriminant Analizi) alternatif olan LR modelleri, sembolik ve sayısal girdi alanlarını işleyebilir ve tüm hedef kategoriler için öngörülen olasılıkları sağlayabilir (Ceyhan, 2020). Ayrıca diğer modelleme teknikleri için de bir temel olarak kullanılabilir.

Lojistik regresyon, girdi alanı değerlerini çıktı alanı kategorilerinin her biriyle ilişkili olasılıklarla ilişkilendiren bir dizi denklem oluşturarak çalışır. Model oluşturulduktan sonra, yeni veriler için olasılıkları tahmin edilmeye çalışılır. Her bir kayıta yönelik, olası her bir çıktı kategorisi için bir üyelik olasılığı hesaplanır. En yüksek olasılığa sahip hedef kategori, o kayıt için tahmin edilen çıktı değeri olarak atanır (IBMCorporation, 2022).

Lojistik regresyon, bağımlı değişkenin log-odds'ini (logit dönüşümü) bağımsız değişkenlerin doğrusal bir kombinasyonu olarak modelleyerek çalışır (Tabachnick & Fidell, 2013). Bağımsız (yordayıcı) değişkenler $X = (X_1, X_2, X_3, \ldots, X_n)$ biçiminde ifade edilirse temel lojistik denklemi;

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

olacaktır. Yukardaki temel lojistik denkleminden yola çıkarak oluşturulan modelde p olayın olasılığıdır ve $\beta_0, \beta_1, \ldots, \beta_n$ model parametreleridir.

Lojistik fonksiyon, verilen tahmin (bağımsız) değişkenlerine dayalı olarak olayın gerçekleşme olasılığını hesaplamak için:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

denklemini kullanır. Model parametreleri (beta katsayıları), genellikle Maksimum Olabilirlik Estimasyonu (Maximum Likelihood Estimation, MLE) kullanılarak tahmin edilir. Beta katsayıları, bağımsız değişkenlerin lojistik olasılık üzerindeki göreceli etkisini ifade eder. Bir beta katsayısının pozitif olması, ilgili değişkenin olayın olasılığını artırdığını gösterirken,

negatif bir katsayı tersi bir etkiyi işaret eder. MLE, gözlemlenen verilerin olasılığını maksimize eden parametre değerlerini bulmayı amaçlar (IBMCorporation, 2022). Lojistik regresyon analizinde, model tahmininde en küçük kareler (ordinary least square) yöntemi yerine, en yüksek olabilirlik (maximum likelihood) yönteminden yararlanılmaktadır. Lojistik regresyon ise, lojistik dönüşüme gerektiren doğrusal olmayan (nonlinear) niteliğinden ötürü, en yüksek olabilirlik yönteminden yararlanmaktadır. Lojistik regresyon sapmaların karesini (en küçük kareler) en aza indirmek yerine, bir olayın gerçekleşme ihtimalini en yüksek yapmaya odaklanır (Cemaloğlu & Duykuluoğlu, 2020).

Modelin uygunluğu, çeşitli istatistiksel testler ve metrikler (örneğin, Wald testi, Hosmer-Lemeshow uyum iyiliği testi) ile değerlendirilebilir (Tabachnick & Fidell, 2013). Ayrıca, sınıflandırma modellerinde genel doğruluk-hata oranı ve modelin tahmin performansını ölçmek için çapraz doğrulama, ROC eğrisi (Receiver Operating Characteristic Curve) ve AUC (Area Under the Curve) gibi metrikler kullanılır.

Lojistik regresyon modeli araştırmadaki deneklerin hangi grubun üyesi olduğunu tahmin etmeye yarayan bir modeldir. Bu bakımdan lojistik regresyon analizi (LRA) yapmanın iki amacı vardır; birincisi yordanan değişkende grup üyeliklerini belirleyen yordayıcı değişkenlerin belirlenmesi, ikincisi grup üyeliklerini en iyi performansla belirleyen lojistik modelin geliştirilmesidir (Hair vd., 2014; akt. Kasap, (2022)).

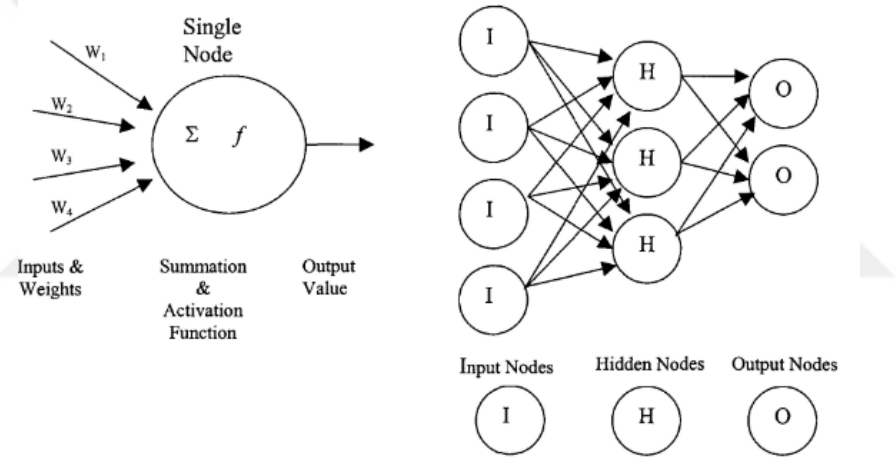
Lojistik regresyon, risk değerlendirme, müşteri terk analizi, hastalık olasılığı tahmini gibi konularda tıptan pazarlamaya, finans sektöründen sosyal bilimlere kadar geniş bir uygulama alanına sahiptir ve sıkça kullanılır.

2.2.1.1.9. Yapay Sinir Ağları -NeuralNet

İlk yapay sinir ağları insan beyninin hesaplama yeteneğinden esinlenerek, elektrik devreleriyle basit bir sinir ağı modeli kurmak isteyen ve sinir hekimi olan Warren McCulloch ile matematikçi Walter Pitts tarafından 1943 yılında geliştirilmiştir (Kabalcı, 2013). Yapay sinir ağları (YSA); insan beyninden esinlenerek, öğrenme sürecinin matematiksel olarak modellenmesi uğraşı sonucu ortaya çıkan bir işleme modelidir. Dolayısı ile insana özgü öğrenme yolu ile yeni bilgiler türetebilme, yeni bilgiler oluşturabilme ve keşfedebilme gibi yetenekleri, herhangi bir yardım almadan otomatik olarak gerçekleştirebilmeyi amaçlar.

SPSS Modeler'da ise " NeuralNet " olarak isimlendirilen YSA algoritması, veri madenciliği ve makine öğreniminde sınıflandırma ve regresyon gibi çeşitli tahmin görevlerinde kullanılır. NeuralNet düğümü, insan beyninin bilgi işleme şeklinin basitleştirilmiş bir modelini kullanır. Buna göre, nöronların soyut versiyonlarını andıran çok sayıda birbirine bağlı basit işleme ünitelerinin simülasyonunu yaparak işe yarar tahminler yapar (IBMCorporation, 2022). Bu sebeple model eğitim ve tahminin yapılmasında çok az istatistiksel veya matematiksel bilgiye ihtiyaç duyar (SPSSModeler, 2020).

NeuralNeT, genellikle karmaşık desenleri tanıma ve tahmin etme kapasitesi nedeniyle tercih edilir. SPSS Modeler'da çok katmanlı algılayıcılar (multi-layer perceptrons, MLP) ve geri yayılım (backpropagation) yöntemi bu algoritmanın temelini oluşturur (SPSSModeler, 2020). Sınıflandırma problemlerinde, çıktı katmanında her sınıf için bir düğüm bulunur ve ağ, verilen girdilere bağlı olarak hangi sınıfa ait olduğunu tahmin eder (Şekil 14).



Şekil 14. Sinir ağı düğüm yapısı. Schumann, J. A. (2005). *Data mining methodologies in educational organizations.* Doktora Tezi, University of Connecticut, United States.

Şekil 14'te görüldüğü gibi model yapısında giriş katmanı (Input Node), veri setindeki özelliklerin (features) her biri için bir nöron içerir. Özniteliklerin önem dereceleri, hangi değişkenlerin model üzerinde daha etkili olduğunu ve hangilerinin az etkili olduğunu gösterir. Bu bilgiler, modeli optimize etmek veya gerekirse girdileri yeniden düzenlemek için kullanılabilir.

Giriş katmanından alınan bilgileri işleyen ve öğrenme sürecinde önemli olan dönüşümleri gerçekleştiren bir veya daha fazla katmana ise gizli katman(lar) denir (Altun, 2019). Gizli

katmanda genellikle sigmoid veya ReLU (Düzeltilmiş Doğrusal Birim) gibi non-lineer aktivasyon fonksiyonları kullanılır (Schumann, 2005).

Çıktı Katmanında ise her bir sınıf için bir çıkış nöronu vardır ve sınıflandırma işlemini gerçekleştirir. Sınıflandırma için genellikle softmax fonksiyonu kullanılır, bu sayede her bir çıkış sınıfının olasılıkları hesaplanır (Çırak, 2012).

Bir sinir ağındaki bilgi, düğümler arasındaki bağlantıların güçlerinde (ağırlık olarak adlandırılır) kodlanır (IBMCorporation, 2022). Ağdaki hata, gerçek çıktılar ile tahmin edilen çıktılar arasındaki farka dayanarak hesaplanır ve bu hata, ağın her bir katmanındaki ağırlıkların (weights) ayarlanması için kullanılır (Han vd., 2011).

Modelin yapılandırılması, kullanılan aktivasyon fonksiyonları ve özelliklerin önemi, modelin hangi verilere dayanarak tahmin yaptığını ve bu tahminlerin nasıl şekillendiğini anlamada kritik öneme sahiptir (Ceyhan, 2020; Kayri, 2015).

Bu çalışmada yer aldığı şekilde gizli katmanlardaki hiperbolik tanjant ve çıkış katmanındaki softmax kullanımı, modelin karmaşık ilişkileri öğrenme ve çok sınıflı sınıflandırma problemlerinde etkili olabileceğini gösterir (Çırak, 2012; IBMCorporation, 2022).

Bir NeuralNet, minimum varsayımlar ile çeşitli tahmin modellerine yaklaşabilir. İlişki biçimi öğrenme yoluyla oluşturulur ve doğrusal ilişkiler için yakın bir yaklaşım elde edilir. Doğrusal olmayan ilişkilerin uygun olduğu durumlarda, sinir ağı doğru model yapısını otonom olarak uyarlar (SPSSModeler, 2020). Bununla birlikte, modelin önemli bir dezavantajı; yorumlama kolaylığından yoksundur, bu da geleneksel istatistiksel modelleri altta yatan süreçleri açıklamak için tercih edilir hale getirir (Han vd., 2011) Ancak NeuralNet algoritmasını kullanarak karmaşık veri yapılarını modelleme ve çeşitli sınıflandırma görevlerinde etkili sonuçlar elde etme imkânı oluşur. Bu, özellikle yüksek boyutlu verilerde veya geleneksel istatistiksel yöntemlerle modellenmesi zor olan veri setlerinde faydalıdır.

2.2.1.1.10. Destek Vektör Makineleri (DVM)

1960'larda bilgisayarlı öğrenme teorileri üzerinde çalışan Vladimir Vapnik ve meslektaşı Alexey Chervonenkis, Vapnik–Chervonenkis (VC) Teorisi adını verdikleri istatistiksel bir öğrenme teorisi geliştirmişlerdir. VC Teorisinden yola çıkarak 1990'ların başında Boser, Guyon ve Vladimir Vapnik, DVM'lerin teorik temellerini ve matematiksel çerçevesini oluşturan çalışmaları yapmıştır (Akpınar, 2017). DVM hem sınıflandırma hem de regresyon analizinde

kullanılabilmesine rağmen sıkılıkla sınıflandırma işlemleri için kullanılan bir denetimli öğrenme modelidir (Metlek & Kayaalp, 2020). Makine öğrenimi alanının temel algoritmalarındandır.

DVM, sınıflandırma stratejisinde istatistiksel bir yaklaşımdan ziyade marj tabanlı geometrik bir ölçüt uygular; dolayısı ile sınıfların istatistiksel dağılımlarının tahmin edilmesine ihtiyaç duymadan, marj maksimizasyonu kavramını kullanarak sınıflandırma modelini tanımlarlar (Metlek & Kayaalp, 2020). Tüm lineer modellerde olduğu gibi, DVM'ler de iki sınıf arasındaki karar sınırı olarak ayırıcı hiperdüzlemleri bulmaya çalışır (Delen, 2010). Bu ayrıcı çizgi, iki sınıfa ait veri noktalarını en geniş marjla ayırır ve hiper düzlemi belirlemek için en yakın veri noktalarını (destek vektörleri) kullanır.

DVM'de verilerin yapısına göre (tam ayrıştırılabilir ya da ayrıştırılamaz olması) sınıflandırmada kullanılacak düzlemlerin yapısı ve sayısı belirlenir. Eğer sınıflandırma işleminde, tam ayrıştırılabilir veriler kullanılırsa genellikle tüm veriler bir hiper düzlemle sınıflandırılabilir; aksi halde aynı boyutta tek bir düzlemle sınıflandırma yapılamaz (Metlek & Kayaalp, 2020). Algoritmanın çalışmasında veri türüne göre çekirdek fonksiyonlar da kullandığı için bir doğru ile ayrılamayan sınıfların boyutunu artırarak bir düzlem vasıtası ile ayırmak mümkün olabilmektedir. Bu durumda çekirdek fonksiyonlardan yararlanılır ve literatürde bunu için Doğrusal, Polinom, Gauss, Laplace, Anova, Sigmoid, Bessel çekirdek fonksiyonlarının kullanıldığı belirtilmektedir (Ceyhan, 2020). DVM'lerde kullanılan verilerin birbiri içinden ayrılıp ayrılmaması, verileri sınıflandırmak için kullanılan düzlemlerin nasıl ve kaç adet olacağını doğrudan etkilemektedir. Temelde ikili sınıflandırmalar için kullanılan model için sonrasında çok boyutlu bağımlı değişkenlere yönelik uygulamalar geliştirmiştir. Buna göre DVM sınıflandırma işlemi; sert marjin lineer sınıflandırma (tamamıyla doğrusal ayrılabilen), soft marjin lineer sınıflandırma ve lineer olmayan sınıflandırma olmak üzere üç gruba ayrılır (Ceyhan, 2020).

DVM, verileri aşırı uyum sağlamadan iki gruptan birine sınıflandırmayı amaçlar (SPSSModeler, 2020). İkili sınıfları ayrıştırmanın temel prensibi; bir düzlem yardımı ile etiketli veri noktalarını iki sınıfa ayırmaktır. Ancak literatürde modelin bazı dezavantajlarından da bahsedilmektedir: bunlardan ilki iki boyutlu sınıf problemleri ve sonuçları için tasarlanmış olması ve sonraki sınıf üyelik olasılıklarını doğal bir biçimde modelleyememesidir. İkinci bir sınırlama hem eğitim hem de test verisinde hız ve boyuttur. Karmaşık ve zaman gerektiren hesaplamalar içerir. DVM'lerle ilgili belki de en ciddi sorun, büyük ölçekli görevlerde yüksek

algoritmik karmaşıklık ve geniş bellek gereksinimleridir (Ceyhan, 2020). Özellikle yüksek boyutlu veri kümeleri için etkili bir sınıflandırma ve regresyon aracı olarak ideal sonuçlar ürettiği gösterilse de model kullanılırken sınırlılıkları da göz önünde bulundurulmalıdır. DVM'ler literatürde yüz algılama, metin-görüntü sınıflandırma, biyoinformatik, protein çaprazlama, uzaktan homoloji tespiti, el yazısı tanıma, jeoloji ve çevre bilimleri, genelleştirilmiş tahmine dayalı kontrol uygulamaları gibi birçok alanda yaygın olarak kullanılmaktadır (Metlek & Kayaalp, 2020).

2.2.1.2. Kümeleme (Clustering)

Kümeleme, verileri benzerliklerine göre gruplara ayırma işlemidir. Bu yöntem, veri setindeki doğal gruplamaları bulmaya yarar ve herhangi bir önceden tanımlanmış kategoriye ihtiyaç duymaz. Müşteri segmentasyonunun belirlenmesi bu yöntem için tipik bir uygulamadır. Kümelemede bağımlı değişkenin önden belirlenmiş bir etiketi bulunmadığı için denetimsiz öğrenme yaklaşımı içinde değerlendirilir.

2.2.1.3. Birliktelik Kuralı- İlişki Analizi (Association Rule Learning)

İlişki analizi, veri öğeleri arasındaki ilginç ilişkileri ve sık sık birlikte görülen öğeleri belirleme sürecidir. Market sepeti analizi, bu türün en bilinen örneklerindendir; burada hangi ürünlerin sıklıkla birlikte satın alındığı analiz edilir.

2.2.1.4. Zaman Serisi Analizi (Time Series Analysis)

Zaman serisi analizi, zamanla değişen veri noktalarını inceleme sürecidir. Bu tür özellikle daha çok finansal piyasalar, hava durumu tahminleri ve envanter çalışmalarında kullanılır.

2.3. İlgili Araştırmalar

Bu başlık altında bu araştırmanın temel amacına uygun olarak alanda yapılan diğer araştırmalar yakından uzağa ilişkili olacak biçimde aşağıdaki başlıklarda sınıflandırılarak sunulmuştur;

i. Açık ve uzaktan öğrenmede veri madenciliği ile yapılan okul terki veya devamlılık araştırmaları

ii. Eğitimde veri madenciliği ile yapılan okul terki veya devamlılık araştırmaları
iii. Açık ve uzaktan öğrenmede (yükseköğrenimde) okul terki ve devamlılık kavramına odaklı araştırmalar
iv. Eğitimde veri madenciliği çalışmaları; öğrenci başarı ve performans sınıflandırmaları

2.3.1. Açık ve Uzaktan Öğrenmede Veri Madenciliği ile Yapılan Okul Terki veya Devamlılık Araştırmaları

Grau-Valldosera (2019) çevrimiçi uzaktan eğitimde öğrenenlerin devamlılık niyetleri ile terk durumlarını tanımlamak amacıyla okulu bırakma olgusunun devam etme olgusuyla paralel gittiğini göz önünde bulundurarak, öğrencilerin 2. dönemdeki devam etme niyetlerinden 3. dönemde yeniden kayıt olarak bu niyetin daha da doğrulanmasına kadarki yola odaklanan boylamsal bir analiz gerçekleştirmiştir. Katalonya Açık Üniversitesinde (UOC) yürütülen araştırmada bir anket aracılığı ile veriler toplanmış; toplanan veriler ile resmi okul terki tanımlarının çevrimiçi (veya uzaktan/açık) kurumlardaki okul terklerinin gerçek doğasını yakalayamadığı vurgulanmıştır. Derinlemesine kayıt verilerinin analizi ile program düzeyinde terk tanımı ve öğrencilerin kayıt ve ayrılma davranışlarının nesnel bir analizinin erken dönemde elde edilebileceği sonucuna varılmıştır. Ayrıca elde edilen sonuçlara göre, olumlu devam niyetine sahip öğrencilerde özellikle yaş ve önceki üniversite deneyiminin sıklığı etkili bir şekilde anlamlı görünmektedir.

Kuzilek vd. (2015) açık üniversitede risk altında bulunan öğrencilerin analizini yaptıkları araştırmalarında demografik veriler ile zenginleştirilmiş bir veri seti ile risk altındaki öğrencileri kurstaki ilk değerlendirmeden önce erken tespit edebilmeyi amaçlamıştır. Araştırmanın amacına uygun olarak iki tür veri kullanılmaktadır: demografik (statik) veriler ve Sanal Öğrenme Ortamı sistemindeki öğrenci etkileşim verileri. Bayesian yaklaşımı kullanılarak Sanal Öğrenme Ortamındaki en ilgili etkinlik türleri (örneğin önemli kaynaklara erişim, öğrenci forumu vb.) seçilmiştir. Her hafta, demografik verilerle birlikte, seçilen etkinlik türleri ile dört tahmin modeli (demografik veriler ve sanal öğrenme ortamı verileri kullanılarak ayrı ayrı çalıştırılan k-En Yakın Komşular algoritması; Sınıflandırma ve Regresyon Ağacı ve hem sanal öğrenme ortamı, hem de demografik verileri kullanan olasılıksal naive Bayes modeli) oluşturmuştur. Çalışmanın sonucunda kurs eğitmenlerine model ile öngörücü bilgilerin sağlanabildiği sonucuna ulaşılmıştır. Bun göre 0. haftada aktif olmayan öğrencilerin %73'ünün X dersinden başarısız olduğu ve bu başarısızlık oranının ilk dört haftanın tamamında aktif

olmayan öğrenciler için %86'ya çıktığı gösterilmektedir. Araştırmada terk tahmini için doğru zamanın tespiti ve değişen kurs tasarımlarının etkisi sebebiyle oluşan zorluklar tartışılmıştır.

Singer (2023) doktora tezini çok aşamalı karma yöntem eylem araştırması olarak yürütmüş; birinci aşamada bir kamu üniversitesinin çevrimiçi kampüsünde kabul edilen başvuru sahiplerinin (n=3.843) kayıt kararlarını anlamak ve tahmin etmek için eğitimde veri madenciliği ve makine öğrenimini araştırmıştır. İkinci aşamada ise üniversitenin kayıt uzmanlarının çevrimiçi öğrenci kayıt kararlarını etkileyen faktörleri nasıl yorumladıklarını ve bunlardan nasıl etkilendiklerini ele almıştır. Önceki araştırmaların ışığında analizlerde en başarılı yöntem olarak belirtilen lojistik regresyon uygulanmıştır. Araştırmada 24 bağımsız değişken değerlendirilerek kayıt kararında etkili 10 değişken (eyalet ikametgahı, başvuru ana dalı, derece adı, birinci nesil statüsü, öğrenci tipi, toplam transfer kredisi, kabul edildiği ay, haftanın kabul edilen günü, derece türü, kabul geri dönüş süresi-gün, başvuru ayı) belirlenmiştir. Tartışma bölümünde, 24 bağımsız değişken içinde yer alan akademik ortalama, cinsiyet, yaş, etnik köken, en son gittiği okul, posta kodu faktörlerinin fiziksel kampüslerde (örgün eğitim) etkili olabileceği ancak online eğitimde yürütülen bu çalışmada modelde anlamlı bulunmadığı belirtilmiştir. ABD'de her yıl giderek artan yükseköğrenim terki veya kayıt problemine odaklanan çalışmada makine öğreniminin, öğrenci kaybını azaltmak amacıyla risk altındaki öğrencileri belirlemek ve onlara müdahale etmek için uygulanabilir olduğu sonucuna ulaşmıştır.

Kotsiantis vd. (2003) makine öğrenimi teknikleri kullanılarak uzaktan eğitimde öğrenci terkinin önlenmesi amacıyla yayımladıkları kitaplarında uzaktan eğitim veren üniversitelerde oldukça sık görülen öğrenci terkini tahmin etmek için en uygun öğrenme algoritmasını belirlemeye çalışmıştır. Bu alandaki en yaygın altı makine öğrenme tekniği olan Karar Ağaçları, Sinir Ağları, Naive Bayes algoritması, Örnek Tabanlı Öğrenme Algoritmaları, Lojistik Regresyon ve Destek Vektör Makineleri kullanılmıştır. Hellenic Open University'nin 'informatics' kursu tarafından sağlanan verilerle bir dizi deney gerçekleştirilmiş ve Naive Bayes algoritmasının başarıyla kullanılabileceği sonucuna varılmıştır. Algoritmanın doğruluk değeri, sadece öğrencilerin demografik verilerine dayanan ilk tahminlerde %63'e ulaşmakta ve akademik dönemin ortasından önce %83'ü aşmaktadır. Dolayısı ile ilerleyen müfredat dönemlerinde veriye eklenen diğer özellikler tahmin doğruluğunu arttırmaktadır. Ayrıca Naive Bayes algoritması uygulanarak, okulu bırakma olasılığı yüksek olan öğrencileri otomatik olarak tanıyan web tabanlı prototip bir destek aracı da oluşturulmuştur.

Yu vd. (2021) hem yatılı hem de tamamen çevrimiçi öğrencileri olan büyük bir ABD araştırma üniversitesinde birden fazla yıla ait tüm öğrenci nüfusu için kapsamlı kurumsal kayıtlara dayanarak (N = 93.457), bir akademik eğitim yılından sonra öğrenci terkini tahmin etmek için makine öğrenimi modelleri ile korumalı öznitelik olarak tanımladıkları (cinsiyet, ırk, birinci nesil öğrenci ve yüksek mali ihtiyaç) değişkenlerin model tahminlerinde genel performansını ve adaletini karşılaştırmışlardır. "Korunan nitelikler üniversite terk tahmin modellerine dahil edilmeli midir?" sorusundan yola çıkarak ampirik inceleme sunulan çalışmada iki tip yükseköğretim için korunan nitelikleri içeren ve içermeyen tahmin modelleri sistematik olarak incelenmiştir. Araştırma sonucunda korunan niteliklerin dahil edilmesinin yaygın olarak kullanılan özellikler (kayıt bilgileri, akademik kayıtlar) zaten modelde yer alıyorsa, genel tahmin performansını etkilemediği ve tahminlerin algoritmik adaletini sadece marjinal olarak iyileştirdiği vurgulanmaktadır.

Adnan vd. (2021) temel demografik değişkenler ve e-öğrenme kursunun ilerleyen zamanlarından elde ettikleri öğrenme aktivitesi verileri ile birden fazla zaman dilimi için risk altındaki öğrencileri Rastgele orman, DVM, K-NN, AdaBoost sınıflandırıcı, Gradient boosting sınıflandırıcı ve YSA modelleri ile belirlemeye çalışmışlardır. Öğrencilerin performansı en başta sadece demografik değişkenler mevcutken tahmin edilmiş ve kurs süresinin %20'sinde bile, Rastgele orman tahmin modeli %79 ortalama hassasiyet puanı, %79 ortalama hatırlama puanı, %79 ortalama f puanı ve %79 ortalama doğruluk puanı ile umut verici sonuçlar üretilmiştir. Diğer zaman dilimlerinde eklenen veriler ile de genel olarak, Rastgele orman modelinin sonuçları, risk altındaki öğrencilerin mümkün olan en erken tahmininde etkili olduğunu göstermişlerdir.

Uzaktan öğrenmenin trendlerinde kitlesel çevrimiçi dersler (MOOC) de önemli bir alan kaplamaktadır. Qiu, Liu & Liu (2018) kitlesel açık çevrimiçi kursları terk eden öğrenciler sorununu çözmek için yürüttükleri çalışmalarında, MOOC'larda bırakmayı tahmin etmek için özellik oluşturma, özellik seçimi ve bırakma tahminini içeren özellik seçimi (FSPred) ile entegre bir çerçeve önermektedir. Klasik yaklaşımın aksine (manuel tespit edilen ve tahmine etkisinin garanti edilmesi zor özellikler yerine) problemi klasik bir ikili sınıflandırma problemi olarak kabul edip, öğrenme ve tahmini gerçekleştirmek için destek vektör makinesi ve lojistik regresyon kullanmaktadır.

Benzer şekilde Liang vd. (2016), Edx'den elde ettikleri öğrencilerin bir aylık öğrenme davranışı günlük verilerini içeren büyük bir veri ile MOOC'larda öğrencilerin önlerindeki on gün içinde

okulu bırakıp bırakmayacaklarını tahmin etmek üzerinde çalışmışlardır. Açık kaynak kodlu Edx platformunu temel alan XuetangX platformundan 39 kurs verisi üzerinde veri çıkarma, veri ön işleme, özellik mühendisliği ile DVM, Lojistik Regresyon, Rastgele Orman ve Gradyan Güçlendirme Karar Ağacı (Gradient Boosting Decision Tree) gibi veri madenciliğindeki dört sınıflandırma modeli eğitilmiş ve çeşitli denetimli sınıflandırma modellerindeki performans testiyle okulu bırakma tahmini yaklaşımında en iyi sonucun Gradyan Güçlendirme Karar Ağacı modeli ile (%88 doğruluk) elde edildiği bulunmuştur.

Makine öğrenimi tekniklerinin kombinasyonu ile e-öğrenme kurslarında öğrenci terkini tahmin etmeye çalışan Lykourentzou vd. (2009) e-öğrenme kursları için üç popüler makine öğrenme tekniğine ve ayrıntılı öğrenci verilerine dayanan bir terk tahmin yöntemi önermiştir. Kullanılan makine öğrenme teknikleri, ileri beslemeli sinir ağları, destek vektör makineleri ve olasılıksal topluluk ile basitleştirilmiş bulanık ARTMAP'dir (Fuzzy Adaptive Resonance Theory ART-MAP). Çalışmada, üç makine öğrenimi tekniğinin sonuçlarını farklı şekillerde birleştiren üç karar şeması da test edilmiştir. Önerilen yöntemin, iki dersin ilk bölümünden itibaren %75-85 genel öğrenci sınıflandırma oranına; son bölümlerde ise %97-100'lük bir orana ulaştığı görülmüştür. Yöntem genel doğruluk, duyarlılık ve kesinlik açısından incelenmiş ve sonuçları ilgili literatürde bildirilenlerden önemli ölçüde daha iyi bulunmuştur.

MOOC'ların bırakılması sorununu çözmede kullanılan makine öğrenimi araştırmalarına kapsamlı bir genel bakış sunmayı amaçlayan Dalipi vd. (2018), literatürde yaygın olarak kullanılan farklı MOOC bırakma tahmin mimarilerini araştırmıştır. Bu makale, yüksek sayıda okul terkine yol açan hem öğrenciyle ilgili faktörleri hem de MOOC ile ilgili faktörleri vurgulamaktadır. Makale ayrıca, öğrenci terki tahmini ile ilgili bazı kritik zorlukları tanımlamakta ve çeşitli makine öğrenimi tekniklerini kullanan araştırmacılara zamanında ve verimli bir şekilde çözmelerinde yardımcı olacak öneriler ve teklifler sunmaktadır.

Ayrıca makalede makine öğrenimi kullanarak okulu terki tahmininde karşılaşılan zorluklar;

- Tıklama akışlarında veri standardizasyonu
- Öğrenci tarafından sağlanan verilerin toplanabilmesi ve güvenliği
- Özellik mühendisliği teknikleri (yapılandırılmamış veri ve özellikler de kullanılabilir)
- Kayıt-katılım sistemi
- Modellerin ve tahmin edicilerin değerlendirilmesi
- Öğrencilerin sosyal etkileşim verilerinin geliştirilmesi başlıkları altında toplanmıştır.

Bu makale, çeşitli öğrenme nesneleri meta veri standartlarına benzer şekilde, bir tıklama akışı verilerinin ve öğrenci tarafından sağlanan verilerin bir standart olarak birleştirilmesi fikrini ortaya koymaktadır. Çalışmanın sonuç bölümünde terk nedenleri, öğrenciyle ilgili faktörler (öğrencinin motivasyon eksikliği, zaman eksikliği, eski öğrenim ve becerilerin yetersizliği) ve MOOC ile ilgili faktörler (kurs tasarımındaki özellikler, MOOC'larda izolasyon hissi ve etkileşim eksikliği, gizli maliyetler) olarak özetlenmiştir.

Okulu bırakmanın bir kullanıcının etkileşim kalıplarıyla derinden ilişkili olduğu analizine dayanarak, Pulikottil ve Gupta (2020) yaptıkları araştırmada MOOC'larda kullanıcı terkini tahmin etmek için 'dikkat meta gömme tabanlı bir derin zamansal ağ' önermektedir. KDDCUP2015 ve XuetangX'tan elde edilen iki farklı veri kümesi üzerinde temel kıyaslamalar dört farklı algoritma (Lojistik Regresyon, Rastgele Orman, Doğrusal Destek Vektör Makinesi, Gradyan Güçlendirme Makinesi) ile gerçekleştirilmiştir. Araştırma metodolojisinde MOOC bırakma oranının, öğrenme etkinliklerinde kendini gösteren öğrencinin katılımıyla doğrudan ilişkili olduğu varsayımı ile yola çıkılmıştır. Yapılan çalışma öğrencilerin etkileşimlerinin türü, zamansallığı ve göreceli hızı gibi çeşitli özelliklerinin analizi ve görselleştirilmesi yoluyla öğrenci terki ile ilişkili faktörler hakkında yeni bilgiler sunmaktadır. Bunun yanında öğrencilerle olan etkileşimlerden otomatik olarak öğrenilen özellikleri kullanarak öğrenci terki tahmini için yeni bir derin dikkat ağı (ONet) önermektedir.

Moldovan vd. (2019) 2008 yılında başlayan ancak yıllar içinde yüksekokul terki ile başarısızlık oranlarından mustarip MOOC'larda, derin öğrenme ve makine öğrenimi teknikleri kullanarak bırakma ile devam eden öğrencilerde başarı tahmini yapmıştır. Araştırma demografik bilgiler, değerlendirme bilgileri ve VLE ile etkileşim gibi farklı öznitelik türlerine dayalı olarak oluşturulan makine öğrenimi modellerinin hem öğrenci terkini hem de sonuç tahmin performansını inceleyerek alana katkıda bulunmaktadır. Dersi bırakma ve nihai sonucu tahmin etmek için; Dağıtılmış Rastgele Orman (Distributed Random Forest-DRF), Gradyan Güçlendirme Makinesi (Gradient Boosting Machine-GBM), Derin Öğrenme (Deep Learning-DL), Genelleştirilmiş Doğrusal Model (Generalized Linear Model-GLM) makine öğrenimi sınıflandırma algoritmaları dikkate alınmıştır. Açık Üniversite (Open University UK) tarafından sağlanan OULAD veri setini kullanan bu çalışmanın sonucunda öğrencinin, Sanal Öğrenme Ortamları ile etkileşimine dayalı makine öğrenimi modellerinin, Gradyan Güçlendirme Makinesi durumunda okulu bırakma tahmini için 0,91 ve sonuç tahmini için 0,93'e varan AUC açısından yüksek performans elde ettiği gösterilmiştir.

Tamada vd. (2019) tarafından yapılan bir araştırma, Sanal Öğrenme Ortamlarında okulu bırakma oranını azaltmanın yollarını öneren metodolojilerin sistematik bir incelemesini sunmaktadır. Amaçları, proaktif eylemleri teşvik eden ve sadece okul terkini tespit eden algoritmaları analiz etmekle sınırlı olmayan tahminler için makine öğrenimi tekniklerini kullanarak uzaktan eğitimde öğrenci terk oranlarını azaltmaya yönelik öneriler ortaya koymaktır. Akademik veri tabanlarında yapılan ilk arama sonucunda 199 makale elde edilmiş ve bunlardan 13'ü çalışmada nihai analize dahil edilmiştir. İnceleme, yayınların tarihsel gelişimini, kullanılan Makine Öğrenimi tekniklerini, kullanılan verilerin özelliklerini ve uzaktan eğitimde okul terkini azaltmak için önerilen çözümleri raporlamaktadır. Birçok çözüm, okulu bırakma tahmini için denetimli makine öğrenimi tekniklerini kullanmıştır. Lojistik regresyon ve destek vektör makineleri en yaygın kullanılan tekniklerdir ve makalelerin öğrenci niteliklerine, tıklama akışı verilerine, forum katılımına ve diğer kaynaklara dayalı olarak Sanal Öğrenme Ortamlarında öğrenci terkini tahmin etmede yüksek doğruluk sunduğu sonucuna varılmıştır.

Kitlesel Açık Çevrimiçi Kurslarda (MOOC) süregelen önemli sorunlardan biri olan öğrencilerin bu kurslardan ayrılma durumuna odaklanan Dominic Ravichandran (2021) araştırmasında, Arizona Eyalet Üniversitesi'ne ait MOOC platformu için Open edX'te sunulan matematik kursu 'College Algebra and Problem Solving'den 2016-2020 yılları arasında elde edilen verilerden yola çıkarak öğrencilerin bir gününe ait bir dizi özellikle kursu bırakma durumlarını tahmin etmeyi amaçlamaktadır. Kullanılan Makine Öğrenimi modeli Rastgele Orman olup bu model doğruluk, kesinlik, geri çağırma, F1-skoru, Eğri Altındaki Alan (AUC), Alıcı İşletim Karakteristiği (ROC) eğrisi gibi doğrulama metrikleri kullanılarak değerlendirilmiştir. Geliştirilen ve önerilen model, MOOC kursunda herhangi bir günde öğrencilerin okulu bırakma veya devam etme durumlarını sırasıyla %87,5 doğruluk, %94,5 AUC, %88 kesinlik, %87,5 hatırlama ve %87,5 F1-skoru ile tahmin edebilmektedir. Katkıda bulunan özellikler ve etkileşimler, modelin tahmini için Shapely değerleri kullanılarak açıklanmıştır.

Kayhan (2019) Uzaktan Eğitim Öğrencilerinin Mezuniyet Durumlarının Veri Madenciliği Yöntemleri ile Tahmini: Amasya Üniversitesi Örneği adlı çalışmasında; Amasya Üniversitesi uzaktan eğitim Çocuk Gelişimi, Tıbbi Dokümantasyon ve Sekreterlik, İnternet ve Ağ Teknolojileri, Mekatronik ve Elektrik Programlarına 2016-2017 güz döneminde kayıt yaptıran ön lisans öğrencilerinin zamanında mezun olabilme durumlarına ilişkin çıkarımlarda bulunmaya çalışmıştır. Tahmin için Karar ağacı, Naive Bayes, Destek vektör makinesi,

Rastgele orman ve Yapay sinir ağları algoritmaları kullanılmıştır. Algoritmalar ile yapılan analiz sonucunda oluşan sınıflandırma performansı ölçütlerine, program bazlı karşılaştırılmalı olarak çalışmada yer verilmiştir. Araştırmada Çocuk Gelişimi bölümünde DVM modelinin; diğer bölümlerde ise karar ağaçları algoritmasının en iyi tahmini yaptığı sonucuna ulaşılmıştır.

2.3.2. Eğitimde Veri Madenciliği ile Yapılan Okul Terki veya Devamlılık Araştırmaları

Yükseköğretimde öğrenci kayıplarının önemli bir problem olduğundan yola çıkarak halihazırda kayıtlı bulunan öğrencileri elde tutmak için makine öğrenimi tekniklerinin karşılaştırmalı bir analizini yapan Delen (2010) önceki araştırmaların aksine terk ve devam durumunun olgusal olarak anlaşılmasından ziyade olası başarısız veya terk edebilecek öğrencileri en iyi şekilde belirleyebilecek yönteme odaklanmıştır. Çalışmasında bir üniversiteye ait beş yıllık kurumsal verilerle çeşitli veri madenciliği teknikleri (hem bireysel hem de topluluklar) kullanarak, birinci sınıf öğrenci kaybının arkasındaki nedenleri tahmin etmek ve açıklamak için analitik modeller geliştirmiştir. Delen, araştırmada öğrenci kaybı sorununun kuruma özgü doğasını açıklamanın yanı sıra terki tahmin edebilen modeller geliştirmek için nicel bir araştırma yaklaşımı önermekte ve bunun yanında veri madenciliği sınıflandırma algoritmalarının çarpık veri setlerindeki performanslarını karşılaştırmaktadır. Yapılan iki deneysel çalışmanın ilkinde verinin tümü olduğu gibi kullanılırken (yani, ~%80 "Evet" ve ~%20 "Hayır" örnekleri); ikincisinde iki sınıfın eşit olarak temsil edildiği iyi dengelenmiş bir veri seti kullanılmıştır. Yani ilk deneydeki 16.066 kayıttan oluşan orijinal veri seti yerine ikincisinde örnekleme işlemlerinin her biri, 3509'u "Hayır" ve 3509'u "Evet" olarak etiketlenen 7.018 kayıttan oluşan bir veri kümesi oluşturulmuştur. Sonuçlar, kullanılan tahmin modelinden bağımsız olarak, dengeli veri setinin (dengesiz/orijinal veri setine kıyasla) ikinci sınıftan önce üniversiteyi bırakma ihtimali olan öğrencileri belirlemek için daha iyi tahmin modelleri ürettiğini göstermiştir. Bu çalışmada kullanılan dört ayrı tahmin modeli arasında en iyi performansı %87,23'lük genel tahmin oranıyla destek vektör makineleri verirken, karar ağacı %87,16'lık genel tahmin oranıyla ikinci olmuş, onu sırasıyla %86,45 ve %86,12'lik genel tahmin oranlarıyla yapay sinir ağları ve lojistik regresyon izlemiştir. Modellerin duyarlılık analizi, eğitim ve finansal değişkenlerin bu olgunun en önemli belirleyicileri arasında olduğunu ortaya koymuştur.

"Bir teknoloji üniversitesinde okulu bırakma riski taşıyan öğrencileri belirlemek için nasıl bir tahmin modeli tasarlanabilir ve geliştirilebilir?" sorusundan yola çıkan Lottering vd. (2020)

araştırma projesi kapsamında EDM'yi kullanarak okulu bırakma tehlikesi olan öğrencileri tanımak için bir tahmin modeli oluşturmuştur. Bu çalışma, boyutsal azaltmanın yeterliliğini araştırmakta ve okulu bırakma tehlikesi olan öğrencilerin belirlenmesi için öğrenci bilgilerinde kapsanan önemli veri değişkenlerini bulmayı amaçlamaktadır. Bunu başarmak için, bir üniversitenin Bilgi İletişim ve Teknoloji (BİT) Fakültesi programlarından elde edilen öğrenci verilerine faktör analizi ve eğitimde veri madenciliği teknikleri uygulanmıştır. Doğruluk, duyarlılık, özgüllük ve F-ölçümü dahil olmak üzere EDM çıktıları, tahmin güçlerini ifade etmek için karşılaştırılmıştır. Özellik seçimi yoluyla ele alınan 19 nitelik arasında tahmin gücü olmayan faktörler (gürültülü veriler) elenmiş geriye kalan 13 nitelik için (konaklama, yaş, toplanan krediler, özürlülük, mali yardım, cinsiyet, ana dil, tamamlanan modüller, tekrarlanan modüller, ilk kez kullanılan modüller, sabır/devamlılık, önceki yıl faaliyetleri, sistemdeki aktif yıllar) farklı makine öğrenimi algoritmaları karşılaştırılmış ve en iyi performansı gösterenler seçilmiştir. Orijinal veri setinde en iyi performans gösteren Destek Vektör Makinesi sınıflandırıcısının okul terki tahminindeki sonuçları umut vericidir, %99.32 F-Ölçüm puanı elde etmiş ve veri kümesindeki gürültü kaldırıldığında bir miktar puan kaybederek %98.99'a düşmüştür. Bu gözlem, azaltılmış veri kümesine uygulandığında da mevcuttur. Küçültülmüş veri setinde Rastgele Orman sınıflandırıcısında da önemli bir iyileşme gözlemlenmiştir.

Guarín vd. (2015) yaptıkları araştırmada Kolombiya Ulusal Üniversitesi'nde akademik yıpranmayı (akademik statü kaybı) modellemek için eğitimde veri madenciliği yaklaşımının uygulanmasının sonuçlarını sunmaktadır. Bu makalede, sadece sosyo-ekonomik verileri değil, aynı zamanda akademik kayıtlar da kullanılarak belirli bir zamanda akademik performans kaybını tahmin etmek için iki veri madenciliği modeli (naive Bayes ve bir karar ağacı sınıflandırıcısı) önerilmektedir. Modeller, öğrencinin ilk dört dönem kaydındaki yıpranmayı tahmin etmeyi amaçlamakta ve kullanılan verilere ilişkin çeşitli senaryolar test edilmektedir. Araştırmada ele alınan öğrenci kabul/kayıt veri seti alanları kısaca aşağıdaki gibi üç kategoride gruplandırılabilir:

- İlk Akademik Bilgiler: lise türü (ör. devlet, özel), erişim türü (ör. normal, özel kabul programı), öğrencinin programı seçtiği seçenek (1'den, ilk seçenekten, 3'e kadar) ve varsa önceki program.
- Demografik ve Sosyo-Ekonomik Bilgiler: kabul yaşı, cinsiyet, gelinen şehir, sosyo-ekonomik sınıflandırma, etnik köken.

- Akademik Potansiyel: beş modülde (yani fen bilimleri, matematik, resim, dil ve sosyal çalışmalar) kabul testi puanları ve temel matematik ile okuryazarlık için sınıflandırma seviyeleri.

Çalışmanın ilk aşamasında modeller tüm veri setini kullanmaktadır; ancak, bu modeller 49 lisans programındaki risk altındaki öğrencileri tahmin etmek için kullanılmak istendiğinde: "sınıflandırıcı bir programa özgü olduğunda mı yoksa popülasyona genel olduğunda mı daha iyi performans gösterir?" sorusunu yanıtlamak için yeni bir deney geliştirilmiştir. Sonuç olarak tüm modellerin program türünden bağımsız olarak aynı değişkenleri kullandığı görülmüştür. Öte yandan Bayes sınıflandırıcısının performansı, ilk kayıt verilerine akademik veriler eklendiğinde artmıştır; ancak ikinci dönemde akademik veriler eklendikten sonra performans düşmüştür. Karar ağacı sonuçlarının bu konuda daha tutarlı ve yeni veriler üzerinde test edilirken daha güvenilir olduğu sonucuna varılmıştır. Deneysel sonuçlar, modellere akademik veriler eklendiğinde akademik statü kaybı tahmininin iyileştiğini göstermektedir; ancak daha fazla akademik veri eklemek sınıflandırıcıyı geliştirmeye devam etmemektedir.

Nouri vd. (2019) yürüttükleri araştırmada bir yandan lisans tez projelerinin neden tamamlanıp/tamamlanmadığını belirlemeyi ve açıklamayı, diğer yandan da makine öğrenmesi algoritmalarını kullanarak tez projelerinin tamamlanma durumunu tahmin etmeyi amaçlamaktadır. Çalışmanın örneklemi, 2010-2017 yılları arasında başlatılan lisans öğrencilerinin tez projelerinden (n=2436) oluşmaktadır. Veriler, tez projeleriyle ilgili verileri kaydetmek için kullanılan iki farklı veri sisteminden elde edilmiştir. Bu sistemlerden, hem öğrenciler hem de danışmanlarla ilgili değişkenleri içeren tez projesi verileri toplanmıştır. Tez projelerinin tamamlanmamasını ve tamamlanmasını etkileyen faktörleri belirlemek için geleneksel istatistiksel analizler (korelasyon testleri, t-testleri ve faktör analizi) yapılmış ve tamamlanma ve tamamlanmamayı öngören bir model oluşturmak için çeşitli makine öğrenimi algoritmaları uygulanmıştır. Sonuçlar, tamamlama ve tamamlamama konusunda öğrencilerin tez sürecindeki performansını üç genel faktör grubunun etkilediğini göstermiştir. Bunlar;

- Öğrencinin lisans programındaki önceki performansı
- Danışmanın tez projesi performansı ve deneyimi
- Süpervizörün araştırma çıktısı

Quadri ve Kalyankar (2010) tarafından yapılan bir araştırma öğrencilerin okulu bırakma özelliğini tahmin etmede kullanılan veri madenciliği çalışmalarını sunmaktadır. Çalışma için hibrit bir yöntem geliştirilmiştir; modelleme süreci karar ağacı teknikleri ve lojistik regresyonun

bir kombinasyonudur. İlk olarak, okul terklerini etkileyen faktörleri belirlemek için karar ağacı algoritması kullanılmaktadır. Faktörler belirlendikten sonra, okul terklerini ve her bir risk faktörünün etkisini ölçmek için lojistik regresyon tekniği kullanılmıştır. Öğrenci terklerinin modellenmesi: Modelleme süreci, öğrenci terkleri ile cinsiyet, devam, önceki dönem notu, ebeveyn eğitimi, ebeveyn geliri, burs, ilk çocuk ve öğrencinin çalışıp çalışmadığı gibi altta yatan risk faktörleri arasındaki ilişkiyi inceleyerek başlamıştır. Nihai analize: cinsiyet, katılım, ön sınıf puanı, ebeveyn eğitimi, ebeveyn geliri, burs, ilk çocuk durumu, çalışma durumu, okulu bırakma durumu değişkenleri dahil edilmiştir. Sonuç olarak okulu bırakma sıklığının, diğer tüm değişkenler arasında en önemli risk faktörü olan ebeveyn gelirine (bu çalışmada) göre değiştiği gösterilmiştir. Düşük gelir düzeyi, orta ve yüksek gelir düzeylerine kıyasla okulu bırakma özelliği üzerinde büyük etkiye sahiptir. Çocuğun ailenin ilk çocuğu olup olmaması da okulu bırakma özelliği üzerinde etkilidir. Önceki dönem notu 5'in üzerindeyse, öğrencinin ilk çocuk olma sorumluluğu (erkek için) nedeniyle okulu bırakma özelliğinin yüksek olduğu görülmektedir. Öte yandan kız öğrenciler de ailenin "ilk çocuğu" olarak evlenme sorunuyla karşı karşıyadır. Karar Ağacı analizine göre, cinsiyet, ebeveyn eğitimi, burs, yarı zamanlı iş gibi nitelikler ilgisiz faktörlerdir. Bu çalışma, en iyi tahmin ve analizi seçmek için karar ağacı tekniğini önermektedir.

Okul terkini öngörebilmek amacıyla öğrencilerin bir kurs içindeki aktivitelerinden elde edilen verilere göre erken uyarı sistemini araştıran başka bir çalışma da Márquez vd. (2016) tarafından Meksika'da 419 lise öğrencisi ile yürütülmüştür. Araştırmacılar geleneksel sınıflandırma yaklaşımlarının aksine –ki bunun için genelde kursların tamamlanması ve verinin böylece toplanması gerekir- bu çalışmada öğrencilerin kursları terk etmesini, en kısa sürede keşfetmek için anlaşılır bir metodoloji ve özel bir sınıflandırma algoritması önermektedir. Araştırmada geleneksel yaklaşım yerine pek çok adımda terk ve devam ile ilgili veriler alınmış; veri madenciliği algoritmaları ile en yüksek tahminin en kısa sürede yapılabildiği zaman araştırılmıştır. Çalışmada veri madenciliğinde geleneksel sınıflandırma algoritması yerine Yorumlanabilir Sınıflandırma Kuralı Madenciliği (Interpretable Classification Rule Mining-ICRM) algoritması ile birlikte okul terkinin en erken adımını belirlemek için her adımda IF-THEN kuralı ile türetilmiş özel bir algoritma kullanımı önerilmektedir. Sonuçta önerilen veri madenciliği algoritması ile kursun ilk 4-6 haftasında öğrencilerin okulu bırakacağını tahmin etmenin mümkün olduğu ve bunun bir erken uyarı sisteminde kullanılabileceği gösterilmektedir.

Ortiz-Lozano vd. (2023) Barselona Üniversitesi'nde İşletme derecesi alan 3583 birinci sınıf öğrencisinden elde edilen akademik verileri kullanarak yaptıkları araştırmada öğrenci katılımının ve okul terkinin potansiyel öngörücü göstergelerini belirlemeyi amaçlamaktadır. Çeşitli değişkenler okulu bırakma kararını etkilerken, hepsi aynı etkiye sahip değildir. Bu değişkenleri inceleyerek, okulu bırakma riski taşıyan öğrencileri tespit edebilecek ve destekleyici müdahalelerin uygulanmasını sağlayabilecek bir 'erken uyarı sistemi' kurulması hedeflenmektedir. Bu araştırmanın amacı iki yönlü belirlenmiştir: ilk olarak, okul terkinin belirlenmesinde sıklıkla kullanılan üç geniş değişken kategorisinden (kişisel, örgütsel ve hedef başarısı) ikincisinin, özellikle de ilk dönemdeki akademik sonuçların, okul terki için en büyük yordayıcı olduğu hipotezini doğrulamak; ikinci amaç olarak ise kolay erişilebilir değişkenler olan akademik performans değişkenlerine dayalı basit bir okul terki tahmin modeli geliştirmektir. Model performanslarını karşılaştırmak için üç farklı algoritma kullanılmıştır: Sinir Ağları, Rastgele Orman ve Lojistik Regresyon. Optimum kesme değeri ROC eğrisine dayalı olarak otomatik olarak belirlenmiştir. Araştırma sonucu, ikinci değişkenin, başarısız olunan ve ilk yarıyılda devam edilmeyen derslerin yüzdesinin, önemli bir tahmin gücü gösterdiğini ortaya koymaktadır. Bu durum, sonuçların sağlamlığını değerlendirmek için Madrid Complutense Üniversitesi'ndeki (İspanya) üç dereceli programların (Hukuk, Lisans ve Ekonomi) 10.784 öğrencisinden oluşan ek bir örneklem ile doğrulanmıştır. Dolayısıyla, sonuçlar tüm durumlarda oldukça tutarlıdır: hem basit bir model (logit) hem de daha sofistike olanlar (sinir ağları ve rastgele orman) yüksek doğruluk (doğru tahmin edilen değerler) ve hassasiyet (doğru tahmin edilen terkler) sergilemektedir. Test setinde sırasıyla %77 ve %69'luk ortalama değerler elde edilmiştir. Gerçekten de modeller, sadece bir dersi geçmenin önemli bir etkiye sahip olmadığını, sınavlara girme eyleminin daha önemli bir etkiye sahip olduğunu ortaya koymaktadır.

Theune vd. (2020) yürüttükleri araştırmada yükseköğretim kurumlarındaki okul terkini mümkün olan en kısa sürede tahmin etmeyi amaçlamaktadır. Bu çalışmada, koşullu çıkarım ağaçlarına dayalı rastgele orman metodu ile birinci sınıf öğrencilerinin öğrenci yaşamının ve eğitim kurslarının çok çeşitli yönlerini kapsayan geniş bir veri seti kullanmıştır. Çalışılan modelde ilk olarak, bir öğrencinin üniversiteye girmeden önceki başlangıç riski, yalnızca öğrenim öncesi belirleyicileri dâhil edilerek modellenmiş; bir sonraki adımda, öğrencinin eğitim-karar aşamasını kapsayan değişkenleri eklenmiştir. Üçüncü adımda, öğrencilerin eğitimlerinin en başındaki erken eğitim aşamasını temsil eden belirleyiciler dâhil edilmiştir. Araştırma sonucunda ön çalışma modellerine erken çalışma aşamasındaki değişkenler dâhil

edildiğinde tahmin performansının güçlü bir şekilde arttığı gözlenmekte ve bu da dikkate alınan tüm çalışma alanları için 0,83 ile 0,88 arasında bir AUC ile sonuçlanmaktadır. Ortaöğretim notunun yüksek önemi her üç modelde de açıktır, ancak erken eğitim aşaması belirleyicileri eklendiğinde önemli ölçüde azalmaktadır. Okul notunun yanı sıra, öğrenci memnuniyeti ve öznel akademik benlik kavramı ve öz değerlendirme ile ilişkili belirleyiciler de okulu bırakma sürecinde önemli bir rol oynamaktadır. Bazı öğrenciler için başlangıç riski yüksek olsa da, yükseköğretimden başarılı bir şekilde mezun olma olasılıklarını artırmanın birçok yolu olduğu sonucuna varılmıştır.

Eğitsel veri madenciliği yoluyla yükseköğretim öğrencilerinin okulu terkinin tahmini için Hegde ve Prageeth (2018) tarafından yapılan bir araştırma Naive-Bayes sınıflandırma algoritmasını kullanarak tahmin için öznitelik seçimi ve öznitelik çıkarımı ile ilgili çeşitli boyut indirgeme tekniklerini içeren bir metodoloji (%72 doğruluk oranı ile sınıflandırmanın başarılı olduğu gösterilmiştir) önermektedir. Çalışmada Naive Bayes sınıflandırma algoritması kullanılmıştır. Pek çok hayal ve beklenti ile okula başlayan öğrencilerin, beklentileri karşılanmadığında veya demografik özellikler gibi faktörlerin de etkisiyle kayıt oldukları programlardan ayrıldıkları görülmektedir. Bunun eğitim kurumları için büyük bir tehdit olduğunu belirten araştırmacılar; demografik faktörler, psikolojik faktörler, sağlık sorunları, akademik performans, öğretmen görüşü, öğrenci davranışı gibi faktörleri tahmin değişkenleri olarak belirlemiştir.

Segura vd. (2022) yayımladığı makalede, Avrupa'nın en büyük üçüncü yüz yüze üniversitesinde ilk yılın sonunda okulu bırakan öğrencilerin verilerini kullanarak, adayların eğitimin başında ya da ilk dönemin sonunda okul bırakma olasılığını tahmin edilmeye çalışmıştır. Okul terkindeki en yüksek oranların ilk yıl veya dönem sonunda olduğundan yola çıkan araştırma, verileri, 2017-18 akademik yılında üniversitede beş farklı alanda (Sosyal Bilimler ve Hukuk, Fen Bilimleri, Sağlık Bilimleri, Mühendislik ve Sanat ve Beşeri Bilimler) eğitim veren 10 bölümün kayıtlı birinci sınıf öğrencilerinden elde etmektedir. Öğrenciler üç kategoride gruplandırılabilecek bir dizi değişkene göre karakterize edilmiştir: sosyo-ekonomik değişkenler, kayıt değişkenleri ve üniversitenin ilk dönemindeki akademik performans değişkenleri. İlk kategori, diğerlerinin yanı sıra cinsiyet, yaş, ebeveynlerin veya velilerin eğitim düzeyi, ikamet edilen belediye ve milliyet gibi değişkenleri içermektedir. İkinci grupta okul türü (devlet veya özel), üniversite giriş notu, öğrencinin tercih ettiği derece sayısı ve giriş çalışması (okul, mesleki eğitim ve giriş uzmanlığı) gibi değişkenler bulunmaktadır. Son olarak, üniversitenin ilk dönemine ait değişkenler arasında kayıt ücretinin miktarı, Avrupa Kredi

Transfer Sistemi sayısı (kayıt olunan ve geçilen), ilk dönem için ortalama not ve öğrencinin burslu olup olmadığı yer almaktadır. Bu çalışmada, ilk aşamada Özellik Seçimi uygulanmış ve ardından beş öngörücü makine öğrenme yöntemi uygulanmıştır: YSA, DVM, KNN, Karar Ağacı ve Lojistik Regresyon. Sonuçlar, okulu bırakma tespitinin sadece kayıt değişkenleriyle çalışmadığını, ancak ilk dönem sonuçlarından sonra iyileştiğini göstermektedir. Akademik performans ise tüm modellerde ilgili bir değişkendir, ancak öğrencinin sonunda çalışabildiği ders üzerindeki tercih düzeyi gibi başka değişkenlerde de vardır. Ayrıca çalışma sonucunda kullanılan makine öğrenimi yöntemlerinin iyi sonuçlar verdiği görülmüştür. Ancak kullanılan tekniklerin başarısı program alanlarına göre değişmektedir.

Christie vd. (2019) tarafından Amerika Birleşik Devletleri'nde zamanında mezun olma oranlarının çok düşük olduğu okullarda makine öğrenmesi ile yürütülen bir çalışmada risk altında olan öğrencilerin eğitim kayıtları ve kayıt sonuçları arasındaki istatistiksel ilişkiler modellenmiş ve öğrenciler için oldukça öngörülü risk puanları belirlenmiştir. Çalışma, önceden uyarı sinyalleri ile kurumlara ve danışmanlara zaman ve pratiklik kazandıracak duyarlı, öğrenci bilgi sistemine entegre edilmiş ve günlük olarak güncellenen bir sistem önermektedir.

Delen (2011) araştırmasında, 8 yıllık kurumsal verilerin yanı sıra veri madenciliği tekniği kullanarak, birinci sınıf öğrenci kaybını tahmin etmek için analitik modeller geliştirmiştir. Bu çalışmada üç popüler sınıflandırma tekniği kullanılmıştır: yapay sinir ağları, karar ağaçları ve lojistik regresyon. Modellerin değişken önem analizi, eğitim ve finansal değişkenlerin bu çalışmada kullanılan yordayıcılar arasında en önemlileri olduğunu ortaya koymuştur. Araştırma sonucu, uygun değişkenlerle yeterli veri sağlandığında, veri madenciliği yöntemlerinin birinci sınıf öğrenci kaybını yaklaşık %80 doğrulukla tahmin edebildiğini göstermektedir. Bu çalışmada karşılaştırılan üç tahmin yöntemi arasında yapay sinir ağları en iyi performansı göstermiş ve okuldan ayrılma örnekleminde %81 genel tahmin doğruluğu elde etmiştir. Bunu karar ağaçları ve lojistik regresyon izlemiştir. Kullanılabilirlik açısından bakıldığında, yapay sinir ağlarının daha iyi tahmin sonuçlarına sahip olmasına karşın, sinir ağlarına kıyasla daha şeffaf bir model yapısı sergiledikleri için karar ağaçlarının kullanılmasının tercih edilmesi önerilmektedir.

Washington üniversitesinde örgün programlar için yükseköğretimde okul terkinin tahminine yönelik yapılan başka bir araştırmada (Aulck vd., 2017) 69.116 öğrenciye ait demografik özellikler (ırk, cinsiyet, doğum tarihi, ikamet bilgisi ve Hispanik kimlik durumu), üniversite giriş sınav puanı ve transkript kayıtları tahmin verisi olarak kullanılmıştır. Düzenli lojistik

regresyon, k-en yakın komşular ve rastgele orman modelllerinden elde edilen sonuçların 0,66 ile 0,73 arasında doğruluk oranına sahip olduğu; matematik, İngilizce, kimya ve psikoloji derslerindeki transkript notları ile kayıt yılı ve doğum yılı gibi değişkenlerin en güçlü belirleyiciler arasında yer aldığı gösterilmiştir. Modeller ROC eğrileri ile karşılaştırılmış lojistik regresyon modelinin en güçlü tahmini sağladığı bulunmuştur.

Serra vd. (2018) denetimli sınıflandırma algoritmalarını kullanarak yükseköğrenimde okul terkini tahmin etmek için yaptıkları araştırmayı 2013-16 eğitim dönemlerinde Bari Aldo Moro Üniversitesi'nden elde ettikleri verilerle yürütmüştür. Veri analizi, üniversite bilgi sistemi içinde her öğrenci için mevcut olan bilgi setine dayanmaktadır. Okul terki tahminleri, denetimli sınıflandırma algoritmaları olarak bilinen uygun makine öğrenimi teknikleri (DVM, lojistik regresyon, Gaussian naive bayes) aracılığıyla gerçekleştirilmiştir. Tahmin değişkenleri olarak ulusal kimlik, doğum bölgesi, kayıtlı bulunan ders türü, ikamet bölgesi, cinsiyet, yaş ile mezuniyet/program çıkış yılı kullanılmıştır. Sonuç olarak lojistik regresyonun terk eden öğrencilerin tahmininde başarılı olmadığı (%31 hassasiyet), Gaussian NB algoritmasının %76 doğruluk ve %69 hassasiyet ile sınıflandırmada daha iyi sonuç verdiği ancak en iyi sınıflandırma performansının destek vektör makinaları (%80 doğruluk ve %97 duyarlılık) ile elde edildiği bulunmuştur.

Gil vd. (2020) okul terkinin altında yatan faktörleri belirlemeye yönelik bir lisede yürüttükleri araştırmada, hazır veri kümelerinden makine öğrenimi algoritmaları ile simülasyonlar çalıştırılmış; iki sınıflandırma algoritması C4.5 ve Naive Bayes, yaş, cinsiyet, medeni durum, yazılı performans ve üç aylık sınav notu, devam yüzdesi, ev mesafesi, veli geliri gibi öğrencinin akademik ve demografik ayrıntılarını içeren bir veri kümesi üzerinde test edilmiştir. En yüksek doğruluk performansının C4.5 algoritması tarafından algılandığı gösterilmiştir.

Dekker vd. (2009) tarafından yürütülen ve okuldan ayrılan öğrencilerin öngörülmesi için yapılan çalışma; elektrik mühendisliği öğrencilerinin öğrenimlerinin ilk döneminden sonra veya öğretim programına girmeden önce eğitsel veri madenciliği ile okulu bırakma durumları ve başarılarına etki eden faktörleri belirlemeyi amaçlamıştır. Deneysel çalışma, Weka aracılığı ile iki karar ağacı algoritması CART ve C4.5, Bayes sınıflandırıcısı (BayesNet), lojistik model, kural tabanlı öğrenici (JRip) ve rastgele orman ile yapılmıştır. Eindhoven Teknoloji Üniversitesi Elektrik Mühendisliği bölümünden alınan 648 öğrenci verisi ile yürütülen çalışmaya göre en basit şekilde karar ağaçları %75 ile %80 arasında doğrulukla tahmin yapabilmektedir. Başarının en güçlü belirleyicisi, genel olarak belirleyici ders olarak

görülmeyen Lineer Cebir dersidir. Diğer güçlü belirleyiciler ise Calculus, Networks ve Fen derslerinin ortalama notlarıdır. En ilgili bilgiler üniversitenin kendisinde toplanmaktadır: üniversite öncesi verilerin özeti kullanılması daha yararlı olabilir. Derinlemesine modeller değerlendirildiğinde araştırma sonucunda verilerin niteliği hususunda üç önemli iyileştirme öne çıkmaktadır; ilk olarak değişen ders organizasyonuna göre oluşan değişimlere bir çözüm bulunmalıdır. Bir dersteki öğrenci performansına ilişkin mevcut bilgilerin veri setindeki tüm öğrenciler için kullanılabilecek şekilde bir araya getirilmesi, şu anda oldukça yaygın olan yanlış sınıflandırma türlerini önleyebilir. İkinci bir iyileştirme, genel olarak notların kodlanması daha standart olmalıdır: bilinmeyen veya mevcut olmayan tüm bilgilerin sıfırla eşleştirilmesinin etkili olmadığı görülmüştür. Öte yandan yine ilişkili olarak kayıp değerlerle ilgilenen daha gelişmiş bir çözüm de bu açıdan düşünülebilir. Yapılan bu araştırma sonucunda yanlış sınıflandırmalar için maliyete duyarlı kapsamlı bir analiz sunulmuş ve okul terki riski bulunan elektrik mühendisliği öğrencileri hakkında herhangi bir ek veri toplamadan tahmin geliştirmenin diğer yolları gösterilmiştir.

Del Bonifro vd. (2020) öğrencilerin okul terkini kursa başlamadan önce ya da ilk yıl içinde tahmin etmek; dolayısı ile terkin sosyal ve ekonomik maliyetlerini hafifletmeye yardımcı olmak amacıyla yaptıkları araştırmada makine öğrenimi tekniklerinden yararlanmıştır. Birinci sınıf öğrencilerini hedefleyen çalışmada tahmin modellerini eğitmek için temel veriler; kişisel bilgiler ve lise akademik kayıtları (örneğin cinsiyet, yaş, lise eğitimi, final notu) ve ilk yılın ilk aylarında alınan kredi sayısı vb. değişkenlerin yer aldığı yaklaşık 15.000 öğrenciye ait veri seti kullanılmıştır. Okul terki olgusunun makul ölçüde doğru tahmin edilmesini sağlayan uygun aracı sağlamak için belirli bir dizi iyi bilinen ve benzer sorunları çözmek için literatürde de yaygın olarak kullanılan modellerden Doğrusal Diskriminant Analizi, Destek Vektör Makinesi ve Rastgele Orman gibi sınıflandırma algoritmaları dikkate alınmıştır. Modeller eğitilirken ilk aşamada mümkün olduğunca erken hareket etme için kurumun ilk kayıt zamanında elde ettiği bilgilere dayanan algoritmalar çalışılmış; sonrasında öğrencilerin belirli bir süre sonra elde ettiği ders kredilerini de dikkate alacak bir dizi test gerçekleştirilmiştir. Eklenen kredi faktörü, modelin nihai performansının büyük ölçüde iyileştirilmesine yardımcı olmaktadır. Temel özellik seti dikkate alındığında, Doğrusal Diskriminant Analizinin DVM sonuçları daha büyük bir varyansla en yüksek performansı elde etmektedir. Her halükârda, elde edilen sonuçlar, ilk aşamada herhangi bir pedagojik veya didaktik değeri olmayan verilerden yola çıkılarak geliştirilen modelin okul terki sorununu azaltma girişiminde pratik olarak yardımcı olabileceğini göstermektedir.

Pradeep vd. (2015) yaptıkları çalışmada eğitsel veri madenciliği kullanarak öğrencilerin akademik performanslarını etkileyen ve okulu bırakmalarının öngörülmesine katkıda bulunan faktörleri analiz etmektedir. Veri seti, Hindistan'ın Kerala eyaletindeki tanınmış bir okulda 2011 ve 2013 yılları arasında kayıtlı öğrencilerin 57 özelliğini içeren 670 öğrenci kaydından oluşmaktadır. Verilere tümevarım kuralları ve karar ağacı gibi çeşitli sınıflandırma teknikleri uygulanmıştır. Özellikle ortaöğretimde okul başarısızlığını tahmin etmeyi amaçlayan çalışmada yüksek boyutlu ve oldukça dengesiz verilerde farklı yaklaşımların kullanılması önerilmektedir. Bu yaklaşımların her birinin sonuçları, yüksek doğruluğa ulaşanı seçmek için karşılaştırılmıştır. Bu çalışmada öğrencilerin sosyal, ekonomik ve kültürel özelliklerinden ziyade akademik notları kullanılmıştır. Araştırma sonucunda sınıflandırma modellerinde sıklıkla ortaya çıkan puan değerleri fizik, İngilizce ve matematik derslerinde "Zayıf", "Çok Zayıf" ve "Sunulmadı" şeklindedir. Yaşın 15'ten fazla olması, çalışma motivasyonunun düşük olması ve öğrencilerin matematiği zor bir ders olarak görmesi, öğrenci başarısızlığına katkıda bulunan faktörlerden bazılarıdır. Nihai sonuçta kural tanımlayarak (tümevarım) ve karar ağacı gibi sınıflandırma algoritmaları ile bir öğrencinin başarısızlığını veya okul terkini başarıyla tahmin edilebileceği gösterilmiştir.

Agrusti vd. (2019) üniversite terkini tahmin için yaptıkları derleme çalışmasında veri madenciliği için ağırlıklı olarak altı sınıflandırma tekniğinin kullanıldığını belirlemişlerdir. Bunlar: karar ağacı, k-en yakın komşu, destek vektör makineleri, Bayes sınıflandırması, sinir ağları ve lojistik regresyona dayalı sınıflandırmalardır. Buna göre 1999 ile 2019 yılları arasında yapılan çalışmaların yaklaşık %67'sinde karar ağacı sınıflandırıcıları; yaklaşık %49'unda Bayes Sınıflandırması, yaklaşık %40'ında sinir ağları ve yaklaşık %34'ünde lojistik regresyona dayalı sınıflandırma kullanılmıştır. Destek vektör makineleri %23, k-en yakın komşu sınıflandırması %12 ve diğer çeşitli sınıflandırma teknikleri %15 oranlarında kullanılmaktadır. Scopus ve Web of Science kataloglarında, üniversite terkini tahmin etmek için hangi veri madenciliği tekniklerinin kullanıldığına odaklanan çalışma, 241 araştırma evreninden ilgili 73 çalışma tespit etmiş; bunların 6'sında veri madenciliği sınıflandırma tekniği, 53'ünde veri madenciliği algoritması ve 14'ünde de veri madenciliği aracı kullanıldığını göstermiştir.

Zhang vd. (2010) öğrenciyi elde tutma ve performanslarını izlemek, öğrencilerin akademik davranışlarını analiz etmek ve etkili müdahale stratejileri için temel sağlamak amacıyla veri madenciliği ve doğal dil işleme teknolojilerini kullanan bir araştırma yürütmüştür. Bu makalede, veri madenciliğinin risk altındaki öğrencileri tespit etmeye, kurs veya modül uygunluğunu değerlendirmeye ve öğrenci kalıcılığını artırmak için müdahaleleri uyarlamaya

nasıl yardımcı olabileceği tartışılmıştır. Araştırma Tinto'nun modelinin akademik perspektifini, performans yönünü kapsayan entegre bir veri madenciliği sistemi oluşturmayı amaçlamaktadır. Veri kaynakları Thames Valley Üniversitesine ait öğrenci kayıtlarını, öğrenci sonuçlarını, ders/modül verilerini, öğrenme becerilerini ve öğrenci etkinliklerini vb. kapsamaktadır. Proje sonucunda öğrenci, modül ve kurs bilgilerini izlemek için bir web sitesinde müdahale bölümü geliştirilmiştir. Bir öğrenci giriş yaptığında, tüm modül sonuçlarının detay bilgilerini, ilerleme grafiğini, risk altında olup olmadığını ve nedenini kontrol edebilmekte; benzer şekilde eğitmenler ve yöneticiler de hangi öğrencilerin risk altında olduğunu ve nedenini bularak öğrenciyle iletişime geçmek isterlerse sistem içinde bir e-postayı kullanabilmektedir. Araştırma sonucu risk altındaki öğrencilerin erken tespitinin yanı sıra yoğun ve sürekli müdahalenin sürdürülmesinin öğrenci kalıcılığını artırmanın anahtarı olduğunu göstermektedir. Ayrıca, üniversitelerin programlarını ve derslerini öğrencilerin hem kişisel hem de akademik olarak en yüksek başarı olasılığına sahip olacak şekilde nasıl hazırlayabileceklerini de açıklamaktadır.

Ribeiro ve Canedo'nun (2020) yayımladıkları araştırmanın amacı, Brasilia Üniversitesi'ni terk eden öğrencileri etkileyen faktörleri belirlemek ve makine öğrenimi ile hangi öğrencilerin lisans derslerini bırakacağını tahmin etmek için bir model sağlamaktır. Bu sayede, okulu bırakma oranını azaltmak için önlemler alınabileceğini ifade eden araştırmacılar öğrenci tutma politikasını iyileştirmek için bir plan hazırlamasına ve en yüksek okul terki oranlarına sahip derslere genel bir bakış sunarak öğrencilerin derslerinin tamamlanması ve performans iyileştirebilmek için nasıl motive edebileceği konusunda tartışmaya yol açmaktadır. Araştırmada geliştirilen model, öğrencilerin lisans derslerini tamamlamadan okulu bırakma olasılığını %54 olarak öngörmüştür. Oluşturulan modele göre, ders kredilerinin miktarı, bir lisans derecesini tamamlamak için gereken minimum ve maksimum süre, öğrencinin Brasilia Üniversitesi'ne giriş formu, ilk ders konularında alınan negatif notların toplamı gibi faktörler, öğrencinin bir lisans derecesini tamamlamadan ayrılmasına neden olan bazı ilgili faktörlerdir. Ayrıca, sosyal faktörlerin de lisans terkine katkıda bulunduğu gösterilmiştir. Öte yandan çalışmanın sonucu, en fazla krediye sahip derslerin (iş yükü), tamamlanması daha uzun süren derslerin (5-6 yıllık dersler) ve öğrencinin daha düşük akademik performansının (zayıf notlar) öğrenci terk oranını etkilediğini ortaya koymuştur. Öğrencinin akademik performansı ve lisans dersinin zorluk derecesi gibi faktörlerin hâlâ okulu bırakmanın ana faktörleri olduğu sonucuna varılmıştır. Tasnim vd. (2019) tarafından eğitsel veri madenciliği yöntemleriyle okuldan ayrılan öğrencileri belirlemeye yönelik yapılan başka bir çalışmada da, mevcut yaklaşımlardan daha

iyi performans gösteren okul terk öğrencilerini belirlemek için eşik tabanlı bir yaklaşım önerilmiştir.

Altıntaş vd. (2021) akademik riskli öğrencilerin kestirimi için yaptıkları çalışmada, yükseköğretim kurumlarının lisans programlarında okuyan öğrencilerin yılsonundaki akademik risklerini kestirebilecek bir Karar Destek Sistemi modeli önerilmiştir. Çalışma Ankara Üniversitesi'nde yürütülmüş; okuyan öğrencilerin akademik başarılarında etkili olan faktörler belirlenmiştir. Örnek çalışma, öğrenci başarısı üzerinde etkili olduğu düşünülen, öğrenci özellikleri ile öğretim süreçlerine ait değişkenler temel alınarak planlanmıştır. Model tahmininde, klasik istatistiksel çözümlemeler ve makine öğrenmesi yöntemlerinden LR, DVM ve YSA'dan yararlanılmıştır. Model performansları ise çeşitli hata kriterleri kullanılarak değerlendirilmiş; sınıflandırma hatası en düşük olan model akademik riskli öğrencilerin kestiriminde kullanılmıştır. Bu model lisans programlarında öğrenim gören öğrencilerin performanslarını kestirirken, akademik riskli öğrencileri sınıflayan modelleri içeren bir Karar Destek Sistemini de içermektedir.

2.3.3. Açık ve Uzaktan Öğrenmede (Yükseköğrenimde) Okul Terki ve Devamlılık Kavramına Odaklı Araştırmalar

Lee ve Choi (2011) 1999-2009 yılları arasındaki çevrimiçi kursların bırakılması hakkında yayımlanan araştırmaları kapsamlı olarak incelemiş; incelenen bu çalışmaların çoğunun çevrimiçi kurslardan ayrılma tanımı üzerinde genel bir anlaşma sağlamadığını ve bunun da ayrılma faktörlerini karşılaştırmayı oldukça zorlaştırdığını belirtmiştir. Araştırmalarda "Öğrenci faktörleri"nin alanyazında en çok atıf alan faktör olduğu (ele alınan tüm makalelerin %55'i), bunu "Çevresel faktörler" ve "Kurs-program faktörleri"nin (sırasıyla %25 ve %20) izlediği sonucuna varmışlardır. Benzer şekilde Grau-Valldosera ve Minguillón (2014) ile Grau-Valldosera (2019) Katolanya Açık Üniversitesi örneğini ele alarak okul terkinde kapsamlı çalışmalar yapmıştır. Lee ve Choi'nin çalışmasını güncellediğini belirten araştırmacılar okul terkinde etkili olan temalar için %63 öğrenci faktörlerine, %45 kurum veya kurs-program faktörlerine ve %3'ü dış faktörlere referans verildiği sonucuna varmaktadır. Grau-Valldosera (2019) "Çevrimiçi uzaktan eğitimde devamlılık niyeti ve etkili yeniden kayıt modelleri için bir terk tanımı" adlı doktora çalışmasında da öğrenci devamlılığını ve daha spesifik olarak, yükseköğretim programlarına ikinci yarıyılda ara veren öğrencilerin yeniden kayıtlarını ölçmek

için e-öğrenme bağlamında okul terkini tanımlamaya odaklanmaktadır. Geliştirilen yeniden kayıt modeli için veri madenciliği tekniklerine dayalı analizlerden yararlanmıştır.

Xavier ve Meneses (2020) de literatür taraması yaptıkları araştırmalarında çevrimiçi yükseköğretimde okul terki tanımlarına odaklanmıştır. 2014-2018 yılları arasında yapılan kapsam taramasında 138 makale ve teze ulaşılmıştır. Bulgular, standart tanımlardan yoksun, karmaşık ancak dağınık bir alanın varlığını ortaya koymaktadır. Reed vd. (2013) da literatür derlemesi yaptıkları çalışmalarında dijital çağda okul terki ve yıpranma ("okulu bırakma" veya "geri çekilme") üzerinde etkili olan nitelikleri ve temel kaygıları araştırmışlar ve sonuç olarak bütüncül bir şekilde okul terkini ele almanın araştırmacıları yanılttığını; yıpranmanın her yönünü açıklamanın yönetilemez olduğunu iddia etmektedir.

Appavoo, Gungea ve Sohoraye (2023), gelişmekte olan bir üniversite olan Mauritius Açık Üniversitesi'nde okulu bırakmanın ana nedenlerini ve okulu bırakma oranını düşük tutmak için alınan önlemleri anlatmış; okulu bırakmaya neden olan içsel ve dışsal faktörleri araştırmayı amaçlamıştır. Sonuç olarak işitsel ve kariyerle ilgili konuların, Mauritius Açık Üniversitesi'nde eğitimden ayrılmaya yol açan başlıca nedenlerden olduğu; alt nedenler arasında ise programların yanlış seçilmesi, yetersiz eğitim desteği ve işveren desteğinin olmaması görülmektedir.

Budiman (2018), Endonezya'daki bir açık üniversite tarafından sunulan uzaktan dil öğrenme programını bırakma nedenlerini inceleyen bir çalışma sunmaktadır. Açık uçlu soru anketleri ve görüşmelerden elde edilen bulgular, öğrencileri okulu bırakmaya iten başlıca nedenlerin İngilizce temel becerilerinin eksikliği, karşılanmayan beklentiler, izolasyon hissi ve iş, aile ve eğitim sorumluluklarını dengeleyememe olduğunu göstermiştir. Çalışma, öğrencilerin uzaktan dil öğrenme programını bırakmalarıyla ilgili faktörleri tanımlamak için teorik bir çerçeve sunmaktadır.

Simpson (2013) akademik çalışmasında, uzaktan eğitimde öğrencilerin düşük mezuniyet oranlarının (ki örgüne göre oran dörtte bir düzeyinde) sebeplerini araştırmış; uzaktan eğitim kurumlarında böyle bir açığın meydana gelmesinin nedenlerini öğrenciler, kurumlar ve bir bütün olarak toplum üzerindeki etkileri perspektifinden analizler ile sunmuştur. Çalışma sonucunda, öğretme ile öğrenmeyi birbirine karıştırmanın kategorik hata olduğunu; kurumların, çevrimiçi olarak öğretim materyallerinin sağlanmasına çok fazla odaklandığını buna karşın öğrencileri öğrenmeye motive etmeye yönelik çok az şey yapıldığını öne sürmektedir. Okulu bırakmanın üstesinden gelmek için proaktif temasın gerekliliğine ve bu teması motive edici

hale getirmenin önemine dair birikmiş kanıtlar olduğunu; son olarak da daha spekülatif olarak, kaynaklar veya organizasyondan ziyade, uzaktan eğitimde öğrenci bağlılığının artırmanın önündeki ana engelin kurumsal yaklaşım olduğunu belirtmektedir. Bu çalışmasından daha önce yayımladığı kitabında Simpson (2003) açık, uzaktan ve çevrimiçi öğrenmede öğrencilerin okul terki ve okulda tutulmasıyla ilgili konuları kapsamlı bir analizler her yönüyle ele almıştır.

Türkiye'de ise Yılmaz (2020) açık ve uzaktan eğitim öğrencilerinin öğrenimi bırakma ve öğrenime devam nedenlerini incelediği doktora tezini nitel desende planlamıştır. Araştırma sonuçlarına göre, öğrencilerin öğrenimi bırakma-devam kararları, içsel nedenler (akademik entegrasyon, sosyal entegrasyon, kaynaklar vb.), dışsal nedenler (maddi durum, aile hayatı, sosyal hayat vb.), öğrenci özellikleri (kişilik yapısı, hedefe bağlılık vb.) ve öğrenci becerileri (öz düzenleme, metabilişsel düşünme vb.) olmak üzere dört ayrı bileşenden etkilenmektedir. Ayrıca bu bileşenler altında, öğrencileri devam etme/bırakma kararına götüren 41 etken belirlenmiştir. Bunun yanında, tezin sonucunda öğrenimi bırakma ve öğrenime devam etme nedenleri birleştirilerek daha önce geliştirilmiş olan bir model için güncelleme önerisi geliştirilmiştir. Benzer şekilde Aguş (2019) da tez çalışmasında açık ve uzaktan eğitimde öğrencilerin okulu bırakma nedenlerini yine Anadolu Üniversitesi bağlamında incelemiştir. Araştırmada okulu bırakma nedenlerinin ve bu nedenlerin demografik gruplar arasındaki farklılıklarının ortaya çıkarılması amaçlanmıştır. Toplam 302 lisans ve ön lisans öğrencisi ile yürütülen araştırmada veriler çevrimiçi anket aracılığı ile toplanmıştır. Araştırma sonucuna göre zaman kısıtlamaları ve motivasyon eksikliğinin okulu bırakmada en önemli faktörlerden olduğu görülmüştür. Öğrenim ücreti, okuldan beklenenlerin bulunamaması, ders çalışma alışkanlığına sahip olunmaması ve akademik zorluk okulu bırakmada orta derecede etkilidir. Tek başına ve iletişim teknolojileri aracılığı ile öğrenme gibi uzaktan eğitime has faktörlerin okul terki için önemi düşük bulunurken, en az öneme sahip faktörler; teknoloji yeterliği ve erişimi, iş ve okul değişikliği, aile ve sağlık sorunları olmuştur. Ayrıca ortaya çıkan bu nedenlerin önem derecelerinin, okula başlama amacı, cinsiyet, yaş, eğitim düzeyi, medeni durum, program türü, çalışma durumu ve ikamet yeri faktörlerine göre de farklılaştığı gösterilmiştir. Okur vd. (2019) da açık ve uzaktan öğrenmede öğrenimi bırakma sebeplerini inceledikleri derleme türündeki araştırmada öğrenimi bırakma nedenlerini kuramlar çerçevesinde ele almıştır. Buna göre okul bırakma veya kayıt yaptırmamanın başlıca nedenleri şu şekilde sıralanmıştır: öğrenim ücretinin yüksek olduğu düşüncesi, dönem uzatma endişesi ve buna bağlı başarısız olma fikri, dersler için gerekli olan öğrenme materyallerinin eksikliği, sınav şartları, akademik veya mesleki kariyer beklentisinin olmaması ve bu sebeple karşılaştıkları

olumsuz olaylar karşısında daha kolay vazgeçmesi ve temel bilgi teknoloji seviyelerinin yetersiz olmasıdır. Aydın vd. (2019) da 2012-2015 yılları arasında Anadolu Üniversitesi Açıköğretim Sisteminde öğrenenlerin okulu bırakma nedenlerini keşfedici sıralı karma yöntemi ile araştırmıştır. Araştırma sonucunda öğrenenlerin okulu bırakma nedenlerinden bazıları: sınav sonuçlarının tatmin edici olmaması, açıköğretim sistemi hakkında bilgi eksikliği, zaman yönetimi sorunları, kayıtlı olunan programın uygun olmaması, günlük yaşam rutinlerinin çalışmalara müdahalesi ve bazı kişisel sorunlar olarak bulunmuştur. Elibol ve Bozkurt (2023) ise veri madenciliği ve analitik yaklaşımlar kullanarak uzaktan eğitim literatüründe öğrenci terkine ilişkin araştırma eğilimlerini değerlendirmişlerdir. Metin madenciliği ve sosyal ağ analizi uygulanarak toplam 164 yayın incelenmiştir. Çalışmanın bulgularına dayanarak, bu makale, uzaktan eğitimde "okul terki" teriminin kesin bir tanımının yapılmasını, öğrenci terkini tahmin etmek için algoritmik yaklaşımların kullanımına yönelik etik ilkeler ve çerçeveler geliştirilmesini de içeren gelecekteki araştırmalar için çeşitli öneriler sunmaktadır. Bunlarında yanında Talan (2020), uzaktan eğitim öğrencilerinin akademik başarı durumlarını çeşitli değişkenler açısından incelediği nicel desendeki çalışmasında; uzaktan eğitim öğrencilerinin akademik başarı durumlarının yaş, cinsiyet ve bilgisayar kullanım süresi değişkenlerine göre istatistiksel olarak farklılaşmadığı sonucuna ulaşmıştır.

Munro (1987), uzaktan eğitimde okul terki söyleminin doğasını incelemek için sosyoloji, antropoloji, siyaset bilimi veya ekonomi gibi alanlardan teorinin anlamlı olabileceği (psikoloji veya müfredat tasarımından ziyade) daha geniş bir sosyal eğitim bağlamı kullanarak, sorunu yeniden kavramsallaştırmanın ve kayıt memurunun bakış açısından ziyade öğrencilerin deneyimleri hakkında konuşacak veriler toplamanın yollarını aramıştır. Mevcut literatürü gözden geçirerek, olumsuz sonuçlarının yanı sıra, okulu bırakmanın mobil veya bağımsız bir öğrenci için sağlıklı bir seçim olabileceğine işaret eden çalışma, kendi döneminde cevaplanmamış bir dizi soruyu gündeme getirmektedir.

Eğitim araştırmacıları yükseköğretimde okul terkini pek çok açıdan ele almışlardır ve kuşkusuz sorun yalnızca açık ve uzaktan öğrenmenin sorunu değildir. Klasik/örgün yükseköğretim programlarında da öncelikle okul terkinin tanımı, kuramsal yapısı ile ilgili model çalışmaları (Hagedorn, 2005; Thalhammer, Hoffmann, Von Hippel vd., 2022; Thomas & Unger, 2015; Tinto, 1975, 1982; Vossensteyn vd., 2015), terk olgusunda etkili olan faktörleri belirlemeye yönelik araştırmalar (Araque, Roldán & Salguero, 2009; Bülbül, 2012; Evans & Tragant, 2020; Gülşen, 2017; Günther & Koeszegi, 2012; Heublein, 2014; Perchinunno, Bilancia & Vitale, 2019) mevcuttur. Ayrıca alanda yapılan çalışmaların gözden geçirilmesi, literatür taraması vb.

ile sistematik analizler yapan araştırmalar da (Ballı & Kartal, 2020; Behr, Giese, Teguim Kamdjou vd., 2020; Guzmán, Barragán & Cala Vitery, 2021; Pinheiro, Ribeiro & Fernandes, 2023; Tinto & John Cullen, 1973; Véliz Palomino & Ortega, 2023) yükseköğretimde okul terki sorununa bütüncül bir bakış sağlayarak önemli katkılar sunmuşlardır.

2.3.4. Eğitimde Veri Madenciliği Çalışmaları; Öğrenci Başarı ve Performans Sınıflandırmaları

Veri madenciliği ile eğitimde artan çalışmalar özellikle ikili sınıflandırma alanında yoğunlaşmaktadır. Bu alanda ise özellikle öğrenci başarısının tahmini, ders başarısın tahmini ve genel olarak öğrenci performansının tahmini çalışmaları yapılmıştır. Literatürün bu bölümünde veri madenciliği ile öğrenci performansının ve başarısının tahmini ile ilgili çalışmalara yer verilmiştir.

Kovacic (2010) makine öğrenimi tekniklerini kullanarak öğrenci başarısının erken tahminini araştırmış; etkili tahmin için eğitim, iş, cinsiyet, statü, engellilik vb. sosyo-demografik özellikleri ve kurs programı, kurs bloğu vb. kurs özelliklerini analiz etmiştir. Veri kümesini içeren bu özellikler Yeni Zelanda Açık Üniversitesi'nden toplanmıştır. Özellik seçimi için makine öğrenimi algoritmaları, öğrencilerin başarısını etkileyen temel özellikleri belirlemek için kullanılmıştır. İlgili bir çalışmada öğrenci başarısının erken tahmini için farklı sınıflandırma ağacı yöntemleri (Ki-kare Otomatik Etkileşim Dedektörü (CHAID), kapsamlı CHAID, QUEST ve sınıflandırma ve regresyon ağacı (CART)) kullanılmıştır. Araştırmadan elde edilen temel bulgu, etnik köken, kurs programı ve kurs bloğunun öğrencilerin başarısını etkileyen ilk üç ana özellik olduğudur.

Dai Shangping (2008), web tabanlı bir eğitim sisteminin kayıtlarından yola çıkarak öğrencilerin final notlarını tahmin edecek bir sınıflandırma yaklaşımı sunmaktadır. Çalışmada birliktelik kurallarından yararlanılmış sonuçta ARMNGA (Yeni Genetik Algoritmada Birliktelik Kuralları Madenciliği) adı verilen yeni bir uzamsal madencilik algoritması önerilmiştir.

Albreiki vd. (2021) yürüttükleri sistematik derlemede, 2009'dan 2021'e kadar olan okul terki ve risk altındaki öğrencilerin belirlenmesiyle ilgili EDM literatürünü incelemiş, çalışmalarda genelde; öğrenci kolejleri/üniversite veri tabanlarından ve çevrimiçi öğrenme platformlarından elde edilen veriler olmak üzere iki tür veri kümesinin kullanıldığı görülmüştür. Ayrıca, çoğu Makine öğrenimi yöntemlerinin, risk altındaki öğrencilerin ve okulu bırakma oranlarının

tahmin edilmesinde önemli rol oynadığı ve böylece öğrencilerin performansını artırdığı doğrulanmıştır. Genel olarak bu derleme, öğrencilerin risk altında olma ve okulu bırakma durumlarını tahmin ederek öğrencilerin performansını artırma hedeflerine ulaşmış ve hem statik hem de dinamik verileri kullanmanın önemini vurgulamıştır.

Benzer şekilde Bonde ve Kirange'da (2018) öğrencilerin akademik performansını tahmin etmek için eğitimsel veri madenciliğini kullanan araştırmaları incelemiş ve yaklaşık 57 araştırmanın yazarı ile birlikte kullanılan veri madenciliği yöntemi ve bulgularını bir tablo halinde özetlemiştir. Sonuç olarak araştırmada, eğitimde veri madenciliği alanında halihazırda yapılmış araştırmalarını ayrıntılarını sunarak hem öğrencilerin takibini yapmak isteyen kurumlara hem de EDM yöntemlerini kullanan araştırmacılara parametrik bir yaklaşım önermektedir.

Nazempour (2023) çevrimiçi öğrenme ortamlarında öğrenme sonuçlarını iyileştirmek amacıyla yürüttüğü tez çalışmasında; eğitimde veri madenciliği ve öğrenme analitiği alanına katkı sunmaktadır. İlk katkısı, farklı öğretim modlarını istatistiksel olarak karşılaştırmak ve öğrencilerin akademik performansı üzerindeki etkilerini araştırmak için bir çerçeve model; ikinci katkısı, öğrencilerin öğrenme stili özelliklerini akademik performanslarını artırmak ve risk altında olanları kursun başlarında tespit etmek için kullanılabilecek bir çerçeve modeldir. Önerilen yaklaşım, öğrencilerin öğrenme stilleri ve demografik özelliklerini içerecek şekilde OULAD adlı kamuya açık bir veri kümesi kullanılarak test edilmiştir.

Hotaman (2020), okul ve öğrenci başarısının değerlendirilmesinde ve ölçme sonuçlarının analizinde karşılaşılan güçlükleri aşmada kullanılabilecek eğitsel veri madenciliği tekniklerini incelemiştir. Kuramsal derleme niteliğindeki çalışma sonucunda, okullarının organizasyonunda ve iyileştirilmesinde eğitsel veri madenciliğinden yararlanılması önerilmiştir.

Şengür ve Tekin (2013) öğrencilerin mezuniyet notlarını erken tahmin edebilecek bir veri madenciliği uygulaması gerçekleştirmek amacıyla yürüttüğü çalışmada, test ettiği iki senaryodan; birinci senaryoda öğrencilerin derslerinde ilk iki yılda aldıkları yılsonu notları; ikinci senaryoda ise ilk üç yılsonunda aldıkları derslerin yılsonu notları kullanılmış ve her ikisi için de Yapay Sinir Ağlarının, Karar Ağaçlarına göre daha iyi tahminler yaptığını göstermiştir.

Atif (2022), veri madenciliği teknikleri kullanılarak eğitim verilerinin analiz edilmesi ile öğrenci performansına katkıda bulunan temel nitelikleri belirlemek amacıyla Birleşik Arap Emirlikleri'nde, öğrencilerin akademik not ortalaması, İngilizce dil yeterliliği, öğrenci ve eğitmenin uyruğu ve yaşı ve öğrenci demografisi ile ilgili verilere odaklanarak yürüttüğü

çalışmada, çeşitli (KNN, Naïve Bayes, SVM, Lojistik regresyon, Karar Ağacı, Rastgele orman, Adaboost, Bagging Sınıflandırıcısı) makine öğrenimi algoritmalarıyla modeller geliştirmiştir. Sonuçlar, Rastgele orman modelinin diğer modellere kıyasla %90,12'lik doğruluk oranı ile daha iyi performans gösterdiğini; notlarla, derse katılım özelliğinin pozitif, sınıftaki öğrenci sayısı özelliğinin negatif korelasyonlu olduğunu ortaya koymuştur.

Ramesh ve diğerlerinin (2013) yürüttüğü çalışmanın kapsamı, final sınavlarında öğrencilerin performansını etkileyen faktörleri belirlemek ve risk altındaki öğrencilere zamanında ve uygun bir uyarı vermek amacıyla öğrencilerin notunu tahmin etmek için uygun bir veri madenciliği algoritması bulmaktır. Ortaöğretim öğrencilerinden birincil ve ikincil kaynaklardan oluşturulmuş veriler ile yürütülen araştırma, okul türünün öğrenci performansını etkilemediğini ve ebeveyn mesleğinin notları tahmin etmede önemli bir rol oynadığını ortaya koymaktadır. Bu makalede, bir öğrencinin performansını değerlendirmek için kullanılan Karar Ağacı ve Naive Bayes, Çok Katmanlı Algılama yöntemleri arasından %72'lik bir doğruluk ile Çok Katmanlı Algılama algoritmasının notları tahmin etmek için en uygun yöntem olduğu bulunmuştur.

Sembiring vd. (2011) öğrencilerin davranışları ile başarıları arasındaki ilişkileri analiz etmek için veri madenciliği yöntemlerini uygulayarak öğrenci performans tahmincileri modelini geliştirmek istemiş; araştırmada Destek Vektör Makinesi sınıflandırması ve çekirdek k-ortalamalar kümeleme (kernel k-means clustering) tekniklerini kullanmıştır. Anket aracılığı ile Universiti Malaysia Pahang'dan toplanan veriler (ilgi, çalışma davranışı, meşguliyet süresi, inanç ve aile desteği değişkenleri) ile öğrencilerin kişisel ve akademik kayıtlarından oluşmaktadır. Araştırma sonucunda oluşturulan modelde akademik performansın en önemli yordayıcısı psikometrik faktörler olarak bulunmuştur.

Yulia (2019) öğrencilerin ilk dört dönemdeki düşük performans nedeniyle okulu bırakmasını tahmin etmek amacıyla eğitimde veri madenciliği uygulamalarına örnek bir çalışma yürütmüştür. Araştırmanın ilk aşamasında K-ortalamalar algoritmasına dayalı bir tanımlayıcı ile çeşitli öğrenci kümelerini seçmiş ve öğrencilerin performansını modellemek için Karar Ağacı ve Naive Bayes sınıflandırma tekniğini kullanmıştır. İlk aşamada öğrencilerden kabul sırasında elde edilen veriler ile model test edilmiş sonrasında üniversite öncesi akademik kayıtların eklenmesiyle ikinci model geliştirilmiştir. Deneysel sonuçlar, öğrencinin sosyoekonomik bilgilerinin akademik performans sonuçlarını etkilediğini, akademik kayıt verileri eklendiğinde ise okulu bırakan öğrencinin tahmininin iyileştirildiğini göstermektedir. Ayrıca eğitim ve doğrulama veri kümesinde Bayes sınıflandırıcısının daha doğru sınıflandırma

yaptığı; karar ağacının ise her iki modelde de 0,5'in altında AUC değeri ile zayıf performans gösterdiği sonucuna ulaşılmıştır.

Taşdemir (2012) öğrencilerin akademik başarısına etki eden faktörlerin veri madenciliği ile tespiti için yaptığı tezinde ÖSYM tarafından gönderilen verilerden ve öğrencinin öğrenimi sırasında aldığı ders başarılarından yararlanmıştır. İki bağımlı değişken ve sekiz bağımsız değişken ile yürütülen analizler için Diskriminant analizi, Naive Bayes, Karar ağaçları, Sinir ağları, Genetik algoritma ve Regresyon analizi sınıflandırma teknikleri kullanılmış, sonuçlarda ise her bir bağımsız değişkenin etkinliği açıklanmıştır. Buna göre yaş ile not ortalaması arasında negatif yönlü; cinsiyet ile not ortalaması arası pozitif yönlü kuvvetli bir ilişki bulunmuştur.

Tüm bu araştırmaların yanında eğitimde veri madenciliği güncel pek çok araştırmada da kullanılmaktadır. Öztürk (2015), açık ve uzaktan öğrenme sistemlerinde Anadolu Üniversitesi açıköğretim programlarına kayıtlı öğrencilerin özelliklerini analiz ederek kümelere ayırmak, öğrenen türleri ve özelliklerine uygun stratejilerin belirlenmesine yön göstermek amacıyla ve Shyamala'da (2008) yükseköğretimde öğrenci verimliliğini tahmin etmek için veri madenciliği tekniklerini kullanmışlardır. Şahin (2018) e-öğrenme ortamlarında başarısızlık riski taşıyan öğrencilere erken müdahale edebilmek için öğrenme analitiklerini kullanarak makine öğrenmesine dayalı bir araç geliştirmiştir.

Bunun yanında eğimde veri madenciliği metodolojilerinin nasıl kullanıldığı ve ilgili algoritmaların teknik özellikleri hakkında çalışmalar (Akdemir, 2021; Alasadi & Bhaya, 2017; Aydın, 2007; Dol & Jawandhiya, 2023; Erdoğan, 2004; Poudyal, 2022; Schumann, 2005) ile dönem dönem eğitimde veri madenciliğinin gelişimine ışık tutan literatür taraması ile sistematik analiz çalışmalarının (Akgün & Özek, 2020; Bhise, Thorat & Supekar, 2013; Bi̇li̇ci̇ & Özdemi̇r, 2021; Dutt vd., 2017; Tekin & Öztekin, 2018) yapıldığı görülmüştür.

BÖLÜM III

YÖNTEM

Bu bölümde araştırma modeli, çalışma grubu ve prosedür, veri toplama yöntemi, verilerin işlenmesi ve analizi ile veri madenciliği başlıkları altında araştırma yöntemine ait öğeler yer almaktadır.

3.1. Araştırma Modeli

Araştırmada, yükseköğretimdeki açık ve uzaktan eğitim öğrencilerinin demografik ve kişisel verileri, akademik başarı puanları değişken olarak alınarak okul terki riski taşıyan öğrencileri erken tespit edebilmek için veri madenciliği algoritmaları ile bir model geliştirilmektedir.

Çalışmada kullanılan veri setinden anlamlı bilgiler çıkarmak için veri madenciliği yöntemleri kullanılmıştır. Veri madenciliği, eldeki veriler ışığında geleceğe yönelik tahmin yapılmasını sağlayacak kuralların belirlenmesi olarak tanımlanabilir (Şengür ve Tekin, 2013).

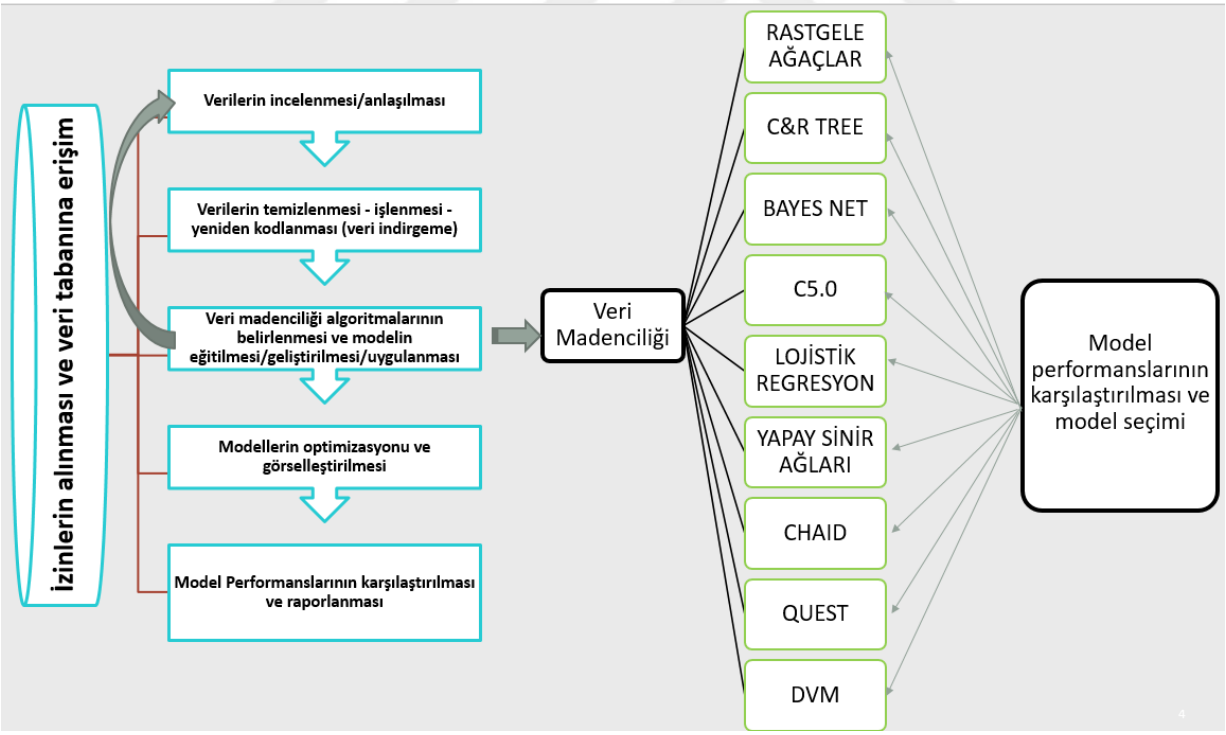
Araştırmanın türü, çalışma grubu, veri toplama araçları ve verilerin analizine yönelik işlemler, Veri Madenciliği için Alanlararası Standart Prosedürler (Cross Industry Standard Process Model for Data Mining: CRISP-DM) sürecinde yer alan aşamalar ile ilişkilendirilerek ele alınmıştır. Bu model, veri madenciliği sürelerini kullanılan endüstriyel alan ve yazılımdan bağımsızlaştırmayı amaçlayan Daimler-Benz, SPSS ve NCR tarafından geliştirilmiştir (Özcan, 2013).

Bu araştırma amacı bakımından nicel araştırma yöntemlerinden biri olan ilişkisel araştırma modelindedir. Korelasyonel araştırmalar, bir veya daha fazla nicel değişken ile bir veya daha fazla değişken arasındaki ilişkiyi araştırır (Fraenkel, Wallen & Hyun, 2012).

3.2. Çalışma Grubu ve Prosedür

Çalışma grubu, CRISP-DM sürecinde Veri Hazırlama aşamasına karşılık gelmektedir. Araştırma, İstanbul Üniversitesi Açık ve Uzaktan Eğitim Fakültesi'nde (İÜ AUZEF) 2021-2022 bahar dönemine kadar en az bir programa kayıt olan tüm öğrencilerin İstanbul Üniversitesi öğrenci bilgi sistemi (İÜ AKSİS) veri tabanından elde edilen veriler ile yürütülmüştür.

Buna göre araştırma süreci veri madenciliği standart prosedürüne uygun olarak Şekil 15'te belirtildiği gibi ilerlemiş ve ilk etapta resmi izinler alınmıştır. Daha sonra sırasıyla, verilerin incelenmesi, anlaşılması, değişkenlerin yorumlanması, eksik, gürültülü, hatalı verilerin temizlenmesi, verinin yeniden kodlanması ve indirgeme (özellik seçimi) işlemlerinden sonra sınıflandırma için uygun veri madenciliği algoritmalarının belirlenmesi, belirlenen modeller için geçerlik kriterlerinin oluşturulması, eğitim ve test verilerinin ayrıştırılması ile modelin eğitilmesi ve uygulanması aşamaları gerçekleştirilmiştir. Ardından modellere ait görseller incelenmiş, model performansları karşılaştırılmıştır. Bu süreç model performanslarını arttırmak için araştırmada yer alan değişkenlerin tekrar ele alınmasıyla optimum bir seviye yakalanana kadar birkaç kez tekrar etmiştir.



Şekil 15. Araştırma prosedürü

3.3. Verilerin Toplanması, Hazırlanması ve İşlenmesi

İÜ AKSİS veri tabanından birden fazla tablo halinde (demografik veriler, akademik başarı notları) CSV dosyası olarak elde edilmiştir. Demografik bilgileri içeren tablo pek çok değişken barındırmakta; bu değişkenlerin bir kısmı aynı verinin birden fazla tipini (örneğin program adı ve program kodu) gösterdiği için dikkate alınmamıştır. Aynı anlama gelecek veri sütunlarından sayısal değerli olanlar tercih edilmiştir. Bunun yanında zamana/yıllara göre biriken veride programlara kaydolan öğrencilerden kayıt esnasında istenen bilgilerin de değiştiği (muhtemel mevzuat gereği) ve bazı kişisel bilgilerin (örneğin medeni hâl bilgisi) son yıllarda alınmadığı görülmüştür. Bu sebeple veriler işlenmeden önce, program ve öğrenim tipi baz alınarak ayrıntılı inceleme yapılmış; bazı değişkenlerin (örneğin ayrılış tarihi) ne anlama geldiği kurumda görevli veri tabanı uzmanları ve kayıt görevlileri ile görüşülmüştür.

Verilerin işlenmesi ve temizlenmesi aşamasında MS Excel ile SPSS Statistics 25 paket programlarından yararlanılmıştır. Buna göre ilk aşamada;

- "Öğrenci numarası, kişi ID, program ID, fakülte ID" vb. tanılayıcı numaraları içeren veriler belirlenmiş ve filtrelenmiştir. Bu aşama kişisel verilerin korunması kapsamındaki veriler silinmiş; "TC kimlik numarası veya öğrenci numarası" gibi alanlarda rakamlar tekilliği koruyacak biçimde manipüle edilmiştir.
- Yüksek korelasyon olan veri tipleri belirlenmiş ve sadece biri kullanılmıştır (uyruk ile vatandaşlık durumu; ikametgâh ile yaşanılan il gibi…).
- Tüm veri setinde eksik vakalar tespit edilmiş ve veri setinden çıkartılmıştır (cinsiyet, uyruk, ikametgâh vb. alanlarında).

İÜ AKSİS sisteminin veritabanında elde edilen verilerde bazı değişkenlerin süreç içinde değişen kayıt kabul mevzuatı veya sistemsel veri kayıpları veya hatalı işlemler sebebi ile eksik veri barındıracağı aşikâr olduğundan eksik veri incelemesi yapılmış ve devamındaki işlemler buna uygun şekilde yürütülmüştür.

Eksik veri için veri tamamlama metodu üç sebepten kullanılmamıştır: bunlardan ilki verisi eksik olan öğrencilerin erken dönemde ayrılmış olduğu tespit edildiği için analizlerde yanlılığa yola açacağı endişesi; ikincisi VM'deki modeller için yeterli sayıda verinin zaten sette fazlasıyla yer alması ve üçüncüsü de doğal süreçte de kurumlara önerilebilecek nihai modelde bu durumun tabi olarak her zaman karşılaşılabilecek olması. Buna ek olarak "Medeni hal" verisinin ilk yıllardaki (2011-12) öğrencilerden temin edildiği ancak son yıllardaki kayıtlı öğrencilerden istenmediği; aynı zamanda süreç içinde değişebilecek bu verinin okul terki olduğu anda gerçek

durumunun tespitinin mümkün olmadığı görüldüğünden bu değişken analize dâhil edilmemiştir.

Verinin temizlenmesi aşamasından sonra çalışmanın amacına ve kullanılan SPSS Modeler 18 programının gerekliliklerine uygun olarak veri yeniden kodlanmıştır:

Program isimlerinde yeniden kodlama: Elde edilen veri dosyasında aynı öğrenim programını ifade eden isimler, tek program adı ile yeniden kodlanmıştır. Örneğin "Çocuk Gelişimi, Çocuk Gelişimi - %25 Burslu, Çocuk Gelişimi - %50 Burslu, Çocuk Gelişimi - Tam Burslu, Çocuk Gelişimi – Ücretli" isimli programların hepsi "Çocuk Gelişimi Programı" olarak kodlanmıştır.

Program düzeyinde yeniden kodlama: Program düzeyine göre yıllarda bazı programların isimleri ile birlikte düzeyleri de kodlandığı için (örneğin, sağlık önlisans, ilahiyat lisan tamamlama vb.) bu veriler yalnızca program düzeyi bilgisini içerecek biçimde yeniden kodlanmıştır. Bu şekilde; önlisans-lisans-lisans tamamlama olan 3 kategori oluşturulmuştur.

Öğrenci kayıt yılı verisinde yeniden kodlama: Veride bu değişken gün-ay-yıl olacak şekilde bulunduğu için kayıt tarihi verisi sadece "yıl" değeri dikkate alınarak yeniden kodlanmıştır.

Geliş/kayıt şekli verisinde yeniden kodlama: Bu değişken ham veride 28 farklı değer alıyorken 6 kategorili olacak şekilde biçimde kodlanmıştır.

Tablo 6
Geliş/Kayıt Şekli Verisinde Yeniden Kodlama İşlemi

Geliş Şekli(ham veri)	Geliş Şekli_1(yeni değer)
Sağlık Lisans Tamamlama	DGS
DGS	DGS
DGS Ek kontenjan	DGS
Af 7143	AF
ÖSS Ek Kontenjan	OSS
Sınavsız İkinci Üniversite	SINAVSIZ IKINCI UNIVERSITE
Standart Kayıt ÖSS	OSS
Merkezi Yerleştirme ile Yatay Geçiş Kurumlar arası	OSS
Merkezi Yerleştirme İle Yatay Geçiş Kurum içi	OSS
Yatay Geçiş Kurumlar arası Yurtiçi	OSS
YÖS	YOS
Yatay Geçiş Kurum içi	OSS
Özel Öğrenci Misafir	OZEL
YÖS Ek Kontenjan	YOS
Af 6569	AF
Af Yatay Geçiş 7143	AF
Özel Durumlarda Yatay Geçiş Ekmadde2	OZEL
Yatay Geçiş Kurumlar arası Yurtdışı	OSS
ÖSYM Tarafından Yerleştirilen 667KHK	OZEL
Fatih Üniversitesi KHK	OZEL
Özel Öğrenci 667khk	OZEL
Af 6111 Yatay	AF
Af 6111	AF
ILITAM	DGS
YÖK Kararı	OZEL
Yatay Geçiş Kurum dışı	OSS
Af KHK 7143	AF

Cinsiyet verisinde yeniden kodlama: Cinsiyet değişkeni 1 (erkek) - 0 (kadın) olarak yeniden kodlanmıştır.

Yaş verisinin yeniden kodlanması: Veri tabanından elde edilen öğrenci yaş verisi yıl bazlı sürekli bir veriyi göstermektedir. Bu veri literatür araştırmasına dayanarak 5 kategoriye ayrılmış ve yeniden kodlanmıştır.

Uyruk verisinin yeniden kodlanması: Uyruk verisi ilk aşamada 102 kategori içermektedir ve ilgili ülke kodları yeniden kodlanmıştır ancak veri daha sonra analize dahil edilmemiştir.

İkametgâh verisinin yeniden kodlanması: MERNİS sisteminden temin edilen ikametgâh bilgisinde il-ilçe-mezra düzeyinde olan veriden öncelikle 81 il verisi elde edilmiş daha sonra öğrenci yoğunluğuna göre veriler yeniden kodlanmıştır. Buna göre ikametgâhı ifade eden il

isimleri plaka kodu ile değiştirilmiş; sonrasında ikametgâh alanı en fazla kişinin yaşadığı yerlerden başlayarak 7 kategori olacak şekilde yeniden kodlanmıştır.

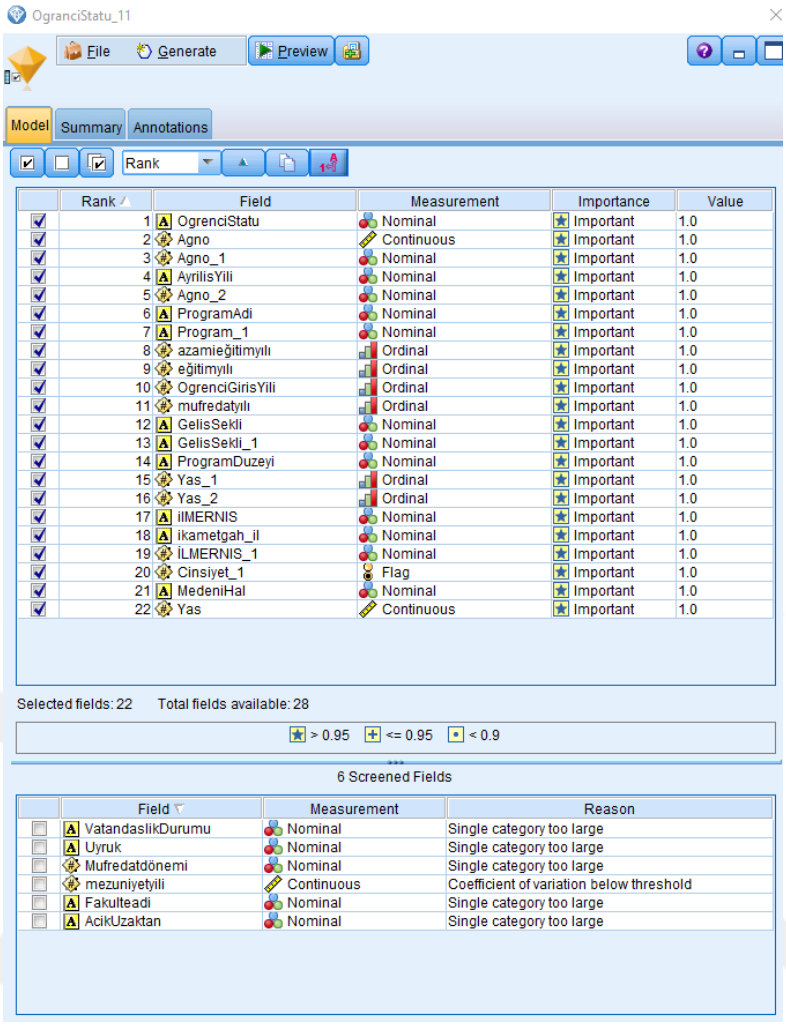
AGNO verisinin yeniden kodlanması: Ağırlık genel not ortalaması verisi 0 ile 4 arasında değişen sürekli bir veri iken terk tahmini için veri AGNO_1 türünde 4 kategorili olarak; Akademik başarı tahmininde yordanan veri olduğu için AGNO_2 sütununda 2 kategori (başarılı-başarısız) olarak yeniden kodlanmıştır.

Öğrenci statüsü verisinin yeniden kodlanması: Veri tabanında 11 kategori olarak yer alan öğrenci statü bilgisi bu araştırmanın yordanan değişkeni olduğu için devam eden ve terk eden olacak şekilde (0-1) iki kategoride yeniden kodlanmıştır.

Tablo 7
Öğrenci Statüsü Verisinin Yeniden Kodlanması

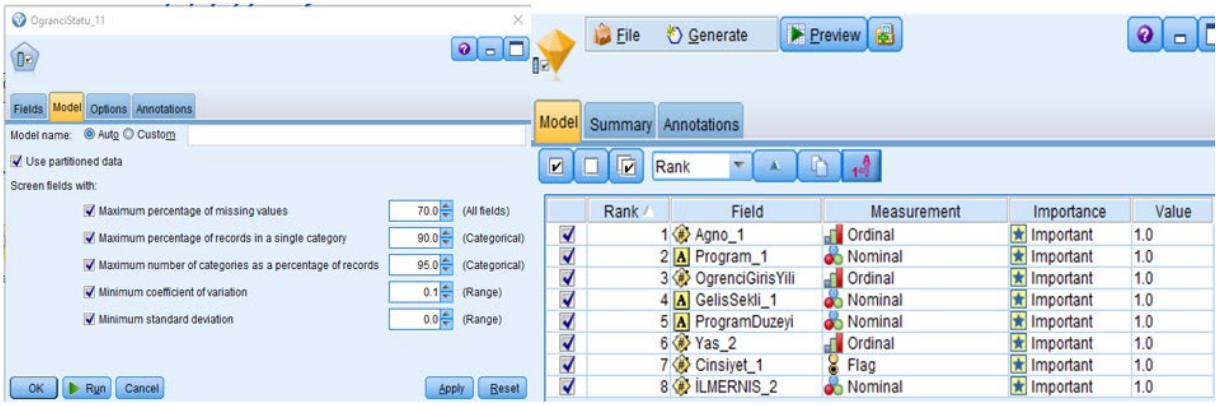
Öğrenci Statüsü (ham veri)	ÖgrenciStatü_1	Frekans
A Koşullu	DEVAMLI	49
B Koşullu	DEVAMLI	157
Bölüm Değişikliği	DEVAMLI	135
Devamlı Öğrenci	DEVAMLI	352992
Mezun	DEVAMLI	95774
Mezun Önlisans	DEVAMLI	311
Program Değişikliği	DEVAMLI	2
Ayrıldı	TERK	41536
Kaydı Silindi	TERK	198146
Kayıt Dondurmuş	TERK	6360
Kayıt Yenilemedi	TERK	203742

Veri alanlarının doğru olarak anlaşıldığına kanaat getirdikten sonra veri indirgeme işlemi yapılmış ve 28 değişken arasından analize dahli olası 22 değişken belirlenmiştir. Her ne kadar YSA yöntemlerinin regresyon analizine göre modele eklenen değişken sayısının fazla olması durumunda daha iyi tahmin yapacağı belirtilse de VM gibi tahmine dayalı yöntemlerde modele daha çok sayıda değişken eklemenin performans kestiriminin doğruluğunda bir artışa neden olmayacağı belirtilmiştir (Huang ve Fang, 2013; Lykourentzou vd., 2009; akt. Aksu & Doğan, 2013). Bu aşamada SPSS Modeler 18 programında yer alan "özellik seçim" fonksiyonu kullanılmıştır (Şekil 16).



Şekil 16. Tüm veri seti için değişkenlerin önemlilik testi

Daha sonra 22 değişkenden yüksek korelasyon içeren ve aynı tür bilgiyi barındıran değişkenler elenmiş ve geriye kalan 11 değişken için tekrar “Future Selection” fonksiyonu kullanılmıştır. Buna göre “uyruk ve medeni hâl” değişkenlerine modelde yer verilmemesine karar verilmiştir. Şekil 17’de yer alan özellik seçim kurallarına göre 9 niteliğin nihai analize dâhil edilmesine karar verilmiştir.



Şekil 17. Özellik seçim fonksiyonuyla elde edilen değişkenler

2011 yılında ilk defa öğrenci almaya başlayan İÜ AUZEF'deki açık ve uzaktan öğretim programlarında program sayısı yıllar içinde artmış olsa da öğrencilerin ilk kayıt oldukları yıl baz alınarak 2011 yılından 2022 bahar kayıt dönemine kadar var olan öğrencilerin tamamının kayıt esnasında beyan edilen (bir kısmı için daha sonraki yıllarda e-devlet üzerinden teyit edilen) kişisel verileri ve süreç içinde oluşan akademik başarı notları değişkenler olarak ele alınmıştır.

2021-22 eğitim öğretim yılı bahar dönemi itibari ile elde edilen veri setinde yaklaşık 899.000 öğrencinin verisi incelenmiş; bunların da bahar dönemi kayıt süreci tamamlandıktan sonra yaklaşık %45'nin diğerlerinin pasif olacak şekilde kayıt statüsünde olduğu görülmüştür. Programların öğrenci kabul etmeye başladığı yıllar değişkenlik göstermektedir. Sisteme dâhil olan öğrencilerin büyük bir kısmının ilk eğitim öğretim döneminden sonra devam veya terk kararı aldıkları gözlemlendiğinden ilk defa kayıt tarihinden bir sonraki dönemdeki öğrenci statüleri incelenmiş ve yordanan değişken olan okul terki, buna göre sınıflandırılmıştır. Araştırma kapsamında incelenecek programlar ve öğrenci kabulüne başladıkları ilk yıllar Tablo 8'de belirtilmiştir.

Bu araştırma temelde iki değişkene (okul terki ve akademik başarı) dair sınıfların belirlenmesine odaklandığı için; ilk iki araştırma sorusu için hedef değişken Öğrenci Statüsü (1,0), son araştırma sorusu için hedef değişken ise AGNO (0,1) verisidir. Buna göre araştırma sorularına göre veri setinde yer alan değişkenler ve tanımları ile olası değerleri Tablo 9'daki gibidir.

Tablo 8

Öğretim Programları ve Programın Öğrenci Kabulüne Başladığı Yıl

Program Adı ve Tipi	Başlangıç Yılı
Acil Durum ve Afet Yönetimi, Önlisans Programı, (Uzaktan Öğretim)	2015
Acil Yardım Ve Afet Yönetimi, Lisans Tamamlama Programı, (Açıköğretim)	2015
Adalet, Önlisans Programı, (Açıköğretim)	2016
Adalet, Önlisans Programı, (Uzaktan Öğretim)	2017
Bankacılık ve Sigortacılık, Önlisans Programı, (Uzaktan Öğretim)	2011
Coğrafi Bilgi Sistemleri, Önlisans Programı, (Uzaktan Öğretim)	2015
Coğrafya, Lisans Programı, (Açıköğretim)	2011
Çocuk Gelişimi, Lisans Programı, (Açıköğretim)	2018
Çocuk Gelişimi, Önlisans Programı, (Açıköğretim)	2018
Çocuk Gelişimi, Önlisans Programı, (Uzaktan Öğretim)	2018
Dış Ticaret, Önlisans Programı, (Uzaktan Öğretim)	2012
Egzersiz ve Spor Bilimleri, Lisans Programı, (Açıköğretim)	2019
Ekonometri, Lisans Programı, (Uzaktan Öğretim)	2012
Endüstri Mühendisliği, Lisans Programı, (Uzaktan Öğretim)	2015
Felsefe, Lisans Programı, (Açıköğretim)	2011
Gazetecilik, Lisans Programı, (Uzaktan Öğretim)	2011
Grafik Tasarımı, Önlisans Programı, (Açıköğretim)	2020
Halkla İlişkiler ve Tanıtım, Lisans Programı, (Uzaktan Öğretim)	2011
Halkla İlişkiler ve Tanıtım, Lisans Tamamlama Programı, (Uzaktan Öğretim)	2011
Hemşirelik, Lisans Tamamlama Programı, (Açıköğretim)	2016
Hukuk Büro Yönetimi ve Sekreterliği, Önlisans Programı, (Açıköğretim)	2016
İktisat, Lisans Programı, (Açıköğretim)	2011
İktisat, Lisans Programı, (Uzaktan Öğretim)	2011
İlahiyat Lisans Tamamlama Programı, Lisans Programı, (Uzaktan Öğretim)	2011
İnsan Kaynakları Yönetimi, Lisans Programı, (Açıköğretim)	2020
İş Sağlığı ve Güvenliği, Lisans Programı, (Açıköğretim)	2011
İş Sağlığı ve Güvenliği, Önlisans Programı, (Açıköğretim)	2011
İş Sağlığı ve Güvenliği, Önlisans Programı, (Uzaktan Öğretim)	2016
İşletme, Lisans Programı, (Açıköğretim)	2016
İşletme Yöneticiliği, Tezsiz Yüksek Lisans Programı, (Uzaktan Öğretim)	2011
İşletme, Lisans Programı, (Uzaktan Öğretim)	2012
Kamu Yönetimi, Lisans Programı, (Uzaktan Öğretim)	2018
Kültürel Miras ve Turizm, Önlisans Programı, (Açıköğretim)	2014
Laborant ve Veteriner Sağlık, Önlisans Programı, (Açıköğretim)	2011
Maliye, Lisans Programı, (Uzaktan Öğretim)	2011
Marka İletişimi, Önlisans Programı, (Açıköğretim)	2016
Medya ve İletişim, Önlisans Programı, (Uzaktan Öğretim)	2017
Perakende Satış ve Mağaza Yönetimi, Önlisans Programı, (Açıköğretim)	2017
Radyo, Televizyon ve Sinema, Lisans Programı, (Uzaktan Öğretim)	2011
Rekreasyon, Lisans Programı, (Açıköğretim)	2019
Sağlık Kurumları İşletmeciliği, Önlisans Programı, (Açıköğretim)	2015
Sağlık Yönetimi, Lisans Tamamlama Programı, (Açıköğretim)	2014
Sivil Hava Ulaştırma İşletmeciliği, Önlisans Programı, (Açıköğretim)	2011
Siyaset Bilimi ve Kamu Yönetimi, Lisans Programı, (Açıköğretim)	2011
Siyaset Bilimi ve Uluslararası İlişkiler, Lisans Programı, (Açıköğretim)	2016
Sosyal Hizmet, Lisans Tamamlama Programı, (Açıköğretim)	2016
Sosyal Hizmetler, Önlisans Programı, (Açıköğretim)	2011
Sosyoloji, Lisans Programı, (Açıköğretim)	2012
Tarih, Lisans Programı, (Açıköğretim)	2018
Tıbbi Dokümantasyon ve Sekreterlik, Önlisans Programı, (Açıköğretim)	2015
Uluslararası Ticaret ve Lojistik, Lisans Programı, (Açıköğretim)	2015
Yaşlı Bakımı, Önlisans Programı, (Açıköğretim)	2019
Yönetim Bilişim Sistemleri, Lisans Programı, (Açıköğretim)	2019

Tablo 9
Araştırmada Yer Alan Değişkenler ve Olası Değerleri

Değişken/Girdi	Değişkenin Tanımı	Olası değerleri
Program adı	Öğretime devam eden ilgili fakültedeki yükseköğretim programı	--
Program düzeyi	Ön lisans-lisans-lisans tamamlama	1, 2, 3
Öğrenim tipi (Acık-Uzaktan)	Açık öğretim, uzaktan öğretim	1, 2
Cinsiyet	Kadın - Erkek	0, 1
İkametgâh illeri (ILMERNIS)	İkamet edilen illeri gösterir değişken; İSTANBUL, ANKARA, İZMİR, BURSA, KOCAELİ, ANTALYA ve DİĞER İLLER olarak yeniden kodlanmıştır.	1, 2, 3, 4, 5, 6, 83
Kayıt/giriş şekli	Öğrencilerin programlara kabul koşullarını gösteren değişken Sınavlar ile (OSS-YOS-DGS), istisna durumlar ile (OZEL-AF) ve ikinci bir üniversite okuma programı (SINAVSIZ IKİNCİ UNIV.) kapsamında	AF, DGS, OSS, OZEL, YOS, SINAVSIZ IKINCI UNIV.
Öğrenci giriş yılı	Öğrencinin fakülteye kaydolduğu yılı gösterir (2010-2021 yılları arası)	2010, 2011,2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021
Yaş_2	Öğrencilerin yaş aralıkları;	
	YAŞ <26	1
	26 ≤ YAŞ ≤ 30	2
	31≤ YAŞ ≤ 35	3
	36≤ YAŞ ≤ 40	4
	YAŞ >40	5
AGNO_1	Ağırlıklı Genel Not Ortalaması aralıkları	
	0	0
	0<AGNO≤1	1
	1<AGNO≤2	2
	2<AGNO≤3	3
	3<AGNO≤4	4
AGNO_2	Ağırlıklı Genel Not Ortalaması 2 ve üstü başarılı; altı başarısız	1, 0
Öğrenci Statüsü	Öğrencinin okul terk veya devam durumu	1, 0

3.4. Verilerin Analizi

Araştırma modeline uygun olarak verilerin analizinde veri madenciliği yöntemleri kullanılmıştır. Bu çalışma bağlamında, veri anlama ve hazırlama süreçlerinin yürütülmesinde SQL veritabanı ve MS Excel 2016 kullanılmıştır. Bunu takiben, IBM SPSS Statistics 25, verilerin yeniden kodlanması, temizlenmesi, eksik değerlerin ele alınması, aykırı değerlerin belirlenmesi ve gürültülü verilerin ele alınması gibi görevleri üstlenmiştir. Çalışmada veri madenciliği algoritmalarının tamamı IBM SPSS Modeler 18 kullanılarak gerçekleştirilmiş ve analizde bu yazılımda bulunan veri madenciliği modelleri kullanılmıştır. Veri tabanından elde edilen verilerden hangilerinin en geniş kapsamda analize dâhil edilebileceği literatürün yönlendirmesiyle belirlenmiştir. Ardından, veri içinde etkili değişkenleri belirlemek ve özellik sayısını azaltmak için SPSS Modeler'daki "özellik seçimi" yöntemi kullanılmış ve daha ileri analizler için modellere dahil edilmek üzere 11 değişken seçilmiştir. Bu yaklaşım, öğrenci başarısızlığı riskini etkili bir şekilde öngörebilecek temel faktörleri belirlemeyi amaçlamaktadır.

Çalışmanın hedefleri ve veri setinin özellikleri doğrultusunda, analizde Rastgele Ağaçlar (RA), C&RT, C5.0, CHAID, QUEST gibi Karar ağaçları, Naive Bayes (NB), Lojistik Regresyon (LR), Yapay Sinir Ağları(YSA) ve Destek Vektör makinesi (DVM) gibi çeşitli sınıflandırma algoritmaları kullanılmış ve bunların performansları değerlendirilmiştir. Model geçerliğinin arttırılması amacıyla veri kümesi, %70'i eğitim verisi ve %30'u test verisi olacak şekilde iki alt kümeye ayrılmıştır. Bu bölünme ile eğitim veri seti dışındaki veriler üzerindeki model performansı değerlendirilmiş ve aşırı uyum azaltmıştır. Literatürde eğitim ile test veri oranlarının toplam veri miktarına göre belirleneceği; eğitim-test veri oranının yaklaşık ikiye bir oranında olmasının uygun ve yaygın bir kullanım olduğu görülmektedir (Aksu, 2018; Özcan, 2013). Son olarak çalışmada, veri madenciliği modelindeki değişkenlerin, okul terkini ve akademik başarıyı sınıflandırmada model doğruluğuna katkısı ve hiyerarşik etkisi incelenmiştir.

Sınıflandırma performansını değerlendirmek için karışıklık matrisine göre hesaplanan kriterlerden Doğruluk, Duyarlılık/Tekrar Arama, Hassasiyet/Özgüllük, F1 puanı, Kesinlik/Pozitif öngörü değeri, Negatif öngörü değeri (NÖD) ve Matthews Korelasyon Katsayısı (MCC) elde edilmiş; ROC eğrisi ile AUC değerleri de hesaplanarak incelenmiştir.

Veri madenciliği modellerinin gelişim süreci iki amaca yönelik olarak ilerlemiştir: büyük veri gruplarının özelliklerini ortaya çıkarmak (tanımlayıcı) veya bu verilere dayanarak tahminlerde

bulunmak (öngörücü). Bu çalışma, öğrencilerin akademik başarılarını "Başarılı" ve "Başarısız" olarak iki grupta kategorize etmiştir. Analizde Rastgele Orman, C&RT, C5.0, CHAID ve QUEST algoritmaları dahil olmak üzere çeşitli karar ağaçları modelleri ve ayrıca Naive Bayes, Lojistik Regresyon, Yapay Sinir Ağı ve Destek Vektör Makinesi gibi sınıflandırma algoritmaları kullanılmıştır. Veri madenciliği algoritmalarının ve sınıflandırma ağaçlarının temel özellikleri literatür bölümünde açıklanmıştır.

Bu araştırmada yararlanılan IBM SPSS Modeller paket programı, CRISP-DM modeli çerçevesinde sınıflandırma ve tahmin modelleri geliştirebilen bir veri madenciliği yazılımıdır. Karar ağaçları, makine öğrenmesi algoritmaları, çeşitli yapay sinir ağaçlarının yanı sıra, faktör analizi, regresyon gibi istatistiksel yönetmeleri ve çözümleri birlikte sunabilen ve veri tabanlarına bağlanarak analizleri geniş veri setleri ile yapabilen bir araçtır. SPSS Modeller programının diğer paket programlara göre görselliği oldukça iyi seviyededir (Ersöz, 2019). Çalışma ekranında sürükle-bırak ile nesne yerleştirme ve nesneleri birbirine bağlama işlemleri gibi pek çok işlem kolaylıkla yapılabilmektedir. Dolayısıyla IBM Modeller veri madenciliği aşamalarını gelişmiş bir teknoloji ile gerçekleştirme imkânı sunar. Verinin anlaşılması aşamasından model performanslarını karşılaştırma ve görselleştirmeye kadar tüm süreç program içinden yürütülebilmektedir. Veri kaynaklarına bağlanma, veriyi tanıma, verinin kalitesini anlama ve verinin grafiksel olarak incelenmesi, hipotezler oluşturmak amacıyla veri gruplarını değerlendirme çalışmalarında modeller, grafikler ve tablolar ile belli bölgelerin seçimini yapma seçeneği sunmaktadır (Ersöz, 2019).

3.5. Sınıflandırma Performansının Değerlendirilmesi ve Geçerlik

Model performanslarının değerlendirmesinde literatürde; bekletme (hold-out), k-katlı çapraz geçerleme (k-fold cross validation), kapsam dışı bırakma (leave-one-out) ve bootstrap olmak üzere dört yöntemin yaygın olarak kullanıldığı görülmektedir. Bekletme yaklaşımı aslında tüm verileri eğitim ve test verisi şeklinde belli oranlarda ayrıştırma işlemidir (Souza, Matwin & Japkowicz, 2002). Bu aşamada veri kümesinden test için belirli bir miktar (% 20 - % 30) tutulur ve bu işleme, bekletme prosedürü denilir. Daha sonra geriye kalan veriler eğitim için kullanılır. Pratik olarak eldeki verinin üçte birini (1/3) test için, geri kalan kısmını ise eğitim için kullanmak oldukça yaygın bir yöntemdir (Ahmed & Elaraby, 2014). Bekletme için verinin yeterli olmaması durumlarında genellikle k- katlı çaprazlama tercih edilir (Aksu & Doğan,

2018). Yeterli sayıda veriye sahip olunduğu için bu araştırma kapsamında bekletme yöntemi kullanılmıştır.

Veri madenciliğinde değerlendirme işleminin iki amacı bulunmaktadır. Bunlardan birincisi nihai modelin gelecekte ne kadar iyi çalışacağının belirlenmesi (hatta kullanılmasının gerekip gerekmeyeceği) ve ikincisi tüm eğitim veri setini en iyi temsil eden modelin ortaya çıkarılmasıdır (Souza vd., 2002). Özellikle sınıflandırma problemlerinde maliyet önemli bir problem olarak araştırmacıların karşısına çıkar; bazı alanlarda pozitif bir örneğin negatif (hasta olan bir kişinin sağlıklı olarak raporlanması) olarak veya negatif bir örneğin (sağlıklı bir kişiye hastalık teşhisi konması) pozitif olarak sınıflandırılması hatanın maliyetini değiştirebilir. Bu sebeple modellerin performansları değerlendirilirken araştırmanın amacı ve kurumsal maliyet riski göz önüne alınır. Buna bağlı olarak veri madenciliğinde elde edilen sonuçları sadece karışıklık matrisinden elde edilen doğruluk değerlerine göre değerlendirmek yerine duyarlık, hassasiyet, ROC eğrisi altında kalan alan, Matthew korelasyon katsayısı, F-ölçütü ve kesinlik gibi standart ölçütlerin de mutlaka raporlanarak elde edilen sonuçların güvenirlik ve geçerliğine ilişkin birden fazla değerin raporlanması önerilmektedir. Bu sayede güvenirliğe ve geçerliğe ilişkin birden fazla delil sunulmuş olacaktır. (Aksu & Doğan, 2018). Buradan yola çıkarak bu araştırmada veri madenciliği modellerinin performansını değerlendirmek için karışıklık (risk-sınıflandırma) matrisinden elde edilen; Doğruluk (Accuracy), Duyarlılık (Sensitivity/recall), Hassasiyet-Belirleyicilik-Özgüllük (Specificity), F1 puanı (F1 score), Pozitif öngörü değeri, Negatif öngörü değeri, Matthews Korelasyon Katsayısı kriterleri ve AUC değeri ile modeller incelenmiş ve raporlanmıştır.

Doğruluk, modelin doğru tahmin ettiği öğrenci sayısının toplam öğrenci sayısına oranıdır. Duyarlılık, gerçekte pozitif olan durumların model tarafından doğru tahmin edilme oranını; Hassasiyet-Belirleyicilik, gerçekte negatif olan durumların model tarafından doğru tahmin edilme oranını; F1 puanı, modelde Hassasiyet ve Duyarlılık arasındaki dengeyi ifade etmektedir. F1 ölçütü ne kadar yüksekse o kadar iyidir. Kesinlik, model tarafından doğru olarak sınıflandırılmış pozitif örnek sayısının, pozitif olarak sınıflandırılmış toplam örnek sayısına oranıdır. Bu ölçüt sınıflandırma modelinin, yanlış pozitifleri eleme kabiliyetini ölçmektedir. Negatif öngörü değeri ise modelin negatif olarak belirlediği durumlardaki gerçek negatif durumların oranını ifade eder. Diğer kriterlere göre literatürde daha az yer alan bir diğer kriter MCC ise özellikle dengesiz veri sınıflarındaki algoritmaların karşılaştırılmasında daha başarılı bulunmuştur (Bulut vd., 2022). Diğer kriterlerin aksine -1 ile 1 arasında değer alabilen MCC'de -1 mükemmel yanlış sınıflandırmayı, 1 ise mükemmel doğru sınıflandırmayı ifade eder. Sınıflar

arasındaki dengesizliği etkili bir şekilde yönetebilmesi sebebiyle, biyoloji ve tıp gibi alanlarda, örneğin hastalık teşhisinde sıklıkla kullanılır. Bu katsayı, bir sınıflandırıcının tahminlerinin kalitesini, sadece doğru sınıflandırmaların sayısına değil, yanlış ve doğru sınıflandırmaların türüne ve frekansına dayanarak değerlendirir.

Sınıflandırma (Confusion) matrisinde gözlem ve tahminlerin kesişim değerleri Doğru pozitif (True Positive -TP), Doğru negatif (True Negative - TN), Yanlış negatif (False Negative - FN), Yanlış pozitif (False Positive - FP) ile ifade edilmesi halinde çalışmada yer alacak değerlendirme kriterlerinde kullanılmak üzere ikili sınıflandırmalar için karışıklık-confusion matrisi Tablo 10'daki gibidir.

Tablo 10
İkili Sınıflandırmalar İçin Karışıklık(Confusion) Matrisi

		Tahmin edilen	
		1	0
Gözlenen (Gerçek)	1	Doğru pozitif –DP (True positive-TP)	Yanlış negatif –YN (False negative -FN)
	0	Yanlış pozitif –YP False positive (FP)	Doğru negatif- DN True negative (TN)

Tabloda gösterilen karışıklık-confusion matrisi değerlerine göre Doğruluk (Accuracy), Duyarlılık (Sensitivity/recall), Hassasiyet (Specificity), F1 puanı (F1 score), Pozitif öngörü değeri, Negatif öngörü değeri, Matthews Korelasyon Katsayısı (MCC) değerleri aşağıdaki gibi (Tablo 11) hesaplanmıştır.

Tablo 11

Sınıflandırma Performansının Hesaplanmasında Kullanılan Formüller

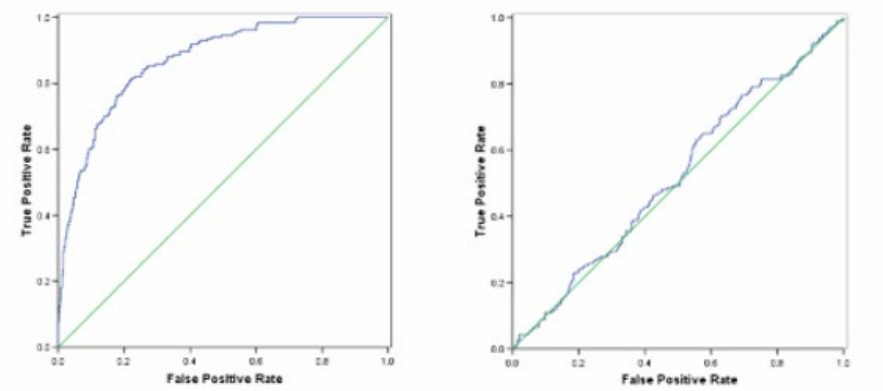
$$Doğruluk = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$
$$Duyarlılık = \frac{TP}{TP + FN}$$
$$Hassasiyet = \frac{TN}{TN + FP}$$
$$F_1 = 2x\frac{Duyarlılık\ x\ Hassasiyet}{Duyarlılık + \ Hassasiyet}$$
$$Pozitif\ öngörü\ değeri = \frac{TP}{TN + FN}$$
$$Negatif\ öngörü\ değeri = \frac{TN}{FN + TN}$$
$$Matthews\ Korelasyon\ Katsayısı: \frac{(TP\ x\ TN) - (FP\ x\ FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Alanda araştırmacılar bu araştırmanın da odak noktasına uygun olacak şekilde terk tahmini için okul terkini pozitif sınıf ve okul terki olmayanları negatif sınıf olarak kabul etmektedir. Burada TP doğru pozitif (gerçekte okuldan ayrılanların okul terki olarak sınıflandırılması), TN doğru negatif (gerçekte okula devam edenlerin devamlı olarak sınıflandırılması), FP yanlış pozitif (gerçekte devam ettiği halde öğrencilerin okul terki kategorisinde sınıflandırılması) ve FN yanlış negatif (gerçekte terk ettiği halde öğrencilerin devamlı kategorisinde sınıflandırılması) durumunu ifade eder. Araştırmanın amacına göre modellerde okulu terk edecek risk altındaki öğrencilerin belirlenmesi önceliklendirildiği zamanlarda TP-doğru pozitiflerin yüksek ve yanlış negatiflerin (okulu terk eden öğrencilerin devamlı öğrenciler olarak tahmin edilmesi) en düşük olduğu, yani daha iyi hassasiyete sahip modeller yerine yüksek duyarlılığa (recall) sahip modeller seçilecektir (Lottering vd., 2020). F-Ölçümü ise duyarlılık ve hassasiyet ölçütlerinin harmonik ortalaması olarak bu ölçütler arasında bir değiş tokuş olarak analiz edilecektir.

Yukarda karışıklık matrisinden elde edilen kriterlerin yanında sınıflandırma modelleri için grafik temelli performans değerlendirme ölçütleri de kullanılmaktadır. Bunlardan en yaygın

olan ROC (Receiver Operating Characteristic) eğrisi, modelin doğru pozitif oranını (True Positive Rate - TPR) yanlış pozitif oranına (False Positive Rate - FPR) karşı grafik üzerinde gösterir. ROC eğrisi, tanı koymak amacıyla kullanılan bir değişkenin değişim aralığı içinde değişkenin tüm değerleri sırasıyla kesim noktası kabul edilmesiyle hesaplanacak duyarlılık değerlerinin, testin yanlış pozitif oranına (1 - özgüllük) karşı noktalanması ile elde edilir (Aksu & Doğan, 2018). Bu eğri, modelin farklı kesme değerlerinde nasıl performans gösterdiğini anlamak için oldukça yararlıdır ve modelin hassasiyet ile özgüllük arasındaki dengesini görselleştirir. ROC eğirişinin altında kalan alan (AUC) da buna bağlı olarak yine bir perfromans göstergesidir. Eğrinin altında kalan alan (AUC - Area Under the Curve), modelin rastgele tahminlerden ne kadar iyi performans gösterdiğini gösterir. AUC değeri 1'e yaklaştıkça, modelin performansının o kadar iyi olduğu anlaşılır (Kayri & Şevgin, 2023). Kesme Değerinin Belirlenmesi, Modelin Duyarlılığının ve Özgüllüğünün Dengelemesi için ROC eğrisi incelenir. Bazı uygulamalar, yanlış alarm oranını minimize ederken doğru algılama oranını maksimize etmeyi gerektirebilir. ROC eğrisi, bu tür bir dengeyi sağlama konusunda rehberlik eder.

Model, hedef tepkinin var olduğu pozitif hız oranında (modelin hedef tepkiyi doğru tahmin ettiği durumlar) ROC eğrisi yükselirken; yanıtın olmadığı durumlarda (modelin yanıtı yanlış tahmin ettiği durumlar) doğru negatif oranı izler. İyi bir model için, eğri sol alt eksenin yakınından keskin bir şekilde yükselerek sağ üst köşeye doğru neredeyse birim kare altındaki tüm alanı kaplayacak şekilde ilerler (Şekil 18). Bilgi sağlamayan bir model için ise, eğri sol alt köşeden sağ üst köşeye doğru bir çapraz çizgi şeklinde ilerler (IBMCorporation, 2022).



Şekil 18. İyi bir model için (sol) ve bilgi sağlamayan bir model için (sağ) ROC eğrisi

AUC, tahmin edici değerlerin benzersiz kombinasyonları olan segmentler olarak tanımlanarak hesaplanır ki bu, hedef değerin tahmin edilen olasılığını taşıyan tüm kayıtların alt kümeleridir.

Bir modelin tahmin edicileri tarafından tanımlanan segmenti, tahmin edilen olasılık sırasına göre azalan düzende sıralanır ve AUC şöyle hesaplanır:

$$AUC = \sum_{i=1}^{S} (f_i - f_{i-1}) \left(\frac{t_i + t_{i-1}}{2} \right)$$

burada f_i , i segmenti için kümülatif yanlış pozitif sayısıdır. Yani i segmenti ve tüm önceki segmentler $j<i$ için yanlış pozitifler, t_i ise kümülatif doğru pozitif sayısıdır ve $f_0=t_0=0$'dır.

ROC eğrisi altındaki alanın beklenen değeri 0.50'dir. Mükemmel bir test için sıfır yanlış pozitif ve sıfır yanlış negatif ile alanın değeri 1.00 olacaktır. Eğri altında kalan alanın yorumlanmasında Aksu & Doğan (2018); 0,90-1,00 = mükemmel, 0,80-0,90 = iyi, 0,70-0,80 = orta, 0,60-0,70 = zayıf ve 0,50-0,60 = başarısız olarak sınıflandırmaktadır. Hosmer vd. (2013) ise Tablo 12'deki kriteri önermiştir (Kayri & Şevgin, 2023):

Tablo 12
AUC Değerlendirme Kriterleri (Hosmer vd., 2013)

AUC Değeri	Yorum
< 0,5	Ayrım belirtmez. Yazı tura atmaktan farksızdır.
0,5 < AUC ≤ 0,7	"Zayıf" ayrım
0,7 < AUC ≤ 0,8	"Kabul edilebilir" ayrım
0,8 < AUC ≤ 0,9	"Mükemmel" ayrım
0,9 < AUC < 1	"Olağanüstü" ayrım

Yapılan araştırmalarda özellikle veri kümeleri doğaları gereği dengesiz ise (örneğin yaklaşık 0,8 ila 0,2 oranında dengesizlik) ROC eğrisi, AUC değeri ve F1 ölçütü en etkili ölçütler olarak görülmektedir (Pulikottil & Gupta, 2020).

BÖLÜM IV

BULGULAR

Bu bölümde çalışmanın bulgularına yer verilecektir. Bulgular problem ve alt problemlere istinaden oluşturulan yapıya göre sunulmuştur. Ayrıca araştırma modeline uygun olarak verinin genel niteliğini anlamaya yönelik elde edilen bulgular verinin keşfi ve tanımlayıcı istatistikler başlığı altında verilmiştir. Buna göre bu bölümdeki alt başlıklar aşağıdaki şekilde belirlenmiştir:

4.1. Verinin keşfi ve tanımlayıcı istatistikler
4.2. Yükseköğretimde açık ve uzaktan öğrenmedeki öğrencilerin okul terkinin tahmininde veri madenciliği sınıflandırma algoritmalarının performansları nasıl değişmektedir?
4.3. Yükseköğretimde açık ve uzaktan öğrenmedeki öğrencilerin okul terki ile ilişkili demografik faktörler (yaş, cinsiyet, medeni durum, ikametgâh ili vb.) ve diğer nitelikler nelerdir?
4.4. AGNO'ya göre öğrencilerin sınıflandırılmasında VM modellerinin performansları nasıl değişmektedir ve öğrenci başarısı ile ilişkili faktörler nelerdir?

4.1. Verinin Keşfi ve Tanımlayıcı İstatistikler

Bu bölümde analizde yer alan değişkenlere ait tanımlayıcı istatistikler sunulmuş (tanımlayıcı veri analizi) buna göre veri madenciliği ile de edilen modelin yorumlanması ve anlaşılması için ön bilgiler elde edilmiştir. Veri madenciliği çalışmalarının ilk aşaması verinin anlaşılması olduğu için bu çalışma bulgular altında ilk adım olarak ele alınmıştır.

Analiz için veri tabanından toplamda 899 204 satırlı bir veri elde edilmiş; uç değerler, eksik ve mükerrer veriler temizlendikten araştırma sorularına uygun olarak veri iki ayrı sete bölünmüştür. Eksik ve mükerrer veri temizlendikten sonra elde edilen tüm veri seti 650.317

kişi; en az bir dersten başarılı olan öğrencilere ait veri seti (AGNO>0) 493.239 kişi olacak şekilde oluşmuştur.

Tablo 13
Veri Setleri ve Toplam Veri Miktarı

Veri seti	Veri sayısı
Ham veri seti	899.204
Tüm veri seti	650.317
AGNO>0 veri seti	493.239

Tablo 13'te araştırma sorularına istinaden veri madenciliği analizlerinin sürdürüleceği tüm veri ve AGNO>0 veri setleri için yapılan işlemlerin ardından elde edilen veri sayıları belirtilmiştir.

Öğrenim tipi değişkeni:

Türk yükseköğretiminde örgün eğitim dışındaki programlar açık öğretim veya uzaktan eğitim olarak iki türde açılabilmekte ve uzaktan eğitim türündeki programlara daha sınırlı öğrenci kontenjanı tanımlanmaktadır. Öğrenci sayılarının öğrenim tipine göre dağılımı Tablo 16'daki; bunların devam-terk durumuna göre oranları ile oluşturulan grafiği Şekil 22'deki gibidir.

Tablo 14
Öğrenim Tipine Göre Öğrenci Sayıları

Program Türü	Devam Eden		Terk Eden		Toplam
	Sayı	%	Sayı	%	
Açıköğretim	314.919	50.4	310.064	49.6	624.983
Uzaktan Öğretim	13.612	53.7	11.722	46.3	25.334
Toplam	328.531	50.5	321.786	49.5	650.317

Tablo 14 incelendiğinde hem açıköğretim hem de uzaktan eğitimde devam eden ve terk eden öğrenci oranlarının birbirine yakın olduğu görülmektedir.

Program değişkeni:

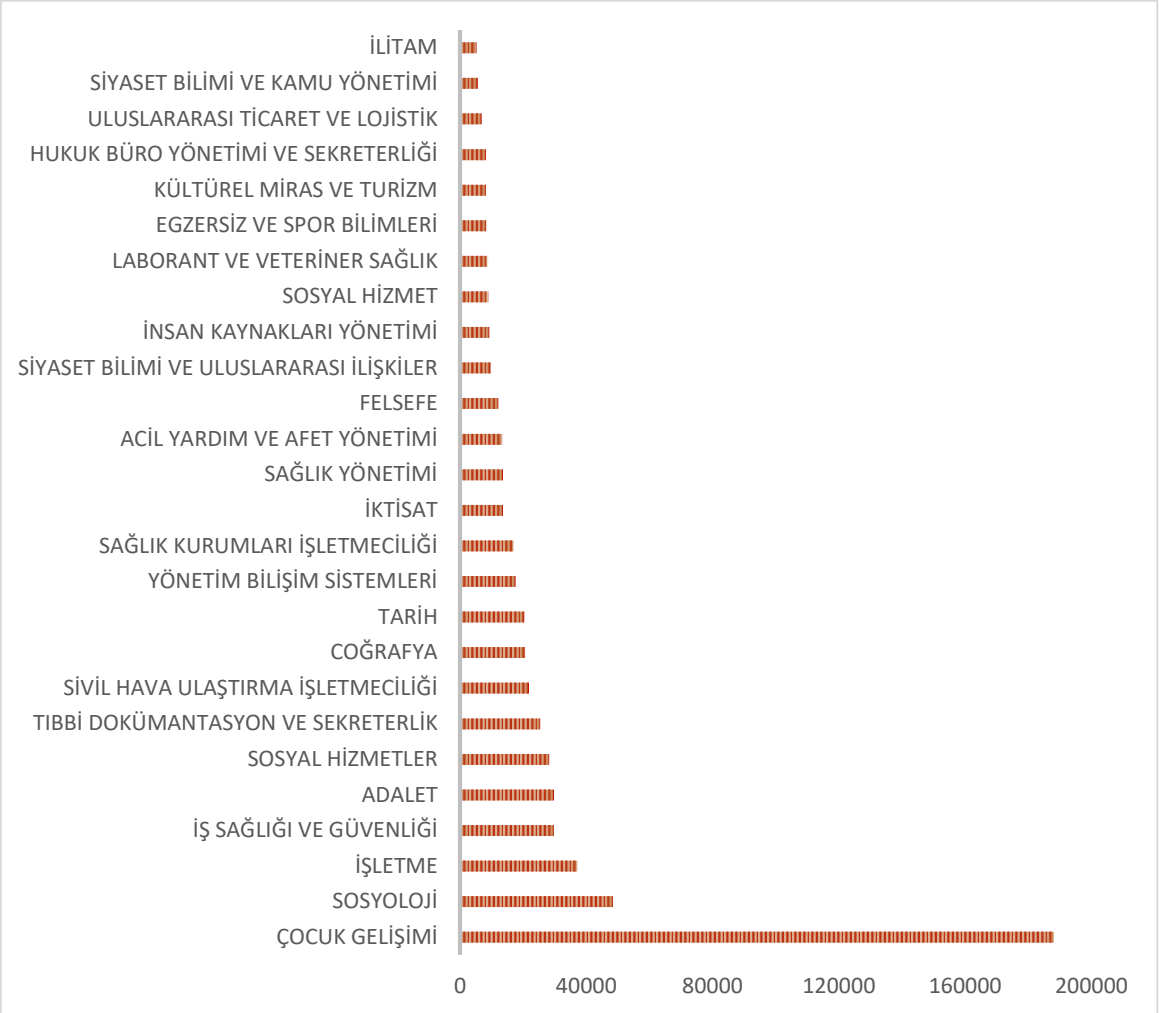
2011 yılından itibaren öğrenci kabul eden programlardaki analize dâhil edilen öğrenci, sayıları ve bu öğrencilerin devam-terk oranları Tablo 15'te yer almaktadır.

Tablo 15
Analizdeki Tüm Veriye Göre Öğretim Programlarındaki Öğrenci Sayıları

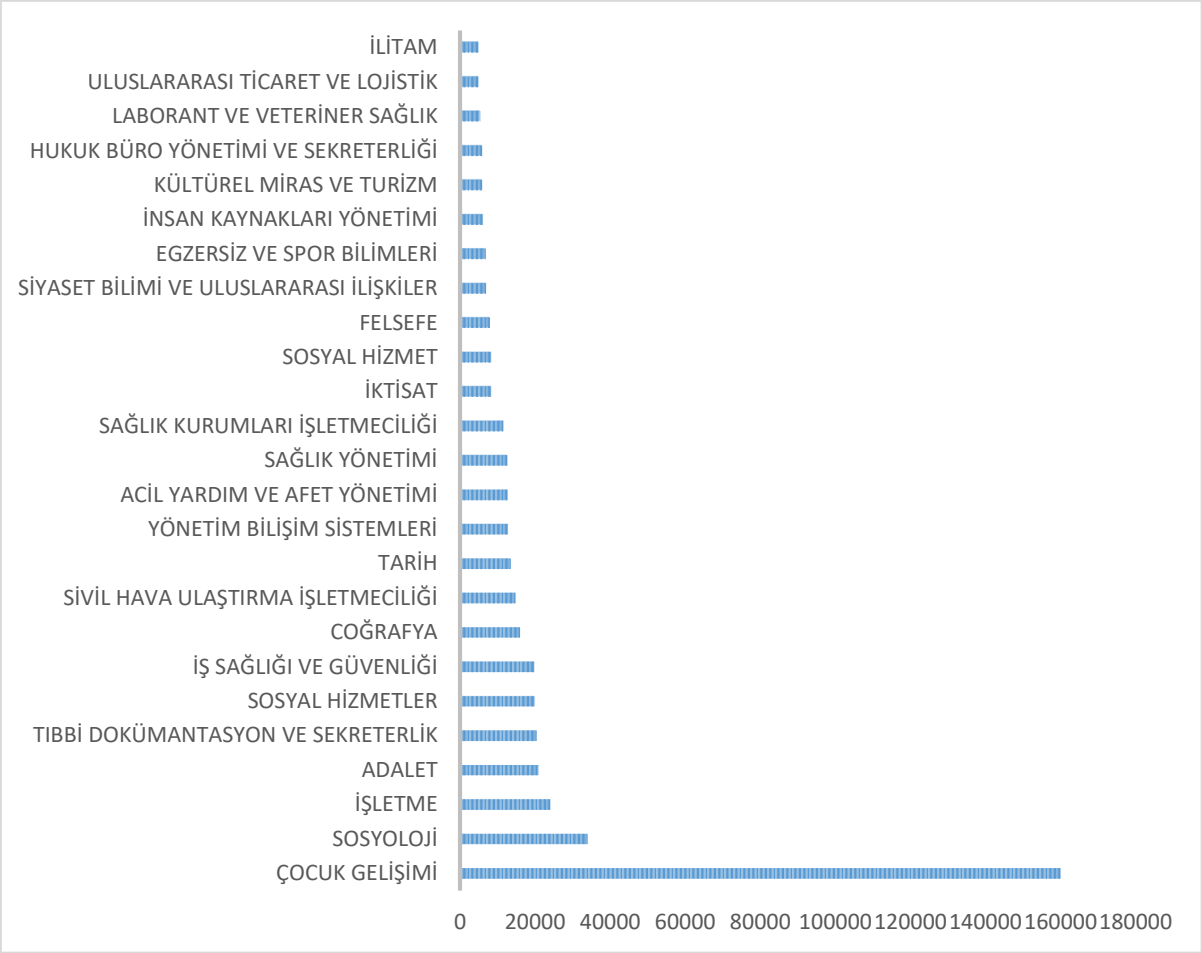
Program	Devam Eden		Terk Eden		Toplam
	Sayı	%	Sayı	Yüzde	
Acil Durum ve Afet Yönetimi	448	39,80	677	60,20	1125
Acil Yardım Ve Afet Yönetimi	11481	86,60	1780	13,40	13.261
Adalet	14.081	47,40	15.657	52,60	29.738
Bankacılık ve Sigortacılık	283	32,70	582	67,30	865
Bilgisayar Programcılığı	0	0,00	2	100,00	2
Bilgisayar ve Öğretim Teknolojileri Öğretmenliği	16	61,50	10	38,50	26
Coğrafi Bilgi Sistemleri	290	33,50	575	66,50	865
Coğrafya	8152	39,20	12.637	60,80	20.789
Çağdaş Türk Lehçeleri ve Edebiyatları	3	14,30	18	85,70	21
Çalışma Ekonomisi ve Endüstri İlişkileri	236	35,60	427	64,40	663
Çocuk Gelişimi	132.781	70,70	55.042	29,30	187.823
Dış Ticaret	359	40,10	537	59,90	896
Egzersiz ve Spor Bilimleri	4370	51,30	4154	48,70	8524
Ekonometri	22	29,70	52	70,30	74
Endüstri Mühendisliği	520	65,50	274	34,50	794
Felsefe	4289	35,50	7790	64,50	12.079
Gazetecilik	599	49,10	622	50,90	1221
Grafik Tasarımı	2013	46,40	2324	53,60	4337
Halkla İlişkiler ve Tanıtım	2572	54,50	2144	45,50	4716
Hemşirelik	1406	78,30	389	21,70	1795
Hukuk Büro Yönetimi ve Sekreterliği	2521	30,30	5802	69,70	8323
İktisat	3497	25,70	10.101	74,30	13.598
İLİTAM	4389	83,90	841	16,10	5230
İnsan Kaynakları Yönetimi	3350	36,30	5870	63,70	9220
İş Sağlığı ve Güvenliği	12.157	40,60	17.769	59,40	29.926
İşletme	10.486	28,20	26.634	71,80	37.120
Kamu Yönetimi	525	50,20	521	49,80	1046
Kültürel Miras ve Turizm	3509	41,40	4971	58,60	8480
Laborant ve Veteriner Sağlık	3350	38,00	5473	62,00	8823
Maliye	245	34,20	471	65,80	716
Marka İletişimi	1139	47,40	1264	52,60	2403
Medya ve İletişim	535	42,00	739	58,00	1274
Perakende Satış ve Mağaza Yönetimi	32	15,70	172	84,30	204
Radyo, Televizyon ve Sinema	652	50,10	650	49,90	1302
Rekreasyon	909	52,80	811	47,20	1720
Sağlık Kurumları İşletmeciliği	5493	32,20	11.569	67,80	17.062
Sağlık Yönetimi	10.861	80,00	2711	20,00	13.572
Sivil Hava Ulaştırma İşletmeciliği	7037	32,30	14.779	67,70	21.816
Siyaset Bilimi ve Kamu Yönetimi	2172	38,70	3443	61,30	5615
Siyaset Bilimi ve Uluslararası İlişkiler	3760	38,50	6016	61,50	9776
Sosyal Hizmet	7208	80,20	1784	19,80	8992
Sosyal Hizmetler	11.000	39,00	17.232	61,00	28.232
Sosyoloji	17.228	35,60	31.186	64,40	48.414
Tarih	7002	34,40	13.365	65,60	20.367
Tıbbi Dokümantasyon ve Sekreterlik	13.001	51,30	12.341	48,70	25.342
Türk Dili ve Edebiyatı	69	43,10	91	56,90	160
Uluslararası Ticaret ve Lojistik	2664	39,10	4143	60,90	6807

Program	Devam Eden		Terk Eden		Toplam
	Sayı	%	Sayı	Yüzde	
Uluslararası Ticaret ve Lojistik Yönetimi	1298	32,50	2699	67,50	3997
Yaşlı Bakımı	1906	52,70	1712	47,30	3618
Yönetim Bilişim Sistemleri	6615	37,70	10.933	62,30	17.548
Toplam	328.531	50,52	321.786	49,48	650317

Tablo 15'te gösterilen öğretim programlarındaki öğrenci sayılarına göre en fazla öğrencisi olan programların dağılımı Şekil 19'daki grafikte (17 program); AGNO notu sıfırdan büyük olan öğrencilerin programlara göre dağılımı ise Şekil 20'de gösterilmiştir.



Şekil 19. Programlardaki öğrenci sayıları (Tüm veri)



Şekil 20. Programlardaki öğrenci sayıları (AGNO>0)

Şekil 19 ve Şekil 20'de gösterilen her iki veri setinde de en fazla öğrenci, sırasıyla; Çocuk Gelişimi, Sosyoloji, İşletme, Adalet, Tıbbi Dokümantasyon ve Sekreterlik, Sosyal Hizmetler, İş Sağlığı ve Güvenliği programlarında iken en az öğrencinin Bilgisayar Programcılığı, Bilgisayar ve Eğitim Teknolojileri Öğretmenliği, Çağdaş Türk Lehçeleri ve Edebiyatları, Ekonometri, Perakende Satış ve Mağaza Yönetimi programlarında olduğu görülmektedir.

Program düzeyi değişkeni:

Program düzeyine göre; lisans, ön lisans ve lisans tamamlama programlarında tüm veri grubunda analize dahil olan öğrencilerin sayısı Tablo 16'da gösterilmiştir.

Tablo 16
Program Düzeyine Göre Öğrenci Sayıları

Program Düzeyi	Devam Eden		Terk Eden		Toplam
	Sayı	%	Sayı	%	
Lisans Programı	206.047	51,20	196.531	48,80	402.578
Lisans Tamamlama Programı	32.772	79,90	8253	20,10	41.025
Önlisans Programı	89.712	43,40	117.002	56,60	206.714
Toplam	328.531	50,50	321.786	49,50	650.317

Tablo 16 incelendiğinde lisans düzeyinde devam ve terk eden öğrenci sınıfları birbirine yakın iken önlisans düzeyinde terk eden öğrenci; lisans tamamlama düzeyinde ise devam eden öğrenci sayısı fazla görünmektedir.

Giriş/kayıt yılı değişkeni:

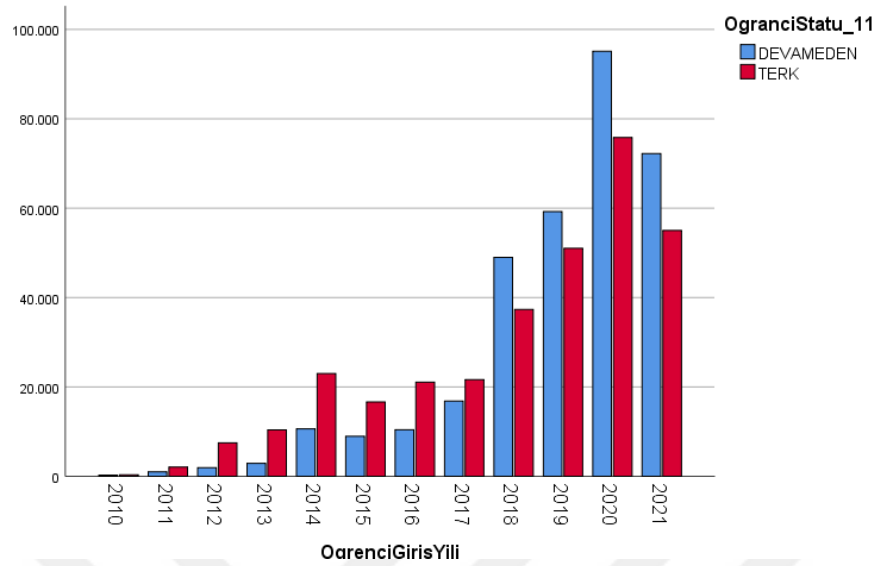
Analize dâhil edilen yıllar içinde programlara kaydolan öğrenci sayılarının 2010 ile 2021 yılları arasındaki değişimi Tablo 17'de gösterilmiştir. Buna göre genel olarak her yıl öğrenci kayıt sayısının arttığı; yıllara göre ise devam eden ve terk eden öğrenci oranlarının değiştiği gözlenmektedir.

Tablo 17
Giriş Yılı Değişkenine Göre Öğrenci Sayıları

Öğrenci Giriş/Kayıt Yılı	Devam Eden		Terk Eden		Toplam
	Sayı	%	Sayı	%	
2010	259	42,80	346	57,20	605
2011	1012	32,80	2074	67,20	3086
2012	1930	20,50	7472	79,50	9402
2013	2891	21,80	10370	78,20	13.261
2014	10.628	31,60	22.983	68,40	33.611
2015	8964	35,00	16.633	65,00	25.597
2016	10.429	33,10	21.082	66,90	31.511
2017	16.849	43,80	21.621	56,20	38.470
2018	48.998	56,70	37.354	43,30	86.352
2019	59.242	53,70	51.024	46,30	110.266
2020	95.106	55,60	75.808	44,40	170.914
2021	72.223	56,80	55.019	43,20	127.242
Toplam	328.531	50,50	321.786	49,50	650.317

Tablo 17'ye göre, 2021 yılında öğrencilerin %43,2'si; 2020 yılında %44,4'ü; 2019 yılında %46,3'ü; 2018 yılında %43,3'ü; 2017 yılında %56,2'si; 2016 yılında %66,9'u; 2015 yılında %65'i; 2014 yılında %68,4'ü; 2013 yılında %78,2'si; 2012 yılında %79,5'i; 2011 yılında

%67,2'si; 2010 yılında ise % 57,2'si okulu terk etmektedir. Yıllara göre devam ve terk dağılımları grafik ile Şekil 21'de gösterilmiştir.



Şekil 21. Giriş yılı değişkenine göre öğrenci devam/terk durumlarının dağılımı

Şekil 21'de de görüldüğü gibi 2017 ve öncesinde terk eden öğrenci sayısının devam edenlere oranla fazla olduğu; 2018'den itibaren genel olarak terk sayıları artsa da devam eden öğrenci sayısına göre daha düşük olduğu gözlenmektedir.

Kayıt/geliş şekli değişkeni:

Öğrenciler mevzuatta tanımlı birden fazla yol ile yükseköğretimde açık ve uzaktan öğretim programlarına kaydolabilmektedir. Özellikle 2017 yılından itibaren herhangi bir sınava girmeden (daha önce en az bir üniversite programına yerleşmiş veya bitirmiş olanlar) açık öğretim programlarına öğrenci kaydının yapılabilmesi kayıt şekillerini çeşitlendirmiştir.

Tablo 18
Açık ve Uzaktan Öğrenme Programlarındaki Öğrencilerin Geliş/Kayıt Şekli

Geliş (Kayıt) Şekli	Devam Eden		Terk Eden		Toplam
	Sayı	%	Sayı	%	
OSS/YKS	67.057	41,30	95.349	58,70	162.406
SINAVSIZ IKINCI UNIV	209.627	50,10	209.039	49,90	418.666
DGS	51.431	75,10	17.030	24,90	68.461
YOS	259	57,70	190	42,30	449
AF	93	58,10	67	41,90	160
OZEL	64	36,60	111	63,40	175
TOPLAM	328.531	50,50	321.786	49,50	650.317

Tablo 18'de öğrencilerin kayıt şekline göre terk ve devam durumları incelendiğinde DGS sınavı ile kaydolan öğrencilerin diğerlerine göre terk oranlarının düşük olduğu, sınavsız ikinci üniversite kapsamında bir programa kaydolan öğrencilerin terk ve devam oranlarının ise birbirine yakın olduğu söylenebilir.

Uyruk değişkeni:

Uyruk değişkeni kategorileri 51 ayrı birim-ülke barındırmış olmasına rağmen öğrencilerin tüm veri içinde %99,85'i "Türk" olduğu; diğer kategoriler yeterli veri içermediği için bu değişken analize dâhil edilmemiştir.

Tablo 19
Uyruk Değişkenine Göre Öğrenci Sayıları

Uyruk/Ülke	Frekans	Yüzdelik Oran
Türkiye	649.350	99,85
Suriye	559	0,1
Türkmenistan	96	0,0
Azerbaycan	71	0,0
İran	30	0,0
Çin	25	0,0
Özbekistan	25	0,0
Mısır	24	0,0
Afganistan	17	0,0

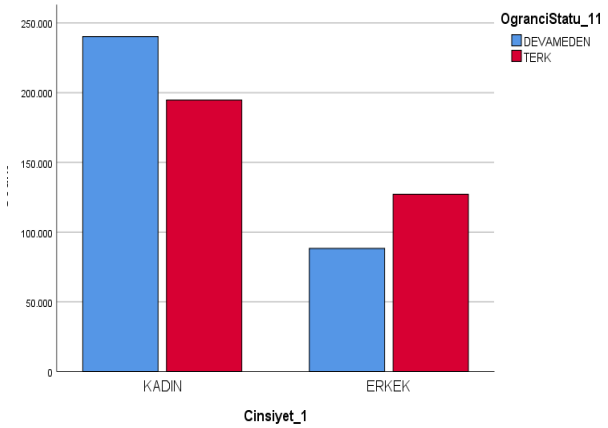
Cinsiyet değişkeni:

Cinsiyet değişkenine göre tüm veri içinde kadınlarda devam eden öğrenci oranının, erkeklerde ise terk eden öğrenci oranının daha yüksek olduğu görülmektedir. Tablo 20 incelendiğinde veri içinde toplam kadın öğrenci sayısının erkek öğrenci sayısının yaklaşık iki katı olduğu tespit edilmiştir.

Tablo 20
Cinsiyet Değişkenine Göre Öğrenci Sayıları

Cinsiyet	Devam Eden		Terk Eden		Toplam
	Sayı	%	Sayı	%	
Kadın	240221	55,20	194656	44,80	434877
Erkek	88310	41,00	127130	59,00	215440
Toplam	328531	50,50	321786	49,50	650317

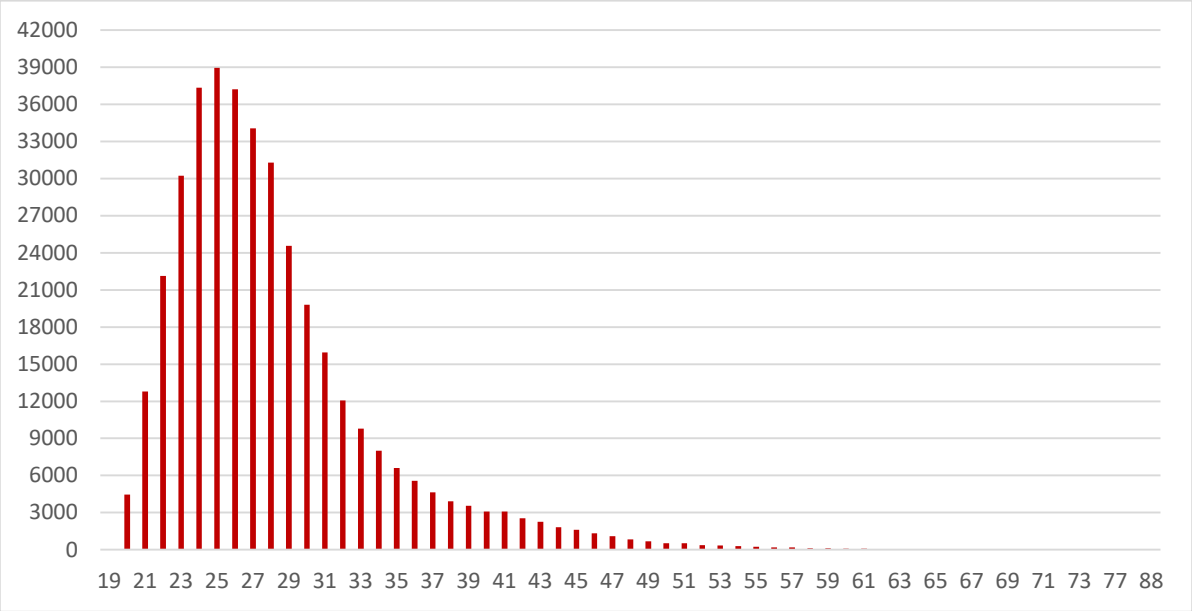
Cinsiyet değişkenine göre öğrencilerin terk ve devam durumu Şekilde 22'de görülmektedir.



Şekil 22. Cinsiyet değişkenine göre öğrencilerin devam/terk durumları

Yaş değişkeni:

Açık ve uzaktan öğrenme programlarının doğası gereği veri için yaş aralığının geniş olduğu görülmektedir. Buna göre sürekli veri olarak ham veride yer alan yaş değişkenin dağılımı Şekil 23'teki gibidir. Yaş verisi sınıflandırma algoritmasında performansın arttırması amacıyla kategorik veri olarak kullanılacaktır.



Şekil 23. Tüm veri için öğrenci yaş dağılımı

Kategorik değişkene çevrilerek beş kategori olacak biçimde yeniden düzenlenen yaş verisine göre tüm veri içinde 26 yaşından küçük 171.238; 26-30 yaş arası 178.447; 31-35 yaş arası

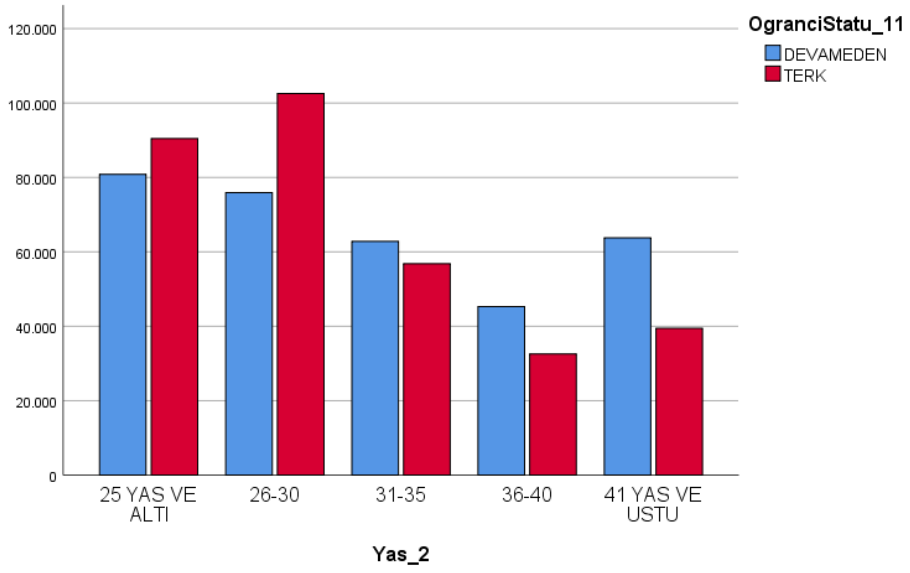
119.597; 36-40 yaş arası 77.825; 40 yaş üstü 103210 kişi bulunmaktadır. Bu kategorilerdeki devam eden ve terk eden öğrenicilerin sayıları Tablo 21'de yer almaktadır.

Tablo 21

Yaş Kategorilerindeki Öğrenci Sayıları (Tüm Veri)

Yaş Aralıkları	Devam Eden		Terk Eden		Toplam
	Sayı	%	Sayı	%	
25 Yas ve altı	80.810	47,20	90.428	52,80	171.238
26-30	75.907	42,50	102.540	57,50	178.447
31-35	62.793	52,50	56.804	47,50	119.597
36-40	45.263	58,20	32.562	41,80	77.825
41 Yas ve üstü	63.758	61,80	39.452	38,20	103.210
Toplam	328.531	50,50	321.786	49,50	650.317

Yaş kategorilerine göre devam eden ve terk eden öğrencilerin oranları Şekil 24'te gösterilmiştir.

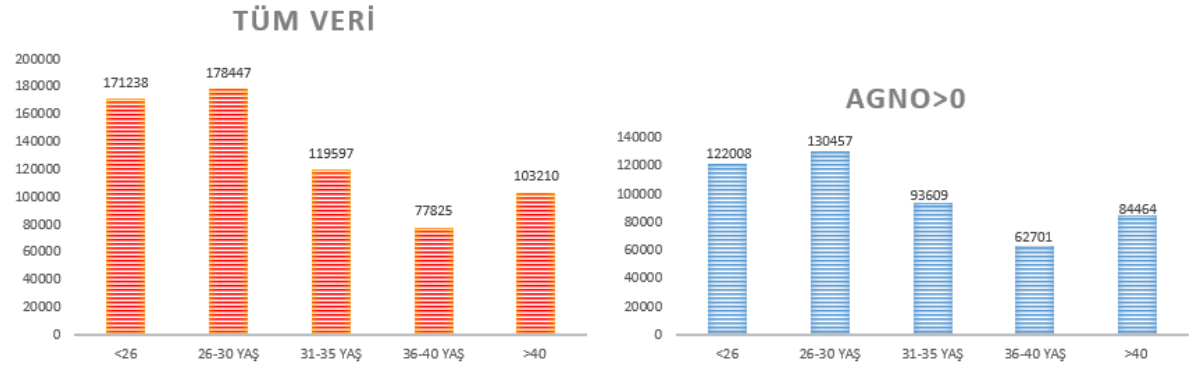


Şekil 24. Yaş kategorilerine göre öğrencilerin terk durumları

Şekil 24'te görüldüğü gibi 5 kategoride ele alınan yaş değişkeninde 30 yaş ve altındakilerde terk eden öğrenci sayısı devam eden öğrenciye göre daha fazladır. 41 yaş ve üstünün ise terk oranları oldukça azalmaktadır.

Ağırlıklı not ortalaması sıfırdan büyük olanlardan yani AGNO>0 veri setinde 122.008'i 26 yaşından küçük; 130.457'si 26-30 yaş arası; 93.609'u 31-35 yaş arası; 62.701'i 36-40 yaş arası

ve 84.464'ü 40 yaş üstüdür. Her iki veri setinde dağılımlara ait sütün grafikleri Şekil 25'te verilmiştir.



Şekil 25. Yaş değişkeninin veri gruplarında dağılımı

Şekli 25'te her iki veri setinde yaş kategorisinin dağılımı görülmektedir. Her iki veri setinde de en fazla öğrenci sayısı 26-30 yaş aralığında iken en az öğrenci sayısı 36-50 yaş aralığındadır.

İl/ikametgâh değişkeni:

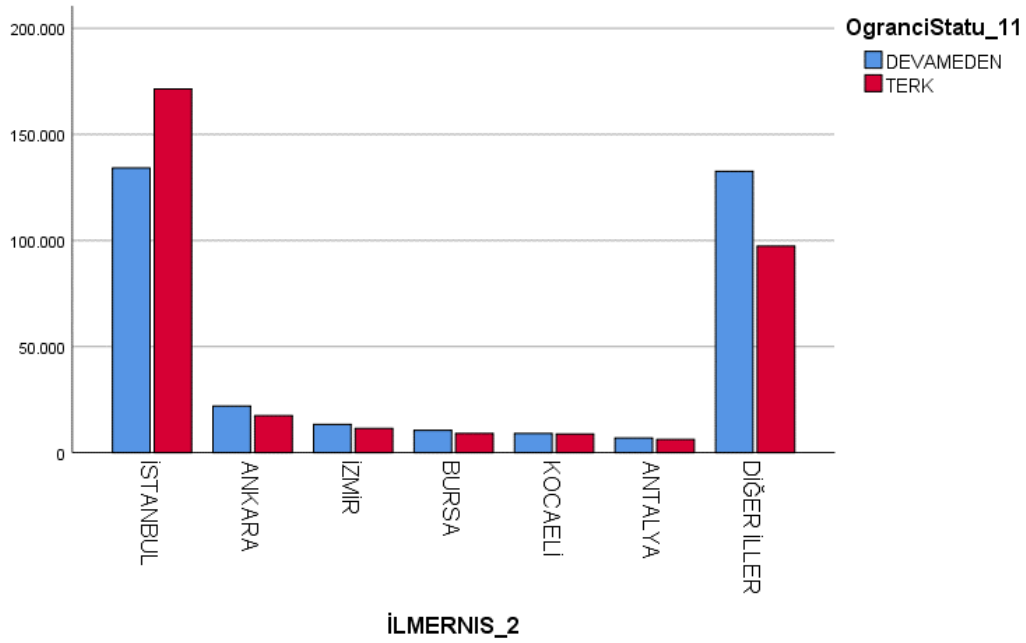
Tüm veri grubunda MERNİS'te (Merkezi Nüfus İdaresi Sistemi) kayıtlı ikamet ili değişkenine göre 81 ilde değişen sayılarda öğrenci bulunmaktadır. Öğrenci sayılarını dikkate alarak yeniden kodlanan veri için kategori azaltılmış ve en yoğun öğrenci barındıran ilk 6 il kendi adı ile ve kalan iller "diğer" kategorisinde belirtilmiştir. Buna göre öğrenci il/ikamet sayılarının dağılımı Tablo 22'deki gibidir.

Tablo 22
İl/İkametgâh Değişkenine Göre Öğrenci Sayıları

İl/İkamet	Devam Eden		Terk Eden		Toplam
	Sayı	%	Sayı	%	
İstanbul	134.140	43,90	171.382	56,10	305.522
Ankara	21.954	55,70	17.486	44,30	39.440
İzmir	13.331	53,80	11.448	46,20	24.779
Bursa	10.583	53,80	9074	46,20	19.657
Kocaeli	9054	50,70	8796	49,30	17.850
Antalya	6882	52,50	6237	47,50	13.119
Diğer İller	132.587	57,70	97.363	42,30	229.950
Toplam	328.531	50,50	321.786	49,50	650.317

Tablo 22 incelendiğinde en çok öğrencinin İstanbul ilinde yer aldığı, aynı zamanda yine İstanbul ilinde terk eden öğrenci oranının devam eden öğrenci oranına göre diğer illerin aksine

yüksek olduğu gözlenmektedir. İllere göre devam ve terk eden öğrenci oranları Şekil 26'da gösterilmiştir.



Şekil 26. İl/ikametgâh değişkenine göre öğrenci terk durumları

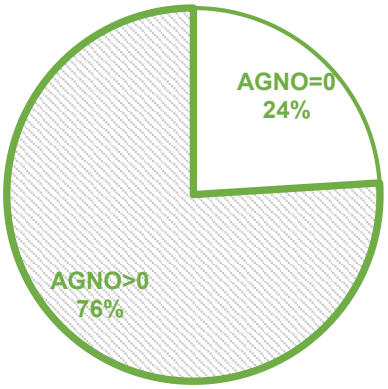
Şekil 26'ya göre İstanbul ilinde terkeden öğrenci sayısının devam eden öğrenci sayısından daha fazla olduğu; diğer illerde ise devam eden öğrenci sayılarının daha yüksek olduğu görülmektedir.

Ağırlıklı Genel Not Ortalaması (AGNO) değişkeni:

AGNO değişkeni öğrenci statü değişkeni ile birlikte araştırma sorularına göre analizlerin bir kısmında bağımlı; bir kısmında bağımsız değişken olarak ele alınmıştır. İlk aşamada sürekli veri olan AGNO değişkeni incelendiğinde ağırlıklı genel not ortalaması sıfıra eşit olan yani hiçbir akademik başarısı bulunmayan öğrencilerin oranının %24 olduğu görülmektedir (Şekil 27).

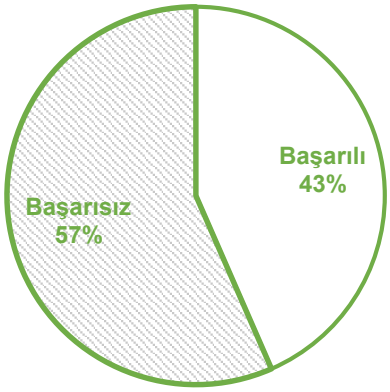
Ağırlık genel not ortalaması değişkeni analizler sürecinde iki farklı şekilde (tahmin edici-bağımsız değişken ve hedef-bağımlı değişken) kullanıldığı içi sürekli AGNO değişkeni beş kategori olarak (0-1-2-3-4) yeniden kodlanıp AGNO1 değişkeni elde edilmiştir. Ayrıca 0 ile 4 aralığında sürekli değer alan AGNO değişkeninde 2 ve üstündeki değerler başarıyı, altındaki

değerler ise başarısızlığı ifade ettiğinden veri tekrar iki kategorili (başarılı-başarısız) olacak şekilde kodlanmış ve böylelikle AGNO2 değişkeni elde edilmiştir (Şekil 28).



Şekil 27.Tüm veri setinde AGNO dağılımı

Şekil 27'de görüldüğü gibi en az bir dersten başarılı (AGNO> 0) olan öğrencilerin oranı ise %76'dır. AGNO'su 2'nin altında olanlar başarısız olarak kodlandığında AGNO değişkeninin 2 kategorideki dağılımı Şekil 28'de gösterilmiştir.



Şekil 28. AGNO2 değişkeninde başarılı ve başarısız öğrencilerin dağılımı

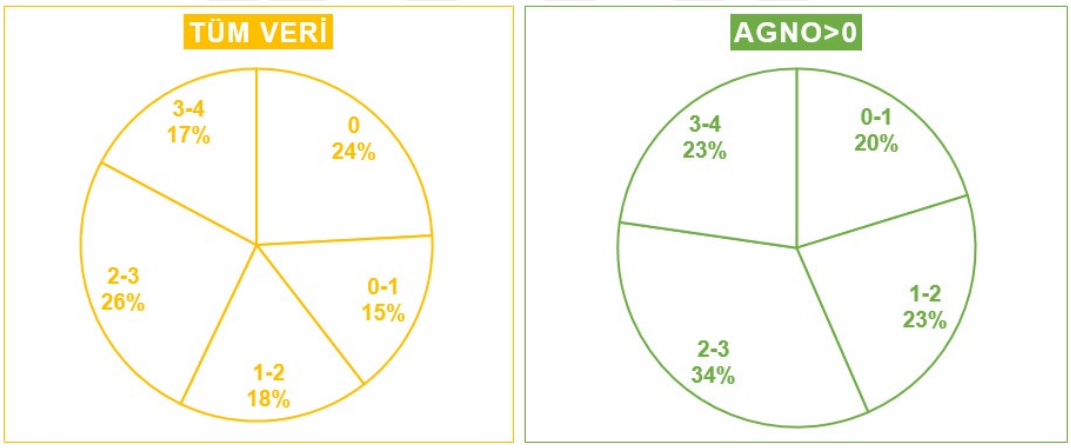
Ağırlıklı genel not ortalaması sıfır olan (yani hiçbir derste başarılı olmayan) öğrenciler tüm veri setinden ayrıldığında (yani AGNO>0 veri setinde) diğer öğrencilerin %56,6'sının başarılı kategorisinde yer aldığı görülmektedir.

Tüm veri seti içinde AGNO 1 değişkeninin tüm kategorilerdeki öğrenci terk oranları incelenmiş ve Tablo 23'teki değerler elde edilmiştir.

Tablo 23
Tüm Veri İçinde AGNO1 Değişkeni Kategorilerindeki Öğrenci Sayıları

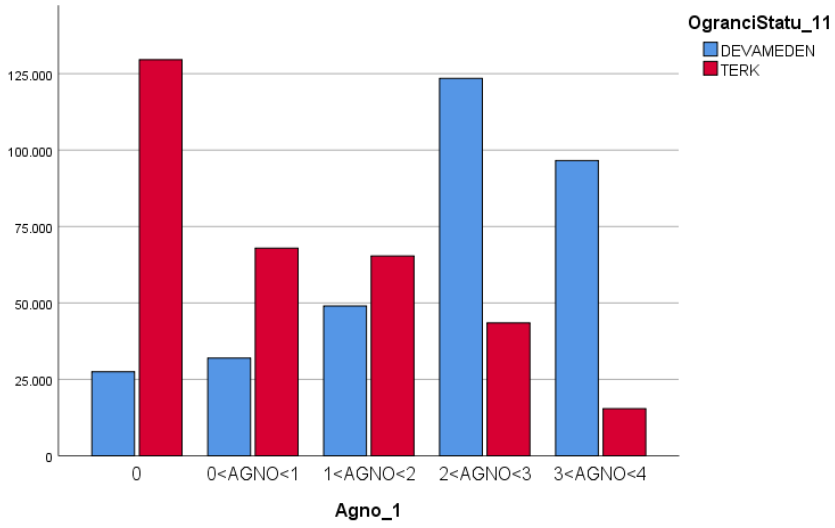
AGNO 1 değişkeninde kategoriler	Devam Eden		Terk Eden		TOPLAM
	Sayı	%	Sayı	%	
AGNO=0	27.512	17,50	129.566	82,50	157.078
0<AGNO<1	31.956	32,00	67.932	68,00	99.888
1≤AGNO<2	49.033	42,90	65.383	57,10	114.416
2≤AGNO≤3	123.444	73,90	43.486	26,10	166.930
3<AGNO≤4	96.586	86,20	15.419	13,80	112.005
TOPLAM	328.531	50,50	321.786	49,50	650.317

Tüm veri seti ve AGNO>0 veri seti içinde yer alan 1-2-3-4 kategorilerindeki ağırlık genel not ortalaması değerinin dağılımı Şekil 29'daki gibidir.



Şekil 29. Tüm veri ve AGNO>0 verisinde not kategorilerinin dağılımı

AGNO1 veri setinde kategorilerde öğrenci terk oranları AGNO>0 veri seti, tüm veri seti içinden ağırlık not ortalaması "0" olanlar çıkartılarak elde edildiği için diğer kategorilerdeki (1-2-3-4) öğrenci devam ve terk oranları AGNO>0 veri setindeki şekliyle korunmuştur. Şekil 30'da AGNO1 değişkeni not kategorilerine göre öğrenci terk oranları gösterilmektedir.

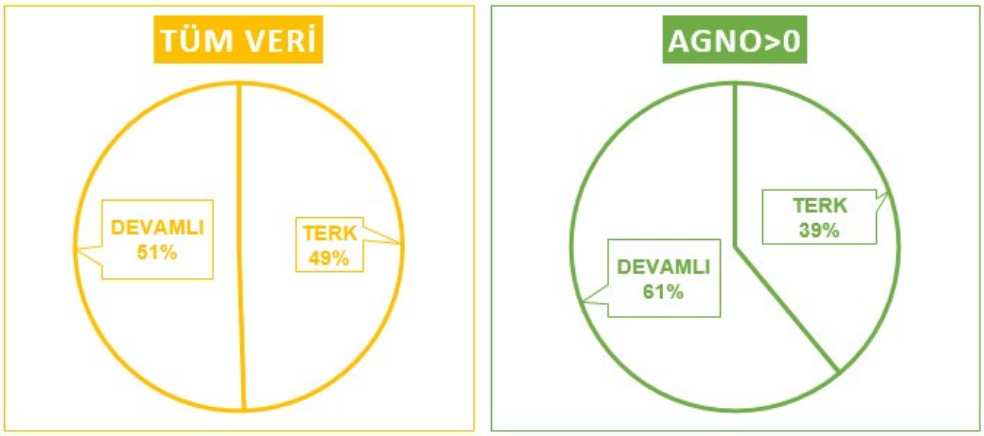


Şekil 30. Ağırlıklı genel not ortalaması değişkenine göre öğrenci terk oranları

Şekil 30 incelendiğinde ağırlıklı genel not ortalaması sıfır olan öğrencilerin büyük oranda (%82,5'inin) terk durumunda olduğu; oran azalsa da ağırlıklı not ortalaması 2'nin altında olan öğrencilerde de terk oranlarının devam oranına göre yüksek olduğu görülmektedir.

Öğrenci statü (terk-devam) değişkeni:

Ayrıca tüm veri ve AGNO>0 veri setinde öğrenci statü değişkenin dağılımı Şekil 31'de verilmiştir. Tüm veri içinde devam eden öğrenci sayısının oranı %51, terk eden öğrenci sayısının oranı ise %49 iken ağırlıklı not ortalaması sıfırdan büyük olan öğrenciler arasında devam eden öğrenci sayısı oranı %61, terk eden öğrenci sayısı oranı %39'dur.



Şekil 31. Öğrenci statü değişkeninin dağılımı; devam ve terk oranları

4.2. Yükseköğretimde Açık ve Uzaktan Öğrenmedeki Öğrencilerin Okul Terkinin Tahmininde Veri Madenciliği Sınıflandırma Algoritmalarının Performansları Nasıl Değişmektedir?

Araştırmanın aşağıda hatırlatılan birinci problemi ve ilişkili alt problemlerine yönelik elde edilen bulgular bu başlık altında sunulmuştur.

Çalışma kapsamında yer alan alt araştırma sorularından 1-i maddesine yönelik karşılaştırılan veri madenciliği modellerinin performansları ve geçerliği eğitim ve test verileri üzerinden Doğruluk, Duyarlılık, Hassasiyet, F1 puanı, Kesinlik, Negatif Öngörü Değeri, MCC ve AUC göstergeleri ile değerlendirilmiştir. İki ayrı veri seti için yürütülen araştırmada tüm veri seti için elde edilen sonuçları eğitim ve test verilerinde ayrı ayrı Tablo 24'teki gibi raporlanmıştır.

Tablo 24
Veri Madenciliği Sınıflandırma Algoritmalarının Okul Terki Tahmininde Tüm Veriye Ait Sınıflandırma Performansları

Algoritmalar	Veri	Doğruluk	Duyarlılık	Hassasiyet	F1 puanı	Kesinlik	NÖD	MCC	AUC
RA	Eğitim	0,784	0,781	0,787	0,784	0,783	0,786	0,568	0,802
	Test	0,783	0,780	0,786	0,783	0,780	0,786	0,566	0,800
CHAID	Eğitim	0,780	0,775	0,784	0,780	0,779	0,780	0,559	0,858
	Test	0,779	0,774	0,784	0,779	0,777	0,781	0,558	0,858
C5.0	Eğitim	0,795	0,800	0,790	0,795	0,789	0,801	0,589	0,854
	Test	**0,790**	**0,795**	0,786	**0,790**	**0,784**	**0,797**	**0,581**	0,853
C&RT	Eğitim	0,778	0,766	0,790	0,778	0,782	0,775	0,556	0,803
	Test	0,778	0,765	**0,790**	0,777	0,781	0,775	0,555	0,803
QUEST	Eğitim	0,754	0,752	0,756	0,754	0,752	0,757	0,508	0,784
	Test	0,754	0,751	0,757	0,754	0,751	0,757	0,507	0,784
BAYES	Eğitim	0,775	0,775	0,776	0,775	0,772	0,779	0,551	0,854
	Test	0,775	0,776	0,775	0,775	0,771	0,780	0,551	0,854
LR	Eğitim	0,777	0,767	0,788	0,777	0,780	0,775	0,555	0,855
	Test	0,776	0,766	0,787	0,776	0,778	0,775	0,553	0,855
YSA	Eğitim	0,788	0,793	0,783	0,788	0,782	0,794	0,576	0,867
	Test	0,788	0,793	0,784	0,788	0,781	0,795	0,577	**0,867**

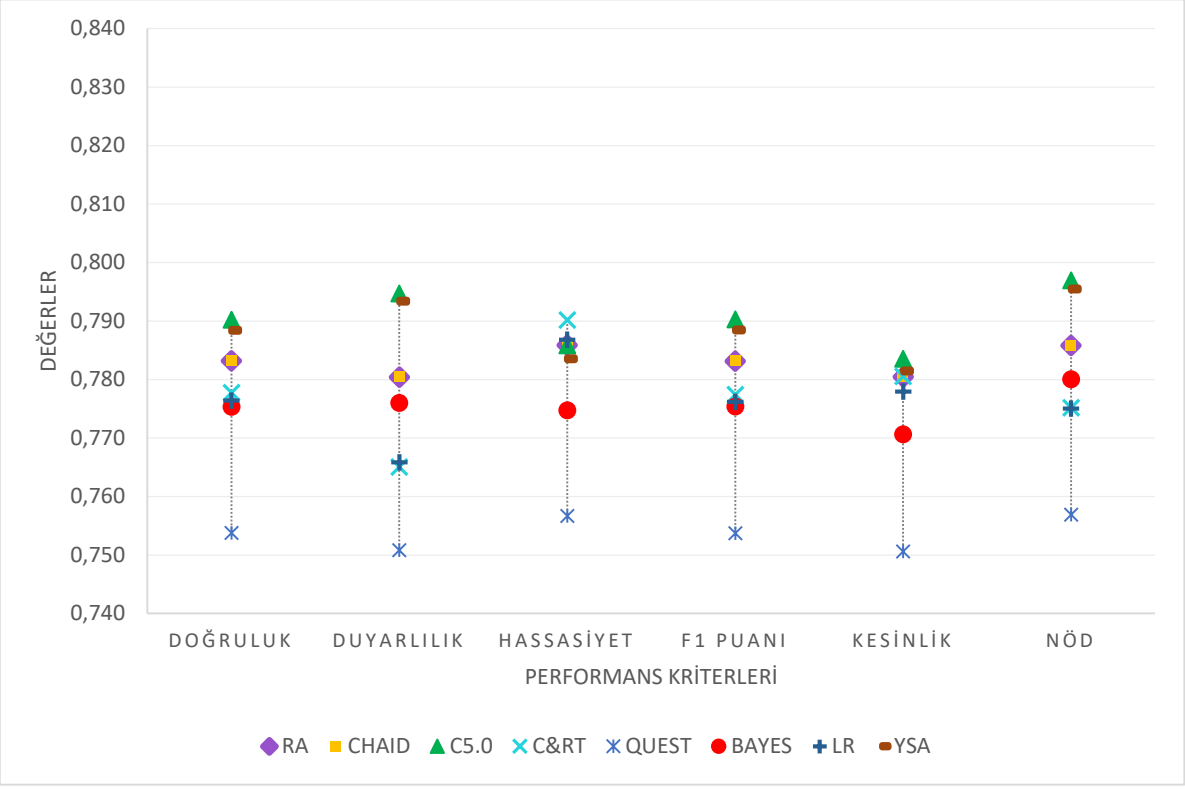
Tablo 24'e göre modellerin eğitim verisinde doğruluk değerleri 0,754 ile 0,795 arasında değişmektedir. Duyarlılık değerleri 0,751 ile 0,795 arasında; Hassasiyet değerleri 0,756 ile 0,790 arasında; F1 puanı 0,754 ile 0,795 arasında; Kesinlik değerleri 0,752 ile 0,789 arasında; Negatif öngörü değerleri 0,757 ile 0,801 arasında; MCC 0,508 ile 0,589 aralığında ve AUC değeri 0,784 ile 0,867 arasında değişmektedir.

Öte yandan veri madenciliği modellerinin esas sınıflandırma performansını incelediğimiz test verisi sonucunda doğruluk değerleri 0,754 ile 0,790 arasında değişmektedir. Performansı en

düşük model dahi %75 genel doğru sınıflandırma oranına erişmiştir. Benzer şekilde Duyarlılık değerleri 0,752 ile 0,80 arasında; Hassasiyet değerleri 0,757 ile 0,790 arasında; F1 puanı 0,754 ile 0,790 arasında; Kesinlik değerleri 0,751 ile 0,784 arasında; Negatif öngürü değerleri 0,757 ile 0,797 arasında; Matthews korelasyon katsayısı 0,507 ile 0,581 aralığında ve AUC değeri 0,784 ile 0,867 arasında gözlenmiştir.

Tablo 24 incelendiğinde tüm veri seti için eğitilen RA, CHAID, C5.0, C&RT, QUEST, BAYES, LR ve YSA algoritmalarında test sonucunda en yüksek doğruluk oranı 0,79 ile C5.0'a; daha sonra 0,788 ile YSA'ya aittir. En yüksek duyarlılık oranı 0,795 ile C5.0'a ve sonra da 0,793 ile YSA'na aittir. Diğer performans değişkenleri incelendiğinde en yüksek hassasiyet oranı 0,79 ile C&RT; en yüksek F1 puanı 0,79 ile C5.0; en yüksek kesinlik değeri 0,784 ile C5.0; en yüksek NÖD 0,797 ile C5.0; en yüksek MCC değeri 0,581 ile C5.0 ve en yüksek AUC değeri 0,867 ile YSA algoritmalarından elde edilmiştir.

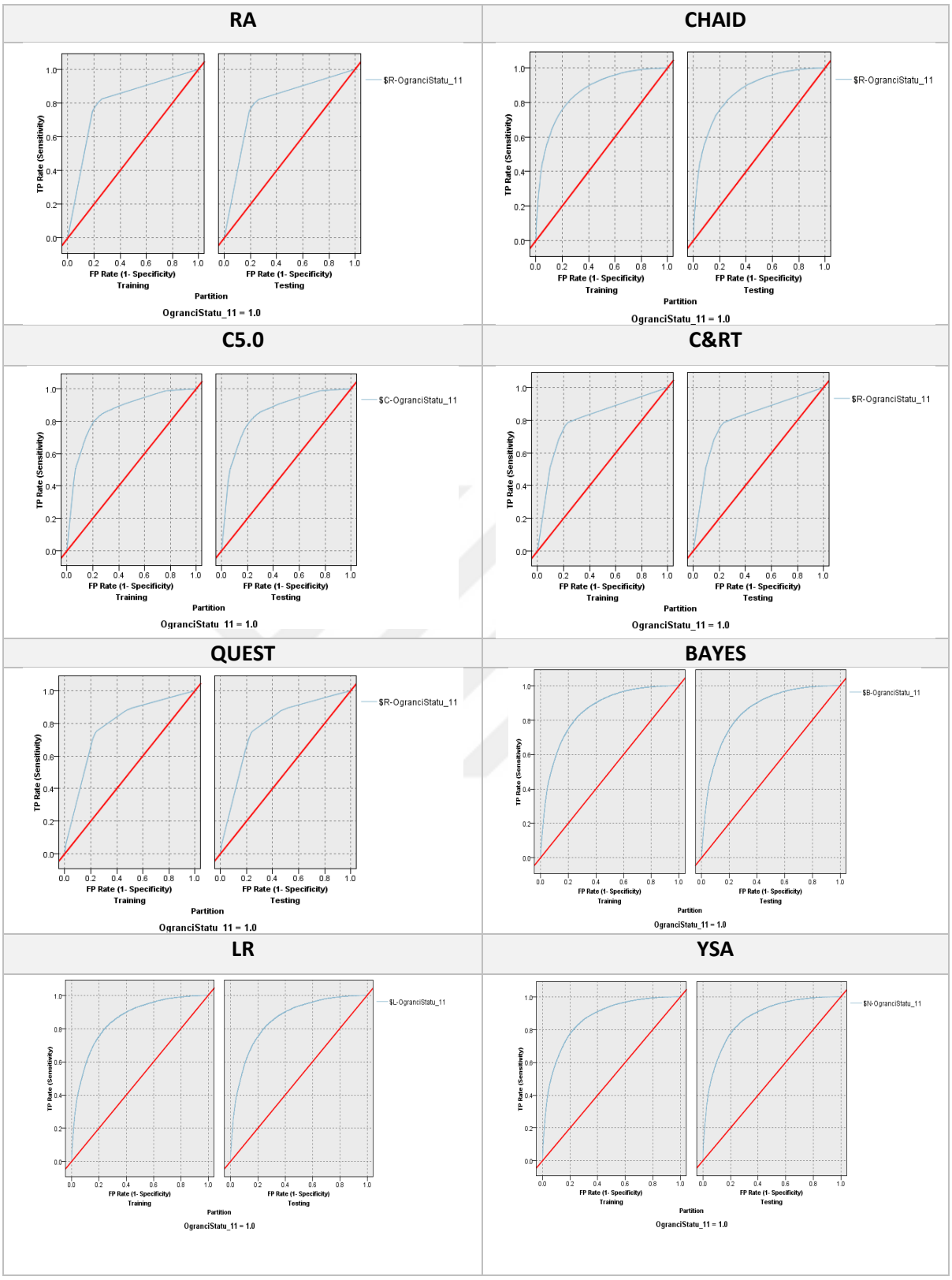
Modellerin sınıflandırma performanslarını karşılaştırabilmek için performans kriterlerine göre her bir modelin değişiminin incelenebileceği grafik (Şekil 32) oluşturulmuştur.



Şekil 32. Sınıflandırma algoritmalarının tüm veri seti için okul terki sınıflandırma performansının karşılaştırılması

Şekil 32’deki grafik incelendiğinde her bir modelin Doğruluk, Duyarlılık, Hassasiyet, F1 puanı, Kesinlik ve NÖD değerlerine göre tepe noktalarının C5.0, YSA ve C&RT algoritmalarında olduğu görülmektedir. Ayrıca QUEST modelinin tüm değerlendirme kriterlerine göre en düşük performansı gösterdiği de belirlenmiştir.

Veri madenciliğinde algoritmaların karşılaştırılmasında yaygın olarak kullanılan değerlerden biri de ROC eğrisidir. Araştırmada yer alan algoritmaların ROC eğrileri Şekil 33’te verilmiştir.



Şekil 33. Tüm veri seti için VM modellerinin ROC eğrileri

Tüm veri seti için modellerin hem eğitim hem de test sonucunda oluşan ROC eğrileri grafiklerle tabloda yer almaktadır. Öncelikle modellerde aşırı öğrenmenin (overfitting) olmadığı gözlenmektedir. Bunun yanında QUEST, RA ve C&RT'nin sınıfları ayırmada diğer modellere nispeten başarısız olduğu görülmektedir.

Tüm veri seti için geliştirilen veri madenciliği sınıflandırma modellerindeki genel doğruluk oranları ve diğer göstergeler iyi ve kabul edilebilir sonuçlar üretse de özellikle tanımlayıcı istatistiklerin incelenmesiyle öğrenci başarı göstergesindeki dengesiz dağılımların doğal olarak öğrenci terkinin tahmininde birincil etken olabileceği düşünülmektedir. Bu sebeple ağırlıklı genel not ortalamasına göre başarı durumları belirlenmiş öğrencilerden özellikle "0" kategorisinde olan öğrencilerin %82,5'inin okulu terk ettiklerinin görülmesi ve bunun bir beklenen durum olması, veri madenciliği algoritmaları ile analizlerin bunun dışındaki öğrenciler için kurulan modeldeki başarısını araştırmayı gerektirmiştir.

Çalışma kapsamında yer alan alt araştırma sorularından 1-ii maddesine yönelik AGNO>0 veri seti için oluşturulan veri madenciliği modellerinde karşılaştırılan algoritmaların performansları eğitim ve test verileri üzerinden Tablo 25'te gösterilmiştir.

Tablo 25.
Terk Tahmini: AGNO>0 için Sınıflandırma Algoritmalarının Okul Terki Sınıflandırma Performansları

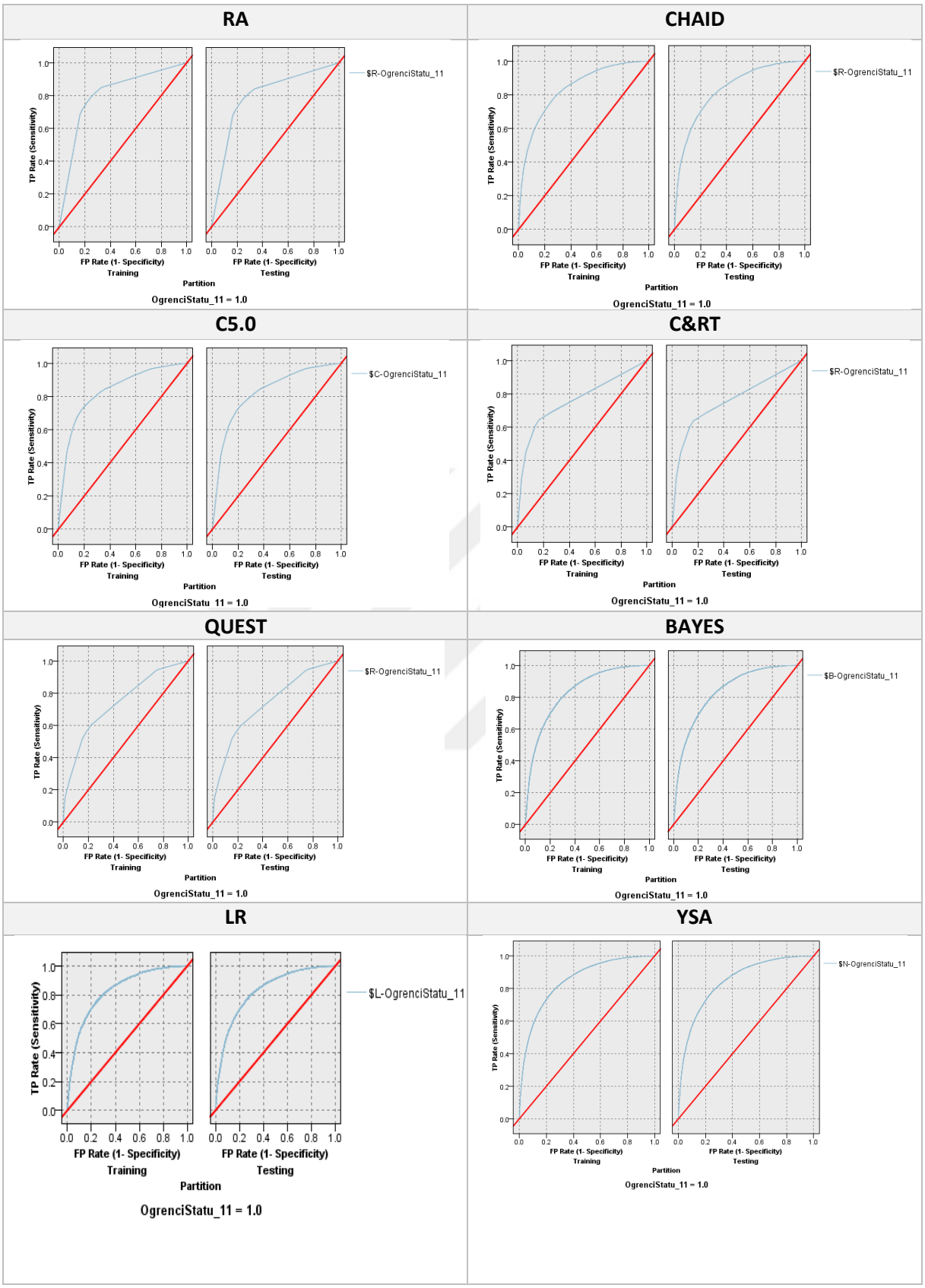
Algoritma	Veri	Doğruluk	Duyarlılık	Hassasiyet	F1 puanı	Kesinlik	NÖD	MCC	AUC
RA	Eğitim	0,771	0,775	0,768	0,772	0,681	0,842	0,533	0,806
	Test	0,766	**0,767**	0,766	**0,767**	0,677	**0,838**	0,524	0,799
CHAID	Eğitim	0,769	0,602	0,876	0,713	0,756	0,775	0,503	0,834
	Test	0,765	0,594	**0,874**	0,707	**0,750**	0,771	0,494	0,832
C5.0	Eğitim	0,785	0,679	0,852	0,756	0,746	0,806	0,542	0,831
	Test	**0,777**	0,668	0,847	0,747	0,736	0,800	**0,525**	0,826
C&RT	Eğitim	0,765	0,605	0,867	0,713	0,745	0,775	0,495	0,763
	Test	0,763	0,600	0,867	0,709	0,742	0,773	0,490	0,759
QUEST	Eğitim	0,719	0,516	0,849	0,642	0,686	0,733	0,391	0,740
	Test	0,719	0,513	0,850	0,640	0,685	0,732	0,389	0,738
BAYES	Eğitim	0,763	0,643	0,840	0,729	0,720	0,787	0,495	0,834
	Test	0,760	0,636	0,839	0,724	0,715	0,783	0,487	0,831
LR	Eğitim	0,765	0,634	0,849	0,726	0,729	0,784	0,498	0,831
	Test	0,763	0,630	0,848	0,723	0,726	0,782	0,493	0,829
YSA	Eğitim	0,779	0,665	0,852	0,747	0,741	0,799	0,528	0,849
	Test	0,776	0,658	0,851	0,742	0,738	0,796	0,521	**0,847**
DVM	Eğitim	0,784	0,667	0,860	0,751	0,752	0,801	0,540	0,822
	Test	0,776	0,656	0,855	0,743	0,743	0,796	0,525	0,799

Tablo 25 incelendiğinde AGNO>0 veri seti için eğitilen RA, CHAID, C5.0, C&RT, QUEST, BAYES, LR, YSA, DVM algoritmalarında model eğitim ve test verilerinde sonuçların birbirine yakın olduğu görülmektedir. Ayrıca tüm modellerde genel sınıflandırma doğruluğu da birbirine oldukça yakındır ve 0,719 ile 0,777 arasında değişmektedir.

Modellerde genel sınıflandırma için en yüksek doğruluk oranı 0,777 ile C5.0'a; daha sonra 0,776 ile YSA'na aittir. En yüksek duyarlılık oranı 0,767 ile RA'a ve sonra da 0,668 ile C5.0'a aittir. Tabloda yer alan diğer diğer performans değişkenlerine göre en yüksek hassasiyet oranı 0,874 ile CHAID; en yüksek F1 puanı 0,767 ile RA; en yüksek kesinlik değeri 0,750 ile CHAID; en yüksek negatif öngörü değeri 0,838 ile RA; en yüksek MCC 0,525 ile C5.0 ve en yüksek AUC değeri 0,847 ile YSA algoritmalarından elde edilmiştir.

Şekil 34'te AGNO>0 veri seti için oluşturulan sınıflandırma modellerinin ROC grafikleri yer almaktadır. Model performansları için grafikler incelendiğinde QUEST ve C&RT algoritmalarında eğri altında kalan alanın diğerlerine göre daha az olduğu; dolayısı ile bu modellerin sınıfları ayırmada nispeten başarısız olduğu söylenebilir. Tersine LR, YSA ve Bayes sınıflandırmalarında eğri belirgin şekilde orta çizgiden ayrıldığı için modeller sınıfları ayırmada başarılı olarak görülebilir.

Ayrıca grafikler hem eğitim hem test verilerine dair sınıflandırma başarısını göstermektedir. Buradan modellerde aşırı öğrenme durumunun oluşmadığı görülmektedir ki bu da model geçerliği için önemli kanıtlardandır.



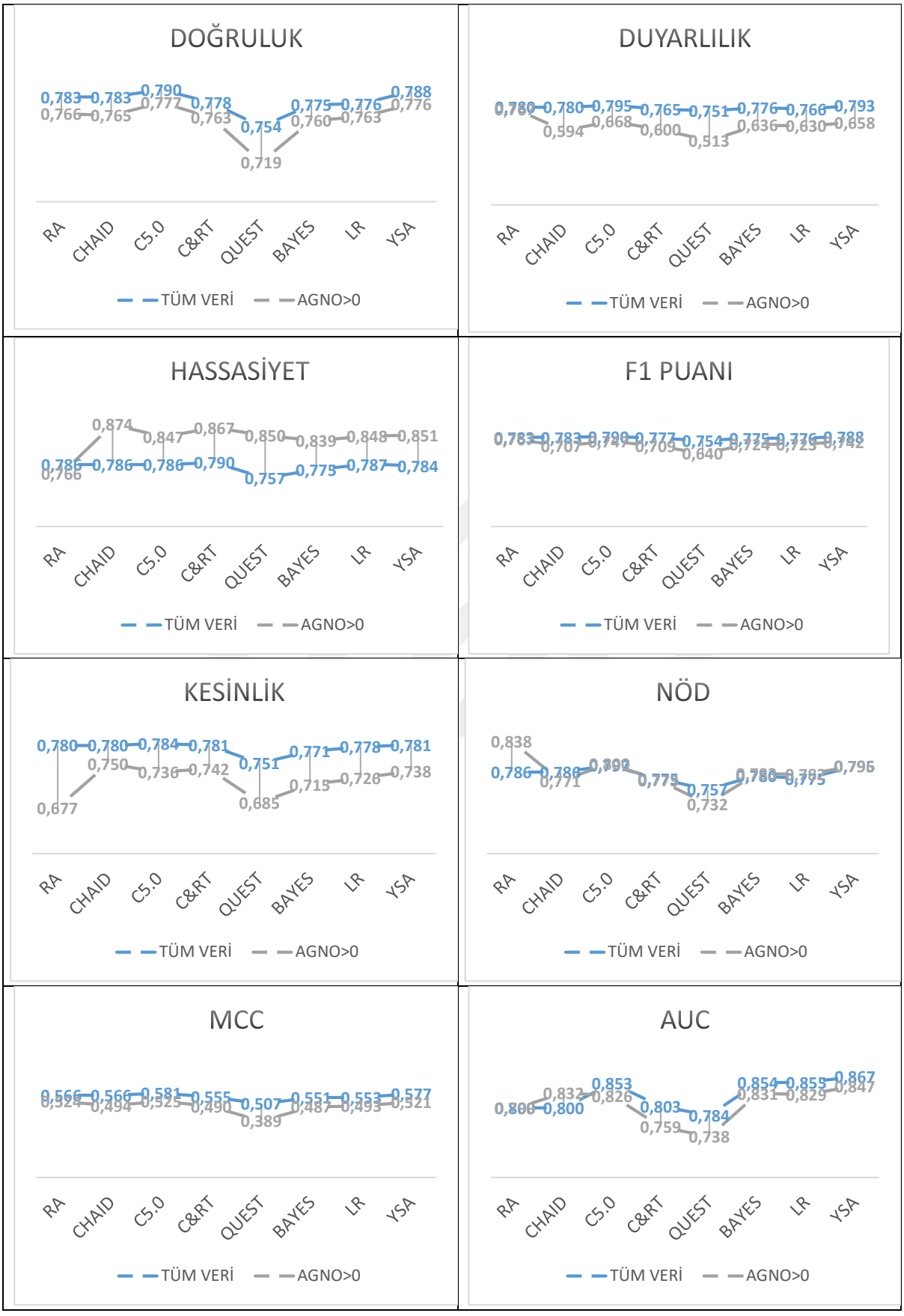
Şekil 34. Terk tahmini; AGNO>0 veri seti için sınıflandırma modellerinin ROC eğrileri

Çalışmanın bu bölümünde yukarıda bulguları yer alan, tüm veri setinde ve ağırlıklı genel not ortalaması sıfırdan büyük olan öğrencilerin yer aldığı veri setinde (AGNO>0) yapılan iki ayrı analizin durumu karşılaştırılmıştır. Dolayısı ile veri madenciliği moddellerinin veri setlerine göre performansları Tablo 26'da özetlenmiş değerlere göre incelenebilir.

Tablo 26
Tüm Veri ve AGNO>0 Verisi ile Yapılan Analizlerde Model Performanslarının Karşılaştırılması

	RA		CHAID		C5.0		C&RT		QUEST		BAYES		LR		YSA	
Performans Kriterleri	TÜM VERİ	AGNO>0	TÜM VERİ	AGNO>0	TÜM VERİ	AGNO>0	TÜM VERİ	AGNO>0	TÜM VERİ	AGNO>0	TÜM VERİ	AGNO>0	TÜM VERİ	AGNO>0	TÜM VERİ	AGNO>0
Doğruluk	0,783	0,766	0,783	0,765	0,790	0,777	0,778	0,763	0,754	0,719	0,775	0,760	0,776	0,763	0,788	0,776
Duyarlılık	0,780	0,767	0,780	0,594	0,795	0,668	0,765	0,600	0,751	0,513	0,776	0,636	0,766	0,630	0,793	0,658
Hassasiyet	0,786	0,766	0,786	0,874	0,786	0,847	0,790	0,867	0,757	0,850	0,775	0,839	0,787	0,848	0,784	0,851
F1 puanı	0,783	0,767	0,783	0,707	0,790	0,747	0,777	0,709	0,754	0,640	0,775	0,724	0,776	0,723	0,788	0,742
Kesinlik	0,780	0,677	0,780	0,750	0,784	0,736	0,781	0,742	0,751	0,685	0,771	0,715	0,778	0,726	0,781	0,738
NÖD	0,786	0,838	0,786	0,771	0,797	0,800	0,775	0,773	0,757	0,732	0,780	0,783	0,775	0,782	0,795	0,796
MCC	0,566	0,524	0,566	0,494	0,581	0,525	0,555	0,490	0,507	0,389	0,551	0,487	0,553	0,493	0,577	0,521
AUC	0,800	0,799	0,800	0,832	0,853	0,826	0,803	0,759	0,784	0,738	0,854	0,831	0,855	0,829	0,867	0,847

Araştırmanın ilk aşamasında tüm veri seti için yapılan analizler ve ikinci aşamda AGNO>0 veri seti için yapılan analizler sunulmuştur. Okul terkinin tahimininde ağırlıklı genel not ortalaması "0" olan ve dolayısı ile kayıt sonrası hiçbir eğitim öğretim faaliyetinde başarılı olmayan öğrencilerin %82,5'inin okul terkettiği bilindiğinden bu grubun veriden çıkartılarak yapılacak analizlerden beklenen sonucun alınıp alınmadığı AGNO>0 veri seti için yeniden incelenmiş ve Tablo 26'da gösterilmiştir. Tablo 26'da her bir veri setine ait performans kriterinin değeri büyüdükçe yeşil tonu koyulaşmakta; ortalama altı değerlerde ise renk beyaza dönmektedir. Buna göre en yüksek model doğruluk oranı 0,79 ile tüm veri seti için C5.0 algoritması ile sağlanmıştır. En yüksek model duyarlılık değeri 0,795 ile tüm veri seti için C5.0 algoritması ile; en yüksek hassasiyet 0,874 ile AGNO>0 veri seti için CHAID algortiması ile; en yüksek F1 puanı 0,79 ile tüm veri seti için C5.0 algoritması ile; en yüksek kesinlik 0,784 ile tüm veri seti için C5.0 algoritması ile; en yüksek negatif öngörü değeri 0,838 ile AGNO>0 veri seti için RA algoritması ile; en yüksek MCC değeri 0,581 ile tüm veri seti için C5.0 algoritması ile; en yüksek AUC değeri 0,867 ile tüm veri seti için YSA algoritması ile elde edilmiştir.



Şekil 35. Tüm Veri ve AGNO>0 verisi için elde edilen VM modellerinin karşılaştırmalı performans grafikleri

Şekil 35'te yer alan grafikler incelendiğinde algoritmaların bazı performans değerlerinde istisnalar olsa da genel olarak okul terki tahmininde tüm veri seti ile çalışılmasının daha yüksek doğruluk oranıyla tahmin değerleri verdiğini söyleyebiliriz. Ancak hassasiyet ve duyarlılık arasında dengeyi gözeten kriterlerden MCC ve F1 puanı her iki veri seti için de algoritmalarda yakın değerler almıştır.

Tüm veri için yapılan analizlere göre AGNO>0 veri setinde yapılan sınıflandırma öngörüsünde modellerin hassasiyet değerlerinin oldukça yükseldiği görülmektedir. Hassasiyet, modelin TN sınıflandırması (doğru negatifler; yani devam eden öğrenci) için bir başarı göstergesi olduğundan AGNO>0 veri setinde devam eden öğrenci öngörüsü için modellerin daha başarılı sonuçlar ürettiği söylenebilir.

Model performanslarının her iki veri seti için bazı kriterlere göre birbirinden ayrıldığı görülmüştür. Buna göre tüm veri setinde genel sınıflandırma doğruluğu en yüksek model C5.0'dır. Ayrıca C5.0 modeli duyarlılık, F1 puanı, kesinlik, negatif öngörü değeri ve MCC'de en yüksek değere sahiptir. En yüksek hassasiyet C&RT modeli; en yüksek AUC değeri de YSA modeli ile elde edildiği için bu üç model aracılığı ile okul terkinin belirlenmesinde etkili olan faktörler ele alınacaktır.

AGNO>0 veri setinde ise C5.0 genel sınıflandırma doğruluğu ve MCC değeri en başarılı model iken; Duyarlılık, F1puanı, negatif öngörü değeri RA modeli; hassasiyet ve kesinlik CHAID modeli ve AUC değeri YSA modeli ile en yüksek değeri almıştır. Buna bağlı olarak da ağırlıklı genel not ortalaması sıfırdan farklı olan (AGNO>0) öğrencilerin okul terkini belirleyen nitelikler bu dört model aracılığı ile incelenecektir.

4.3. Yükseköğretimde Açık ve Uzaktan Öğrenmedeki Öğrencilerin Okul Terki ile İlişkili Demografik Faktörler (Yaş, Cinsiyet, Medeni Durum, İkametgâh İli vb.) ve Diğer Nitelikler Nelerdir?

Araştırmanın ikinci problemi ve ilişkili alt problemlerine yönelik elde edilen bulgular aşağıdaki başlıklar altında sunulmuştur.

i. Tüm değişkenlerin yer aldığı veri setinde veri madenciliği ile öğrencilerin okul terkinin tahmininde hangi faktörler rol oynamaktadır?
ii. AGNO>0 veri setinde yer alan öğrencilerin veri madenciliği ile okul terkinin tahmininde hangi faktörler rol oynamaktadır?

4.3.1. Tüm Değişkenlerin Yer Aldığı Veri Setinde Veri Madenciliği ile Öğrencilerin Okul Terkinin Tahmininde Hangi Faktörler Rol Oynamaktadır?

Tüm veri seti ile yapılan okul terki öngörüsünde performans değerleri karşılaştırıldığında C5.0, C&RT ve YSA algortimalarının diğerlerine göre daha başarılı sonuçlar ürettiği görülmüştür. Buna göre geliştirilen bu modellerde öğrencilern okul terki ile ilişkili değişkenler aşağıda ayrıntılı olarak ele alınmıştır.

4.3.1.1. C5.0

C5.0 algoritması karar ağacı ile yapılan sınıflandırma çalışmalarında sıkça kullanılan ve oldukça başarılı sonuçlar veren algoritmalardandır. Bu çalışmada terk eden ve devam eden öğrencilerin tahmininde 0,790 doğruluk, 0,795 duyarlılık, 0,786 hassasiyet, 0,790 F1 puanı, 0,784 Kesinlik, 0,797 NÖD, 0,581 MCC ve 0,853 AUC değeri ile en başarılı model olarak görülmüştür. Modelde yer alan değişkenlerin sınıflandırma üzerinde önemlilik dereceleri Tablo 27'deki gibidir.

Tablo 27
Tüm Veri Seti İçin C5.0 Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri

Faktör/Nitelik	Etki katsayısı
AGNO1	0,8468
Öğrenci Giriş Yılı	0,0985
Yaş Kategorisi	0,0373
Program Türü	0,0076
İkametgâh İli	0,0053
Program Tipi	0,0043
Cinsiyet	0,0003
Program Düzeyi	0
Geliş/Kayıt Sekli	0

Tablo 27'de yer alan değişkenler incelendiğinde C5.0 karar ağacı ile kurulan modelde okulu terk eden öğrencilerin belirlenmesinde en önemli etken AGNO yani 5 kategorili ağırlıklı not ortalaması değişkenidir. Daha sonra sırası ile öğrenci giriş yılı, yaş, program, ikematgah ili, program tipi, cinsiyet değişkenleri etkili değişkenler olarak bulunmuştur. Program düzeyi ve geliş şekli değişkenlerinin ise C5.0 sınıflandırma algoritmasında öğrencilerin terk veya devam durumlarının belirlenmesinde bir etkisi gözlenmemiştir. Ayrıca AGNO (ağırlıklı genel not ortalaması) değişkeninin katsayısının terkin tahmininde diğer etkili olan öğrenci giriş yılı, yaş,

program, ikametgâh, öğrenim tipi, cinsiyet, kayıt/geliş şekli ve program düzeyi değişkenlerine göre oldukça yüksek olduğu görülmektedir.

C5.0 algoritmasında doğruluk ve hassasiyet oranları yüksek bir performans gösterse de modelde 1093 düğüm oluşmuştur. Karar ağacında ilk düğüm AGNO değerine göre oluşmaktadır, devamında ise alt düğümler sırasıyla, öğrenci giriş yılı ve yaş değişkenlerine göre dallanmaktadır. C5.0 modelinin düğümlerinin fazla; ağaç yapısının çok geniş olması modelin anlaşılmasını zorlaştırabilmektedir. Ek 1'de tüm veri için C5.0 algoritmasının ilk dört düzeyi şema olarak gösterilmiştir. İlgili ağaçta da okul terkinin öngörülmesinde etkili olan ilk değişkenin AGNO olduğu; buna göre AGNO değeri "0 ve 1" olanlarda okul terkinin (%70,7) "3 ve "4 olanlara (%21) göre daha yüksek olduğu grafiksel gösterimden de anlaşılmaktadır.

C5.0 ağacında ilk düzeyde ağırlıkı not ortalamasına göre dallar oluşmakta ve öğrencilerin yaklaşık %57'sinin AGNO değeri "0 ve 1" ; %43'ünün AGNO değeri "3 ve 4" olarak sınıflandırılmaktadır. Ayrıca ağacın ilk 4 düzeyine ait model ve değişkenlere göre aşağıdaki kurallar oluşmuştur.

Burada modeldeki ilk düzey düğümde etkili değişken olan AGNO değeri "0-1-2" olanlardan;

- Giriş yılı "2010-2011-2012-2013-2014-2015-2016-2017-2018-2019" olup AGNO değeri "0 ve 1" olanlar (98.244) %89 oranla okulu terk etmekte; AGNO değeri "2" olanlardan giriş yılı "2010-2011-2012-2013-2014-2015-2016-2017" olan %84 oranla okulu terk etmektedir.
- Giriş yılı "2020-2021" olanlardan AGNO değeri "0" giriş yılı"2020" olanlar %91 oranla okulu terk etmekte; AGNO değeri "1-2" giriş yılı"2021" olanlar ise %76 oranla okula devam etmektedir.

Ağacın diğer bir dalına göre ise AGNO değeri "3-4" olanlardan;

- Giriş yılı "2021" olanlar %95 oranla okula devam etmektedir. Giriş yılı "2010-2011-2012-2013-2014-2015-2016-2017-2018-2019-2020"olanların sınıflandırılmasında en etkili değişkenin yaş faktörü olduğu; buna göre yaş kategorisi "1-2" olanlar ile "2-4-5" olanların sınıflarında doğruluk oranlarının değiştiği ve sonraki en etkin değişken olan program faktörüne göre yüzlerce ağaç oluştuğu gözlenmektedir.

4.3.1.2. C&RT

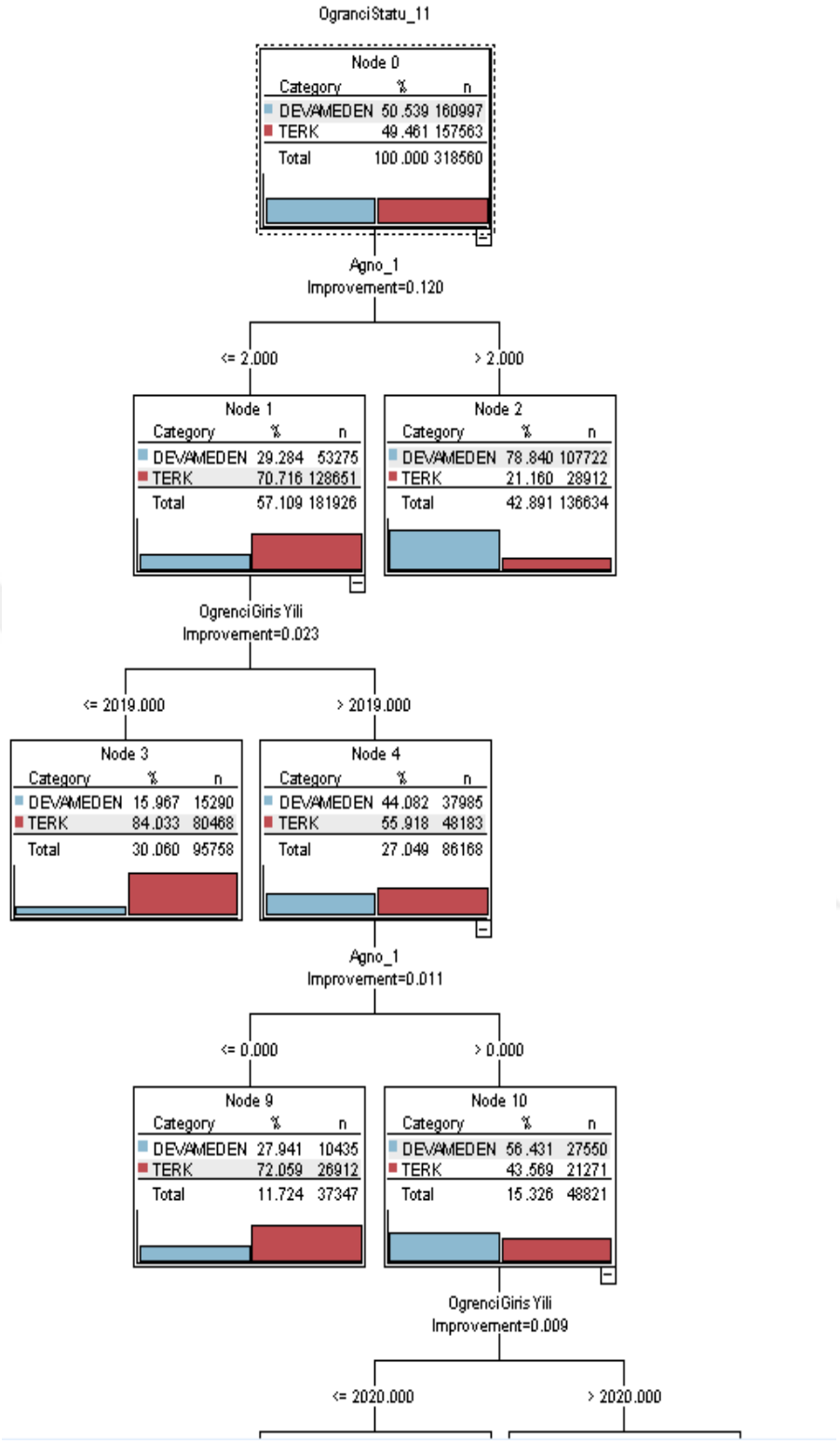
C&RT modeli regresyon ve karar ağacı algoritmasını birlikte uyguladığı için çok yaygın kullanıma sahiptir. Bu araştırmada terk eden ve devam eden öğrencilerin tahmininde 0,778 doğruluk, 0,765 duyarlılık, 0,790 hassasiyet, 0,777 F1 puanı, 0,781 Kesinlik, 0,775 NÖD, 0,555 MCC ve 0,803 AUC değerleri ile yüksek performans değerlerine sahip modellerden biridir. Modelde yer alan değişkenler ve önemlilik dereceleri Tablo 28'deki gibidir.

Bulgular incelendiğinde modelde 8 değişkenin yer aldığı ve C&RT Modelinin C5.0'ın aksine 38 düğüm ve 4 düzeyde bir karar ağacından oluştuğu görülmektedir.

Tablo 28
Tüm Veri Seti İçin C&RT Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri

Faktör/Nitelik	Etki katsayısı
AGNO1	0,8726
Öğrenci Giriş Yılı	0,0823
Program Düzeyi	0,0089
Yaş	0,0089
İkametgâh İli	0,0089
Açık-Uzaktan	0,0089
Cinsiyet	0,0089
Geliş/Kayıt Şekli	0,0006

Tablo 28 incelendiğinde okul terki ile ilişkili en önemli değişkenin C5.0 modelinde de olduğu gibi AGNO1 değişkeni olduğu; bunu ise sırasıyla öğrenci giriş yılı, cinsiyet, öğrenim tipi, ikametgâh, yaş, program düzeyi ve kayıt/geliş şekli değişkenlerinin takip ettiği görülmüştür.



Şekil 36. Tüm veri seti için C&RT modelinin ağaç yapısı

C&RT modeline ait ağaç yapısı Şekil 36'daki gibi oluşmuştur. Modelin genel sınıflandırma doğruluğu %77,8 iken %79 ile diğer modellere göre en yüksek hassasiyete sahip olduğu

bilindiğine göre modelin negatif sınıfı (devam eden öğrencileri) tahminde daha etkili olduğu söylenebilir. Ayrıca ağaç yapısı da diğer modellere göre daha sade ve anlaşılırdır.

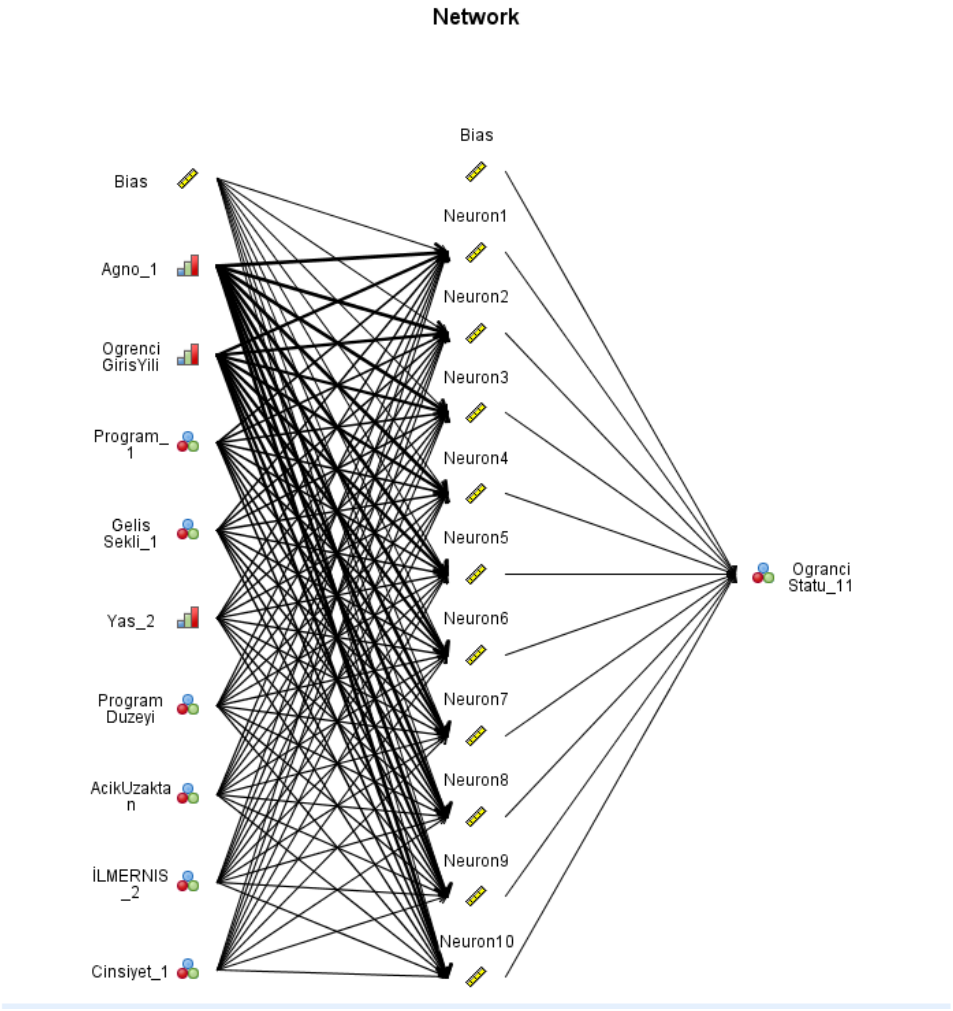
Şekil 36'da yer alan karar ağacı yapısına göre C&RT modelinde yapılan sınıflandırmada AGNO değeri "0, 1 ve 2" olanların (181.926 kişi) AGNO değeri "3 ve 4" olanlara göre daha yüksek oranda okulu terk ettiği; AGNO değeri "3-4" olanların %78 oranla okula devam ettiği görülmektedir. Ağacın sonraki düzeyleri incelendiğinde AGNO değeri "0, 1 ve 2" olanlardan;

- Giriş yılı "2010-2011-2012-2013-2014-2015-2016-2017-2018-2019" olanlar %84 oranla okulu terk etmekte;
- Giriş yılı "2020, 2021" olup AGNO değeri "0" olanlar %72 oranla okulu terk etmekte;

AGNO değeri "1-2-3-4"; giriş yılı "2010-2011-2012-2013-2014-2015-2016-2017-2018-2019-2020" ve geliş şekil "ÖZEL-SINAVSIZ İKİNCİ ÜNİVERSİTE" olanlar %62 oranla okulu terk etmekte iken geliş şekli "DGS-OSS-YOS" olanlar %54 oranla okula devam etmektedir. Öte yandan giriş yılı "2021" olanlar ise %76 oranla okula devam etmektedir.

4.3.1.3. YSA

Yapay sinir ağlarının klasik doğrusal modelleme anlayışının dışına çıkıp tüm etkenleri ve etkileşimlerini göz önüne alarak yaptığı sınıflandırma pek çok uygulamada başarılı bulunmaktadır. Bu araştırmada ise öğrencilerin terk ve devam durumlarına göre sınıflandırılmasında 0,867 AUC değeri ile diğerlerine göre en yüksek performansı göstermiştir. Ayrıca modelin genel doğruluk oranı 0,788'dir. Modelde 0,793 duyarlılık, 0,784 hassasiyet, 0,788 F1 puanı, 0,781 Kesinlik, 0,795 NÖD ve 0,577 MCC değeri gözlenmiştir. Modelde AUC değerinin 1'e yakın olması, modelin pozitif (öğrencinin terk edeceği) ve negatif (öğrencinin devam edeceği) sınıflar arasında iyi bir ayrım yapabildiğini; 0,867 gibi yüksek bir AUC değeri, modelin güvenilir tahminlerde bulunduğunu ve karar sınırının doğru bir biçimde belirlendiğini gösterir. Yapay sinir ağları modelinde giriş katmanında yer alan değişkenler ile oluşturulan nöronlar aracılığında çıkış katmanında öğrenci statüsüne bağlı sınıflandırma yapılmışıtr. Yapay sinir ağı modelindeki ağ yapısı Şekil 37'deki gibidir.



Şekil 37. Tüm veri seti için YSA modelinde sinir ağları şeması

Şekilde 37'deki gibi modelde farklı girdi değişkenlerinin çıktılar üzerindeki etkileşimleri görsel olarak sunulmuştur. Şekilde ayrıca çok sayıda nöron içeren bir gizli katman görülmektedir (Nöron1'den Nöron10'a kadar). Her bir nöron, girdi katmanından aldığı bilgileri işler ve sonraki katmana aktarır. Gizli katmandaki nöron sayısı, modelin karmaşıklığını ve öğrenme kapasitesini gösterir. Çok sayıda nöron, modelin daha karmaşık ilişkileri öğrenebileceğinin, ancak aşırı uyuma eğilimin de olabileceğinin bir göstergesidir. Değişkenlerin şekildeki sıralaması modeldeki önem dereceleri için de bir göstergedir. Buna göre en önemli etkenin diğer iki modelde de olduğu gibi AGNO1 değişkeni olduğu; bunu ise sırasıyla öğrenci giriş yılı, program, kayıt/geliş şekli, yaş, program düzeyi öğrenim tipi, ikametgâh ve cinsiyet değişkenlerinin takip ettiği görülmüştür.

Tablo 29'da YSA modelinde etkili olan değişkenlerin önem katsayıları verilmiştir.

Tablo 29
Tüm Veri Seti İçin YSA Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri

Faktör/Nitelik	Etki katsayısı
AGNO 1	0,3274
Öğrenci Giriş Yılı	0,2595
Program	0,1128
Geliş/Kayıt Şekli	0,0923
Yaş	0,0634
Program Düzeyi	0,0562
Açık-Uzaktan	0,0360
İkametgâh ili	0,0325
Cinsiyet	0,0199

Özellikle AGNO ve öğrenci giriş yılı gibi girdiler yüksek öneme sahip görünmektedir. Bu özniteliklerin model tahminlerinde belirleyici rol oynadığı anlaşılıyor. Buna bağlı olarak diğer öznitelikler daha düşük öneme sahiptir.

Tüm veri setinde okul terkine göre öğrencilerin sınıflanırılmasında en başarılı olan üç modelin tahmin için kullandığı değişkenler yukarıda açıklanmıştır. Aşağıdaki tablo ile tümü için bir karşılaştırma yapılabilir.

Tablo 30
Okul Terkine Etki Eden Faktörlerin Modellerde Karşılaştırması (Tüm Veri)

TÜM VERİ SETİ					
C5.0		C&RT		YSA	
Faktör	*Etki katsayısı*	*Faktör*	*Etki katsayısı*	*Faktör*	*Etki katsayısı*
AGNO1	0,8468	AGNO1	0,8726	AGNO1	0,3274
Öğrenci Giriş Yılı	0,0985	Öğrenci Giriş Yılı	0,0823	Öğrenci Giriş Yılı	0,2595
Yaş	0,0373	Program Düzeyi	0,0089	Program_1	0,1128
Program_1	0,0076	Yaş	0,0089	Geliş/Kayıt Şekli	0,0923
İkametgâh İli	0,0053	İkametgâh İli	0,0089	Yaş	0,0634
Açık-Uzaktan	0,0043	Açık-Uzaktan	0,0089	Program Düzeyi	0,0562
Cinsiyet	0,0003	Cinsiyet	0,0089	Açık-Uzaktan	0,036
		Geliş/Kayıt Şekli	0,0006	İkametgâh İli	0,0325
				Cinsiyet	0,0199

Tablo 30 inclendiğinde her üç modelin de okul terki için sınıflandırma yaparken ilk iki sırada ağırlı genel not ortalamasını ve öğrenci kayıt yılı faktörünü kullandığı görülmektedir. Daha sonra faktörler farklılaşmakta; C&RT, program düzeyi ve yaş değişkenini daha etkili bir faktör olarak almışken; C5.0, yaş ve program türünü; YSA ise program türü ve geliş şeklini sırasıyla önemli faktörler olarak değerlendirmişlerdir. Ayrıca her üç modelde de sınıflandırma için öğrenim tipi, ikamet ili ve cinsiyet etkili faktörler iken modele etki düzeyi en düşük

değişkenlerdendir. Bunun yanında C5.0 modeli program düzeyi ve geliş şekli faktörlerini; C&RT ise program türü faktörünü sınıflandırmada hiç kullanmamıştır.

Modellerin sınıflandırma performanslarında genel doğruluk oranlarının birbirine çok yakın olduğunu da dikkate alırsak okul terkinin belirlenmesinde en önemli faktörün ağırlıklı genel not ortalaması olduğu; öğrenci kayıt yılı, program türü ve düzeyi ile yaş faktörünün önemli diğer nitelikler olduğu görülmektedir. Bunu yanında ikamet ili de okul terkinde oldukça etkili bir faktör olarak ortaya çıkmıştır.

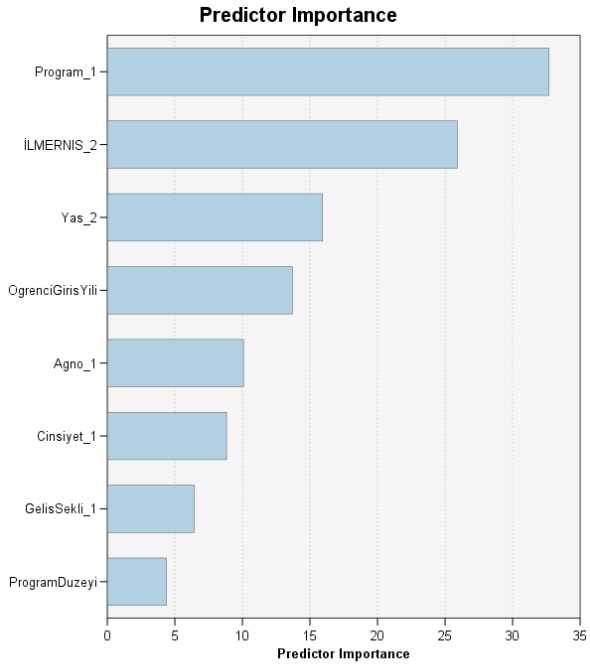
4.3.2. AGNO>0 Veri Setinde Yer Alan Öğrencilerin Veri Madenciliği ile Okul Terkinin Tahmininde Hangi Faktörler Rol Oynamaktadır?

Tüm veri seti ile yapılan analizlerin sonucunda özellikle AGNO1 değişkeninde yer alan "0" ketogorisinin tahminde etkili olduğu ancak bu değişkenin sistemde hiçbir eğitim öğretim faaliyeti göstermemeyi ortaya koyduğu ve bu kategoride olan öğrencilerin okulu terk etme oranı (%82,5) diğer başarı kategerilerine göre doğal olarak bariz şekilde farklılaştığı için analize bu grup öğrenci dahil edilmeden model başarısı araştırılmaya değer görülmüş ve AGNO> 0 veri setinde yer alan öğrenci grubunun sınıflandırılmasında veri madenciliği algortimalarının başarısı analiz edilmiştir.

AGNO>0 veri setinde okul terki öngörüsünde modellerin performans değerleri karşılaştırıldığında RA, CHAID ile yine C5.0 ve YSA algortimalarının diğerlerine göre daha başarılı sonuçlar ürettiği görülmüştür. Buna göre bu modlelerle oluşturulan sınıflarda etkili olan değişkenler aşağıda ele alınmıştır.

4.3.2.1. Rastgele Ağaçlar

Rastgele ağaçlar modeli rastgele değişkenlerin etkisini inceleyerek model oluşturan temel karar ağacı modellerindendir. RA, AGNO>0 veri seti için terk eden ve devam eden öğrencilerin tahmininde 0,766 doğruluk, 0,767 duyarlılık, 0,766 hassasiyet, 0,767 F1 puanı, 0,677 Kesinlik, 0,838 NÖD, 0,524 MCC ve 0,799 AUC değeri ile yüksek performans değerlerine sahip modellerden biridir. Modelde yer alan değişkenlerin önem dereceleri Şekil 38'teki gibidir.



Şekil 38. AGNO>0 veri seti C5.0 modelinde tahmin değişkenleri ve önemi

Şekil 38 incelendiğinde en önemli değişkenin program faktörü olduğu; bunu sırasıyla ikametgâh ili, yaş, öğrenci giriş yılı, AGNO, cinsiyet, kayıt/geliş şekli ve program düzeyi değişkenlerinin takip ettiği görülmüştür.

RA modelinde sınıflandırma öngörüleri için yüksek oranlı doğru sonuçlar üreten kuralların bir kümesi de modelin çıktısı olarak verilmektedir (sınıflar için en iyi karar verme kuralları Ek 2'de incelenebilir). Üretilen modelde yer alan ilk 4 kural "terk" durumunu (1 kategorisi) tahmin etmede 0,849-0,895 arasında doğruluk oranına sahiptir ki bu oranlar, modelin genel doğru sınıflandırma oranından daha yüksektir. Buna göre Ek 2'de yer alan kurallar incelendiğinde;

Giriş yılı "2014-2017-2020" olan ve "Acil Durum ve Afet Yönetimi, Bankacılık ve Sigortacılık, Bilgisayar ve Öğretim Teknolojileri Öğretmenliği, Coğrafya, Coğrafi Bilgi Sistemleri, Dış Ticaret, Ekonometri, Endüstri Mühendisliği, Felsefe, Gazetecilik, Halkla İlişkiler ve Tanıtım, Hukuk Büro Yönetimi ve Sekreterliği, Kamu Yönetimi, Kültürel Miras ve Turizm, Maliye, Medya ve İletişim, Perakende Satış ve Mağaza Yönetimi, Radyo, Televizyon ve Sinema, Sağlık Kurumları İşletmeciliği, Sosyal Hizmetler, Sosyoloji, Sivil Hava Ulaştırma İşletmeciliği, Tarih, Türk Dili ve Edebiyatı, Uluslararası Ticaret ve Lojistik Yönetimi, Çalışma Ekonomisi ve Endüstri İlişkileri, Çağdaş Türk Lehçeleri ve Edebiyatları, İktisat, İşletme" programlarında

okuyan öğrencilerden AGNO "1-2" kategorisinde yer alanların (yani ortalama başarının altı) %89,4 doğruluk oranı ile okul terki olarak sınıflandırıldığı;

AGNO "1-2" kategorisinde olup "Adalet, Bankacılık ve Sigortacılık, Coğrafya, Coğrafi Bilgi Sistemleri, Dış Ticaret, Ekonometri, Felsefe, Hukuk Büro Yönetimi ve Sekreterliği, Kültürel Miras ve Turizm, Maliye, Medya ve İletişim, Perakende Satış ve Mağaza Yönetimi, Sağlık Kurumları İşletmeciliği, Sosyal Hizmetler, Sosyoloji, Sivil Hava Ulaştırma İşletmeciliği, Tarih, Çalışma Ekonomisi ve Endüstri İlişkileri, Çağdaş Türk Lehçeleri ve Edebiyatları, İktisat, İş Sağlığı ve Güvenliği, İşletme" programlarında okuyan ve 2016 yılı ve öncesinde kaydolan öğrencilerin %87,4 doğruluk oranı ile okul terki olarak sınıflandırıldığı;

Giriş yılı 2020 ve öncesinde; AGNO "1-2" kategorisinde olan ve "Acil Durum ve Afet Yönetimi, Bankacılık ve Sigortacılık, Bilgisayar ve Öğretim Teknolojileri Öğretmenliği, Coğrafya, Coğrafi Bilgi Sistemleri, Dış Ticaret, Egzersiz ve Spor Bilimleri, Ekonometri, Felsefe, Gazetecilik, Grafik Tasarımı, Halkla İlişkiler ve Tanıtım, Hemşirelik, Hukuk Büro Yönetimi ve Sekreterliği, Kültürel Miras ve Turizm, Laborant ve Veteriner Sağlık, Maliye, Marka İletişimi, Medya ve İletişim, Perakende Satış ve Mağaza Yönetimi, Radyo, Televizyon ve Sinema, Rekreasyon, Sağlık Kurumları İşletmeciliği, Sağlık Yönetimi, Sosyal Hizmet, Sosyal Hizmetler, Sosyoloji, Sivil Hava Ulaştırma İşletmeciliği, Siyaset Bilimi ve Kamu Yönetimi, Siyaset Bilimi ve Uluslararası İlişkiler, Tarih, Tıbbi Dokümantasyon ve Sekreterlik, Türk Dili ve Edebiyatı, Uluslararası Ticaret ve Lojistik, Yaşlı Bakımı, Yönetim Bilişim Sistemleri, Çalışma Ekonomisi ve Endüstri İlişkileri, Çağdaş Türk Lehçeleri ve Edebiyatları, İktisat, İLİTAM, İş Sağlığı ve Güvenliği, İşletme" programlarında okuyan öğrencilerin %86,3 doğruluk oranı ile okul terki olarak sınıflandırıldığı;

Giriş yılı 2020 ve öncesi olup 25 yaşından küçük olan öğrencilerden AGNO "1-2" kategorisinde yer alanların "Acil Durum ve Afet Yönetimi, Adalet, Bankacılık ve Sigortacılık, Bilgisayar ve Öğretim Teknolojileri Öğretmenliği, Coğrafya, Coğrafi Bilgi Sistemleri, Dış Ticaret, Egzersiz ve Spor Bilimleri, Ekonometri, Felsefe, Gazetecilik, Halkla İlişkiler ve Tanıtım, Hukuk Büro Yönetimi ve Sekreterliği, Kamu Yönetimi, Kültürel Miras ve Turizm, Laborant ve Veteriner Sağlık, Maliye, Medya ve İletişim, Perakende Satış ve Mağaza Yönetimi, Radyo, Televizyon ve Sinema, Rekreasyon, Sağlık Kurumları İşletmeciliği, Sosyal Hizmetler, Sosyoloji, Sivil Hava Ulaştırma İşletmeciliği, Siyaset Bilimi ve Kamu Yönetimi, Siyaset Bilimi ve Uluslararası İlişkiler, Tarih, Tıbbi Dokümantasyon ve Sekreterlik, Türk Dili ve Edebiyatı, Uluslararası Ticaret ve Lojistik, Uluslararası Ticaret ve Lojistik Yönetimi, Yönetim Bilişim

Sistemleri, Çalışma Ekonomisi ve Endüstri İlişkileri, Çağdaş Türk Lehçeleri ve Edebiyatları, İktisat, İnsan Kaynakları Yönetimi, İş Sağlığı ve Güvenliği, İşletme" programlarında okuyanlarının %84,9 doğruluk oranı ile okul terki olarak sınıflandırıldığı görülmektedir.

4.3.2.2. CHAID

CHAID, AGNO>0 veri seti için terk eden ve devam eden öğrencilerin tahmininde 0,765 doğruluk, 0,594 duyarlılık, 0,874 hassasiyet, 0,707 F1 puanı, 0,750 Kesinlik, 0,771 NÖD, 0,494 MCC ve 0,832 AUC değeri ile yüksek performans değerlerine sahip modellerden biridir. Modelde yer alan değişkenlerin önemlilik dereceleri Tablo 31'deki gibidir. Model AGNO>0 veri seti için en yüksek hassasiyet ve kesinlik (pozitif öngörü değeri) oranlarına sahiptir.

Aşağıda yer alan Tablo 31 incelendiğinde de modeldeki sınıflandırma için en yüksek etkinin AGNO; en düşük etkinin cinsiyet değişkeni ile oluştuğu görülmektedir.

Tablo 31
AGNO>0 Veri Seti İçin CHAID Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri

Faktör/Nitelik	Etki katsayısı
AGNO1	0,6418
Program	0,1724
Öğrenci Giriş Yılı	0,1622
Yaş	0,0092
Geliş Şekli	0,0083
İkametgâh İli	0,0027
Program Düzeyi	0,0020
Cinsiyet	0,0014

Tablo 31'e göre AGNO değişkeni, öğrenci statüsünü (terk veya devam) tahmin etmede kritik bir rol oynamaktadır. Program ve öğrenci giriş yılı değişkenleri de modelde önemli yer tutar ancak AGNO kadar etkili değildir. Sınıflandırmada ağırlıklı genel not ortalamasından sonra öğrencinin kayıtlı olduğu programın ve öğrencinin hangi yıl kaydolduğunun önemli olduğu görülmektedir. Ayrıca cinsiyet ve program düzeyi gibi değişkenlerin modelin tahmin kabiliyetine nispeten daha az katkı sunduğu söylenebilir.

Ayrıca modelde ağaç yapısının da oldukça genişlediği ve 92 düğümün oluştuğu görülmüştür. Değişkenlerin modelde sınıflandırmaya etkisinin ayrıntılı incelenmesiyle ilk düzeydeki dalların

AGNO değişkenine göre oluştuğu; AGNO değeri “1 ve 2” olanlarda terk durumunun; AGNO değeri “3 ve 4” olanlarda ise devam durumunun daha yüksek olduğu söylenebilir.

CHAID modelinin 3. düzeye kadar ağaç yapısı ve değişkenlerin sınıflandırmadaki etkisi incelendiğinde aşağıdaki kuralların oluştuğu görülmektedir.

- AGNO değeri “1”;

Giriş yılı “2014” ve daha önceki yıllarda olan öğrencilerin 30 yaş ve altındakiler (Yas <= 2) %93; 30 yaş üstündekiler %89 sınıflandırma doğruluğu ile terk kategorisinde sınıflandırılmıştır.

Giriş yılı “2014 sonrası ile 2017 ve öncesi” olan öğrencilerin cinsiyete göre “erkek” olanlardan lisans ve lisans tamamlama programında okuyanlar %87; ön lisans programında okuyanlar %84 sınıflandırma doğruluğu ile terk kategorisinde sınıflandırılmıştır.

Giriş yılı “2018” olan öğrenciler %82 sınıflandırma doğruluğu ile terk kategorisinde sınıflandırılmıştır.

Giriş yılı “2019” olan öğrencilerin ikametgâh ili 1 veya 3 (İstanbul- İzmir) olanlar %73; ikametgâh ili İstanbul ve İzmir dışındaki diğer illerde olan öğrenciler ise %79 sınıflandırma doğruluğu ile terk kategorisinde sınıflandırılmıştır.

Giriş yılı “2020” olan öğrencilerin lisans ve lisans tamamlama programında okuyanlar %68; ön lisans düzeyinde okuyanlar %63 sınıflandırma doğruluğu ile terk kategorisinde sınıflandırılmıştır.

Giriş yılı “2021” olan öğrencilerden "DGS" "OSS" "YOS" ile kayıt olanlar %76 sınıflandırma doğruluğu ile devamlı kategorisinde sınıflandırılmış; "SINAVSIZ IKINCI UNIV" " ile kayıt olanlardan 25 yaş ve daha küçükleri %62; 25 yaşından büyükler ise %68 sınıflandırma doğruluğu ile devamlı kategorisinde sınıflandırılmıştır.

- AGNO değeri “2”;

Giriş yılı “2014” ve daha önceki yıllarda olan öğrencilerin 30 yaş ve altındakiler %89; 30 yaş üstündekiler ise %86 sınıflandırma doğruluğu ile terk kategorisinde sınıflandırılmıştır.

Giriş yılı “2014”olan öğrenciler %84; giriş yılı “2016 ve 2017” olan öğrenciler %78; giriş yılı “2018” olan öğrenciler ise %60 sınıflandırma doğruluğu ile terk kategorisinde sınıflandırılmıştır.

Giriş yılı “2019” olan öğrencilerden "Acil Durum ve Afet Yönetimi" "Acil Yardım Ve Afet Yönetimi" "Adalet" "Bankacılık ve Sigortacılık" "Coğrafi Bilgi Sistemleri" "Felsefe"

"Gazetecilik" "Hukuk Büro Yönetimi ve Sekreterliği" "Kültürel Miras ve Turizm" "Laborant ve Veteriner Sağlık" "Maliye" "Medya ve İletişim" "Sağlık Kurumları İşletmeciliği" "Sosyal Hizmet" "Sosyal Hizmetler" "Sosyoloji" "Sivil Hava Ulaştırma İşletmeciliği" "Siyaset Bilimi ve Kamu Yönetimi" "Siyaset Bilimi ve Uluslararası İlişkiler" "Tıbbi Dokümantasyon ve Sekreterlik" "Uluslararası Ticaret ve Lojistik Yönetimi" "Yaşlı Bakımı" "Yönetim Bilişim Sistemleri" "İnsan Kaynakları Yönetimi" "İş Sağlığı ve Güvenliği" "İşletme" Programlarında Okuyanlar %58 Sınıflandırma Doğruluğu İle Terk Kategorisinde; "Coğrafya" "Dış Ticaret" "Halkla İlişkiler ve Tanıtım" "Hemşirelik" "Kamu Yönetimi" "Radyo, Televizyon ve Sinema" "Sağlık Yönetimi" "Tarih" "Çocuk Gelişimi" "İktisat" "İLİTAM" programlarında okuyanlar %56 sınıflandırma doğruluğu ile devamlı öğrenci kategorisinde sınıflandırılmıştır.

Giriş yılı "2020" olan öğrencilerin "SINAVSIZ IKINCI UNIV" İle "Adalet" "Egzersiz ve Spor Bilimleri" "Sağlık Kurumları İşletmeciliği" "Sosyal Hizmetler" "İş Sağlığı ve Güvenliği" Programlarına Kayıt Olanlar %59; "Coğrafya" "Laborant ve Veteriner Sağlık" "Sosyoloji" "Siyaset Bilimi ve Kamu Yönetimi" "Tarih" "İnsan Kaynakları Yönetimi" programlarına kayıt olanlar %64; "Felsefe" "Rekreasyon" "Sivil Hava Ulaştırma İşletmeciliği" "Siyaset Bilimi ve Uluslararası İlişkiler" "Uluslararası Ticaret ve Lojistik" "Yönetim Bilişim Sistemleri" "İktisat" "İşletme" programlarına kayıt olanlar %69 sınıflandırma doğruluğu ile terk kategorisinde sınıflandırılmıştır.

Modelde AGNO değişkeninde kategorisi "3 ve 4" olan öğrencilerin ise ağırlıklı olarak devamlı öğrenci olduğu; sınıflandırma doğruluğunun ise %98 ile %58 arasında değiştiği görülmüştür.

4.3.2.3. C5.0

C5.0 karar ağacı, AGNO>0 veri setinde yapılan sınıflandırma işleminde terk eden ve devam eden öğrencilerin tahmininde 0,777 doğruluk; 0,668 duyarlılık; 0,847 hassasiyet; 0,747 F1 puanı, 0,736 Kesinlik; 0,800 NÖD; 0,525 MCC ve 0,826 AUC değerleri ile diğer bir başarılı modeldir. C5.0 karar ağacı modelinin genel doğru sınıflandırma oranı ile MCC diğerlerine göre yüksektir. Modelde yer alan değişkenler ve önemlilik dereceleri Tablo 32'deki gibidir.

Tablo 32
AGNO>0 Veri Seti İçin C5.0 Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri

Faktör/Nitelik	Etki katsayısı
AGNO1	0,7031
Öğrenci Giriş Yılı	0,1966
Program	0,0750
Yaş	0,0096
Cinsiyet	0,0081
İkamet ili	0,0068
Açık-Uzaktan	0,0008
Geliş/Kayıt Şekli	0
Program Düzeyi	0

Tablo 32'de yer alan değişkenler incelendiğinde öğrencilerin okul terki durumunun sınıflandırılmasında en etkili değişken başarı ile ilişkili olan AGNO iken; kişisel değişkenler olan yaş, cinsiyet ve ikamet ili faktörleri nispeten daha az etkili görünmektedir. AGNO değişkenini sırasıyla öğrenci giriş yılı, program, yaş, cinsiyet, ikametgâh ve öğrenim tipi değişkenlerinin takip ettiği görülmüştür. Öte yandan önem katsayılarına bakılırsa program düzeyi ve programa kayıt şeklinin modelde sınıflandırmaya etkisinin olmadığı söylenebilir.

Ayrıca Modelin ağaç yapısının 827 düğüm ile oldukça karmaşık oluştuğu söylenebilir. Buna göre ilk 4 düzeydeki karar ağacı şeması Ek 3'teki gibidir.

Karar ağacı ile modelden elde edilen kurallar bütünü incelendiğinde ilk kırılmanın AGNO'ya göre oluştuğu, daha sonraki düzeylerde bunu Öğrenci giriş/kayıt yılının ve program değişkeninin takip ettiği görülmekte ve buna göre aşağıdaki bulgular ortaya çıkmaktadır.

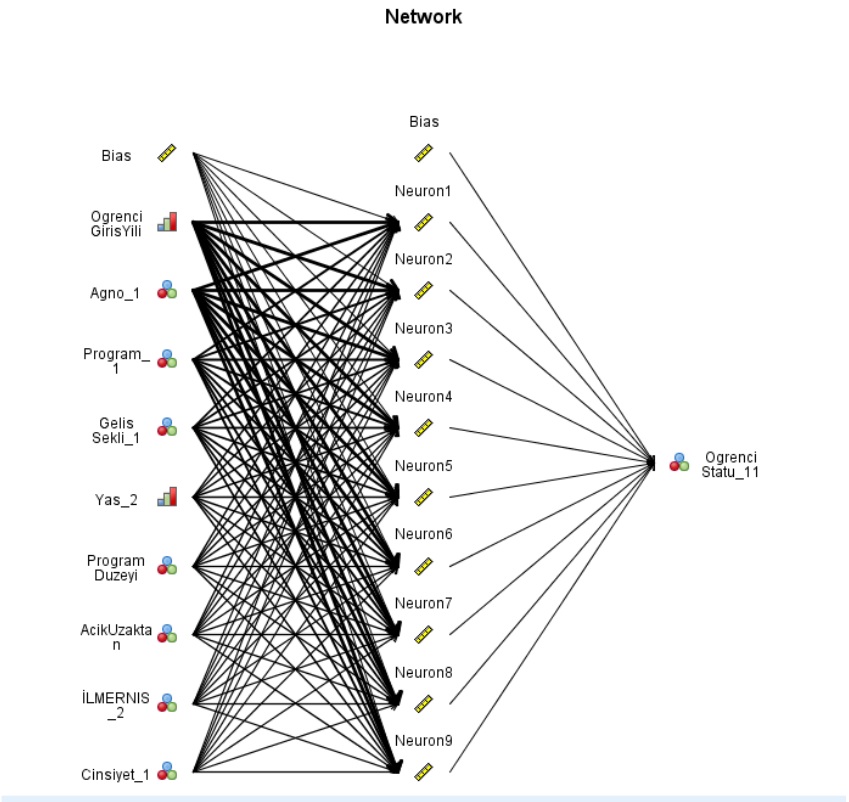
AGNO değeri "1" olanlardan giriş yılı "2010-2011-2012-2013-2014-2015-2016-2017-2018-2019-2020" olanlar %80 oranla okulu terk etmekte; giriş yılı "2021" olanlar %69 oranla okula devam etmektedir.

AGNO değeri "2" olanlardan giriş yılı giriş yılı "2010-2011-2012-2013-2014-2015-2016-2017-2018-2019-2020" olanlar %68 oranla okulu terk etmekte; giriş yılı "2021" olanlar %85 oranla okula devam etmektedir.

AGNO değeri "3 ve 4" olanların ise çoğunlukla okula devam ettiğine ilişkin sınıflandırma kuralları üretilmiş ve bu tahminlerin doğruluk oranlarının da alt düzey dallarda programlar yer aldığı için program türü faktörüne göre değiştiği gözlenmiştir.

4.3.2.4. YSA

Tüm veri seti için AUC kriterine göre performans değeri yüksek olan dolayısıyla öğrencilerin terk ve devam durumlarını sınıflandırmada başarılı olan YSA modelinin AGNO>0 veri seti için de yüksek performansı gösteren modellerden olduğu görülmüş ve bu başlıkta modelde yer alan değişkenlerin sınıflandırmaya etkisi incelenmiştir. YSA modeli 0,847 AUC değeri ile diğerlerine göre en yüksek performansı göstermiştir. Ayrıca modelin genel doğruluk oranı 0,776'dır. Modelde 0,658 duyarlılık, 0,851 hassasiyet, 0,742 F1 puanı, 0,738 Kesinlik, 0,796 NÖD ve 0,521 MCC değeri gözlenmiştir. Modelde AUC değerinin 0,847 ile 1'e yakın olması, sınıflar arasında iyi bir ayrım yapabildiğini gösterir. YSA modelindeki faktörlerin ve ara katmanda oluşan nöronların ağ yapısı Şekil 39'daki gibidir.



Şekil 39. AGNO>0 veri seti için YSA modelinde sinir ağları şeması

Şekil 39 incelendiğinde YSA modelinde yer alan değişkenler içinde "Öğrenci giriş yılı" ve AGNO1'in en öncelikli değişkenler olduğu bunu sırasıyla; program, geliş şekli, yaş, program düzeyi, öğrenim tipi, ikametgâh ili ve cinsiyet değişkenlerinin takip ettiği görülmektedir. Ayrıca modelde yer alan değişkenlerin etki katsayıları Tablo 33'te verilmiştir.

Tablo 33
AGNO>0 Veri Seti İçin YSA Modelinde Okul Terkini Etkileyen Değişkenler ve Önem Dereceleri

Faktör/Nitelik	Etki katsayısı
Öğrenci Giriş Yılı	0,2710
AGNO1	0,2622
Program	0,1390
Geliş Şekli	0,1051
Yaş	0,0692
Program Düzeyi	0,0538
Açık-Uzaktan	0,0451
İkametgâh ili	0,0295
Cinsiyet	0,0251

Tablo 33'e göre modelde AGNO ile öğrenci giriş yılının okul terk eden öğrencilerin sınıflandırılmasında en çok ve birbirine yakın düzeyde etkiye sahip olduğu, cinsiyet ve ikamet ili değişkenlerinin etkisinin ise düşük düzeyde olduğu söylenebilir.

Araştırmanın bu aşamasında AGNO>0 veri setinde okul terki, ile ilişkili bulunan ve 4 ayrı veri madenciliği modelleri ile incelenen değişkenler için bir özet tablo sunulmuştur. Buna göre RA, C5.0, CHAID ve YSA modellerinde gözlenen etkili faktörler Tablo 34'te yer almaktadır.

Tablo 34
AGNO>0 Veri Seti için Okul Terkine Etki Eden Faktörlerin Karşılaştırılması

AGNO>0 VERİ SETİ							
RA		C5.0		CHAID		YSA	
Faktör	*Etki katsayısı*	*Faktör*	*Etki katsayısı*	*Faktör*	*Etki katsayısı*	*Faktör*	*Etki katsayısı*
Program	...	AGNO1	0,7031	AGNO1	0,6418	Öğrenci Giriş Yılı	0,271
İkametgâh İli	...	Öğrenci Giriş Yılı	0,1966	Program	0,1724	AGNO1	0,2622
Yaş	...	Program	0,0750	Öğrenci Giriş Yılı	0,1622	Program	0,139
Öğrenci Giriş Yılı	...	Yaş	0,0096	Yaş	0,0092	Geliş/Kayıt Şekli	0,1051
AGNO1	...	Cinsiyet	0,0081	Geliş/Kayıt Şekli	0,0083	Yaş	0,0692
Cinsiyet	...	İkamet ili	0,0068	İkametgâh İli	0,0027	Program Düzeyi	0,0538
Geliş/Kayıt Şekli	...	Açık-Uzaktan	0,0008	Program Düzeyi	0,002	Açık-Uzaktan	0,0451
Program Düzeyi	...			Cinsiyet	0,0014	İkametgâh İli	0,0295
						Cinsiyet	0,0251

Tablo 34'e göre, sınıflandırma yapan modellerin farklı değişkenleri önceliklendirdiği görülmektedir. RA modeli sırasıya program türü, ikamet ili ve yaş faktörlerini ilk üç sırada tahmin değişkeni olarak ele alırken; C5.0, CHAID ve YSA modellerinde sıralama değişse de önemli olan ilk üç faktör, AGNO, öğrenci giriş yılı ve program türü değişkenleridir. C5.0

modelinde geliş şekli ve program düzeyi faktörlerinin sınıflandırmada etkisi "0" bulunmuşken; RA ve CHAID modellerinde öğrenim tipi (açık-uzaktan) değişkeni hiç yer almamaktadır. Tüm tahmin değişkenlerini model için etkili bulan YSA modelinde ise okul terki sınıflandırmasında en düşük etkiye sahip faktörler öğrenim tipi, ikamet ili ve cinsiyet faktörleridir.

Tüm bunların yanında tüm veri seti için oluşturulan modellerde etkili olan faktörler ile AGNO> 0 veri seti için oluşturulan modellerdeki faktörler incelendiğinde AGNO değişkeninin etki katsayısının düştüğü görülecek; hatta AGNO>0 veri setinde genel sınıflandırma doğruluğu en yüksek model olan RA'da AGNO değişkenin ağacın ilk düğümlerinin oluşmasında etkili olmadığı (ancak 5. sırada yer aldığı) görülmektedir. Bunun yerine AGNO>0 veri setinde sınıflandırma için yine bir akademik faktör olan program türünün önemli olduğu görülmektedir.

4.4. AGNO'ya göre Öğrencilerin Sınıflandırılmasında VM Modellerinin Performansları Nasıl Değişmektedir ve Öğrenci Başarısı ile İlişkili Faktörler Nelerdir?

AGNO değişkeni, önceki araştırma sorularına dair bulgular incelendiğinde okul terkinin belirlenmesinde yüksek etkiye sahip faktörlerden biridir ve tüm veri madenciliği modellerinde önemli etkenlerden biri olarak ortaya çıkmıştır. Araştırmanın bu aşamasında öğrencilerin ağırlıklı genel not ortalaması 2 ve üstünde olanlar başarılı (1-kategorisi) 2'nin altında olanlar ise başarısız (0 kategorisi) olarak etiketlenmiş ve öğrencilerin sınıflandırılmasında modellerin başarısı incelenmiştir.

Yükseköğretimde açık ve uzaktan öğrenmedeki öğrencilerin Ağırlıklı Genel Not Ortalamasının (AGNO) tahmininde;

i. veri madenciliği sınıflandırma algoritmalarının performansları nasıl değişmektedir?
ii. Veri madenciliği modellerine göre AGNO ile ilişkili faktörler nelerdir?

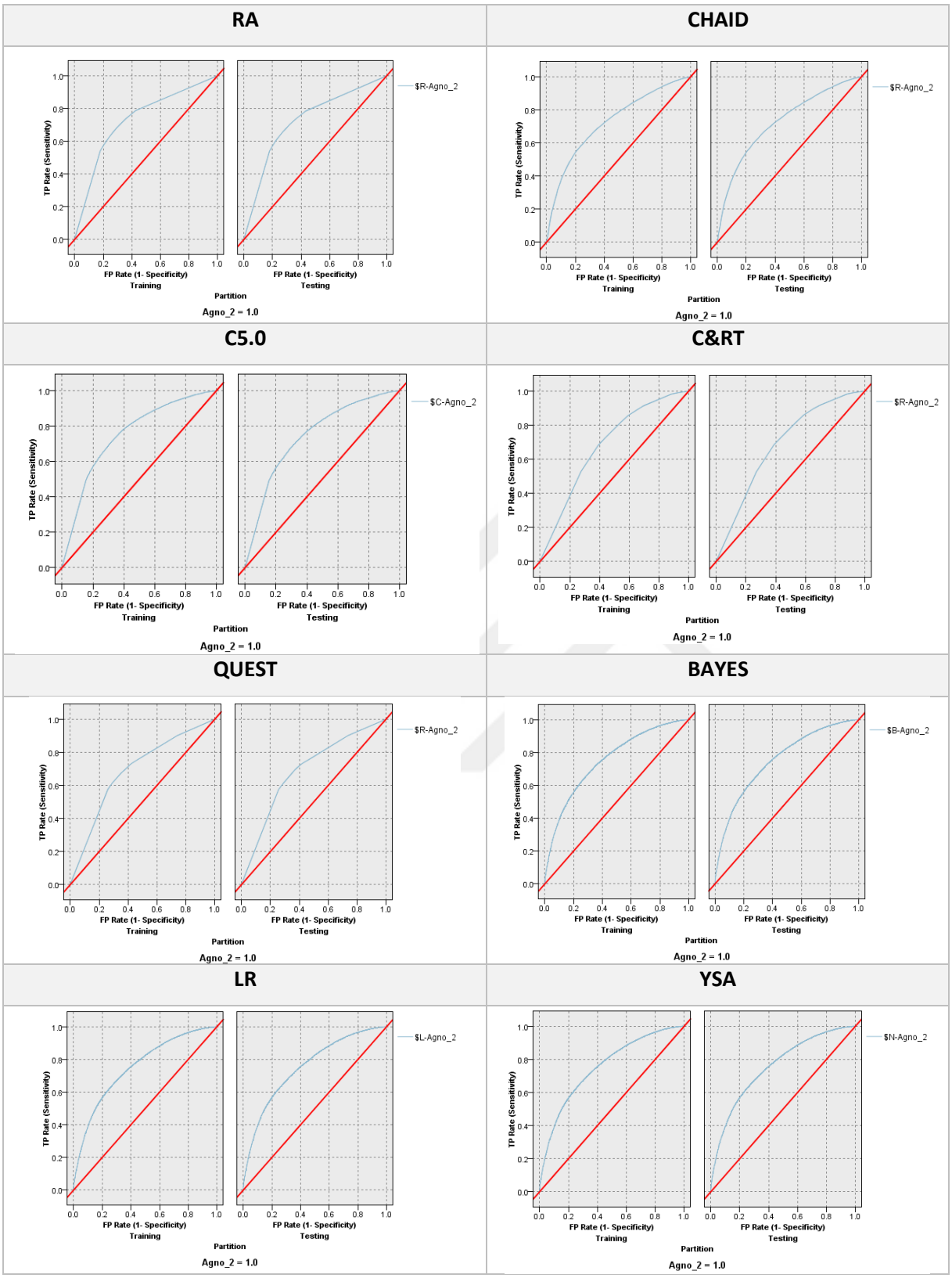
Araştırma sorusunda 3-i alt problemine yönelik oluşturulan modellerde eğitim ve test verilerine ait Doğruluk, Duyarlılık, Hassasiyet, F1 puanı, Kesinlik, NÖD, MCC, AUC kriterlerinin performans değerleri Tablo 35'te yer almaktadır.

Tablo 35
Öğrenci Başarısının (AGNO) Tahmini için VM Algoritmalarının Performanslarının Karşılaştırılması

Algoritma	Veri	Doğruluk	Duyarlılık	Hassasiyet	F1 puanı	Kesinlik	NÖD	MCC	AUC
RA	Eğitim	0,689	0,656	0,731	0,692	0,761	0,621	0,385	0,729
	Test	0,688	0,655	**0,730**	**0,691**	**0,760**	0,619	**0,382**	0,727
CHAID	Eğitim	0,669	0,674	0,662	0,668	0,722	0,610	0,334	0,726
	Test	0,671	0,676	0,665	0,670	0,725	0,611	0,338	0,728
C5.0	Eğitim	0,703	0,774	0,611	0,683	0,721	0,675	0,391	0,750
	Test	**0,696**	0,767	0,602	0,675	0,716	0,665	0,375	0,745
C&RT	Eğitim	0,661	0,856	0,408	0,553	0,653	0,686	0,299	0,683
	Test	0,665	**0,858**	0,413	0,557	0,656	**0,689**	0,305	0,687
QUEST	Eğitim	0,665	0,726	0,585	0,648	0,695	0,622	0,314	0,689
	Test	0,666	0,728	0,586	0,649	0,697	0,622	0,317	0,690
BAYES	Eğitim	0,689	0,744	0,617	0,675	0,716	0,650	0,363	0,752
	Test	0,690	0,745	0,617	0,675	0,718	0,650	0,365	0,754
LR	Eğitim	0,688	0,744	0,615	0,673	0,715	0,649	0,361	0,753
	Test	0,689	0,743	0,617	0,674	0,717	0,648	0,363	0,754
YSA	Eğitim	0,688	0,746	0,613	0,673	0,715	0,650	0,362	0,754
	Test	0,689	0,747	0,613	0,674	0,716	0,650	0,363	**0,756**
DVM	Eğitim	0,703	0,784	0,599	0,679	0,717	0,681	0,390	0,736
	Test	0,696	0,778	0,592	0,672	0,714	0,671	0,377	0,726

Buna göre öğrenci başarısının sınıflandırılması için en yüksek genel doğruluk oranı C5.0 ve DVM modeli (0,696) ile; en yüksek duyarlılık C&RT modelinde (0,858) ortaya çıkmıştır. En yüksek hasasiyet değeri ise 0,730 ile RA modeline aittir. Bu göstergelerin yanında modeller grafiksel bir kriter olan ROC eğrileri ile de karşılaştırılmış, sonuçlar Şekil 40'ta sunulmuştur.

Genel olarak tüm modellerde MCC'nin düşük olduğu ve hasasiyet ile duyarlılık değerleri arasında farklar oluştuğu görülmektedir. Dolayısı ile bazı modeller pozitif (yani başarılı) sınıfın tahmininde gösterdiği başarıyı negatif sınıfın belirlenmesinde gösteremeyebilir. Bu sebeple çalışmanın amacına uygun olarak en yüksek doğruluk oranına sahip C5.0 modeli, en yüksek duyarlılık ve negatif öngörü değerine sahip C&RT modeli ile en yüksek hassasiyet ve MCC değerine sahip RA modellerinin istenilen sonuçları vereceği düşünüldüğünden bu modellerin yapısı ve öğrenci başarının öngörülmesinde etkili değişkenlerin neler olduğu ayrıntılı ele alınmıştır.

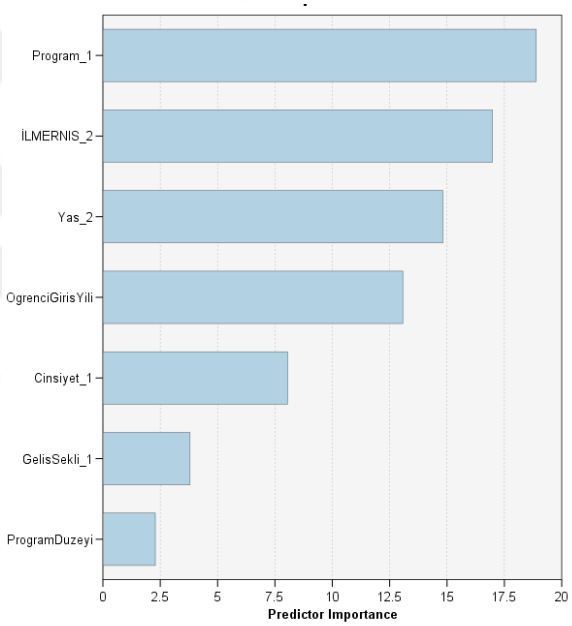


Şekil 40. Öğrenci başarısı sınıflandırmasında modellerin ROC eğrileri

Buna göre araştırma sorularında yer alan 3-ii alt problemi için öğrenci başarısının sınıflandırılmasında etkili olan faktörler ve veri madenciliği modelerinin yapısına yönelik elde edilen bulgular aşağıdaki gibidir.

4.4.1. Rastgele Ağaçlar

Rastgele ağaçlar algoritması akademik başarının tahmininde özellikle Hassasiyet 0,730; F1 puanı 0,691; Kesinlik 0,760; ve MCC 0,382 değeri ile diğer modellerden daha iyi sonuçlar üretmiştir. Modelde yer alan akademik başarının belirlenmesine etki eden faktörler ve etki onarları Şekil 41'de yer almaktadır.



Şekil 41. Öğrenci başarısının tahmininde RA modelinin tahmin değişkenleri ve önemi

Şekilde 41 incelendiğinde öğrencilerin başarılı-başarısız olarak sınıflandırılmasında sırasıyla, program, ikamet ili, yaş, öğrenci giriş yılı, cinsiyet, geliş şekli ve program düzeyi değişkenleri etkili olmaktadır. Rastgele ağaçlar modeli ayrıca en sık kullanılan kuralları içerecek en yüksek doğruluk oranına sahip kuralların bir listesini oluşturmaktadır (ilgili tablo Ek 4'te yer almaktadır).

Modelde "Başarılı" kategorisinin tahmininde en yüksek doğruluk oranına sahip ilk beş kural aşağıdaki gibidir.

“AF, DGS, OZEL” kategorilerin “Acil Durum ve Afet Yönetimi, Acil Yardım ve Afet Yönetimi, Egzersiz ve Spor Bilimleri, Hemşirelik, Medya ve İletişim, Rekreasyon, Sağlık Yönetimi, Sosyal Hizmet, Çocuk Gelişimi, İLİTAM” programlarına 2016 -2020 yılları arasında kayıt olan öğrenciler %89,9 doğruluk oranı ile “1-Başarılı” kategorisinde;

Giriş yılı 2016 -2020 yılları arasında olup “AF, DGS, OZEL” yolu ile “Acil Durum ve Afet Yönetimi, Acil Yardım ve Afet Yönetimi, Bilgisayar ve Öğretim Teknolojileri Öğretmenliği, Egzersiz ve Spor Bilimleri, Endüstri Mühendisliği, Hemşirelik, Radyo, Televizyon ve Sinema, Rekreasyon, Sağlık Yönetimi, Sosyal Hizmet, Çocuk Gelişimi, İLİTAM” programlarına kayıt olan öğrenciler %89 doğruluk oranı ile “1-Başarılı” kategorisinde;

“AF, DGS, OZEL” kayıt yolu ile programlara giren, “İstanbul ve İzmir” dışındaki diğer illerde ikamet eden, lisans ve lisans tamamlama düzeyinde “2020” öncesinde “ Acil Durum ve Afet Yönetimi, Acil Yardım ve Afet Yönetimi, Bilgisayar ve Öğretim Teknolojileri Öğretmenliği, Egzersiz ve Spor Bilimleri, Endüstri Mühendisliği, Hemşirelik, Medya ve İletişim, Radyo, Televizyon ve Sinema, Rekreasyon, Sağlık Yönetimi, Sosyal Hizmet, Çocuk Gelişimi, İLİTAM” programlarına kayıt olan öğrenciler %88,4 doğruluk oranı ile “1-Başarılı” kategorisinde;

“AF, DGS, OZEL” yolları ile programlara giren, “İstanbul” dışındaki diğer illerde ikamet eden, lisans ve lisans tamamlama düzeyinde “2020” öncesinde “Acil Durum ve Afet Yönetimi, Acil Yardım Ve Afet Yönetimi, Bilgisayar ve Öğretim Teknolojileri Öğretmenliği, Egzersiz ve Spor Bilimleri, Endüstri Mühendisliği, Hemşirelik, Medya ve İletişim, Radyo, Televizyon ve Sinema, Rekreasyon, Sağlık Yönetimi, Sosyal Hizmet, Yaşlı Bakımı, Çocuk Gelişimi, İLİTAM” programlarına kayıt olan öğrenciler %87,8 doğruluk oranı ile “1-Başarılı” kategorisinde;

“2020” öncesinde “AF, DGS, OZEL” kategorilerinden biri ile “Acil Durum ve Afet Yönetimi, Acil Yardım ve Afet Yönetimi, Egzersiz ve Spor Bilimleri, Endüstri Mühendisliği, Hemşirelik, Radyo, Televizyon ve Sinema, Rekreasyon, Sağlık Yönetimi, Sosyal Hizmet, Türk Dili ve Edebiyatı, Çocuk Gelişimi, İLİTAM” programlarına kayıt olan öğrenciler %86,2 doğruluk oranı ile “1-Başarılı” kategorisinde sınıflandırılmıştır.

Bu kurallar bütününde program türlerinin (ve tabi diğer değişkenlerin) etkisi açık şekilde görüldüğü için özellikle kurumlara rehberlik etmesi beklenebilir.

4.4.2. C5.0

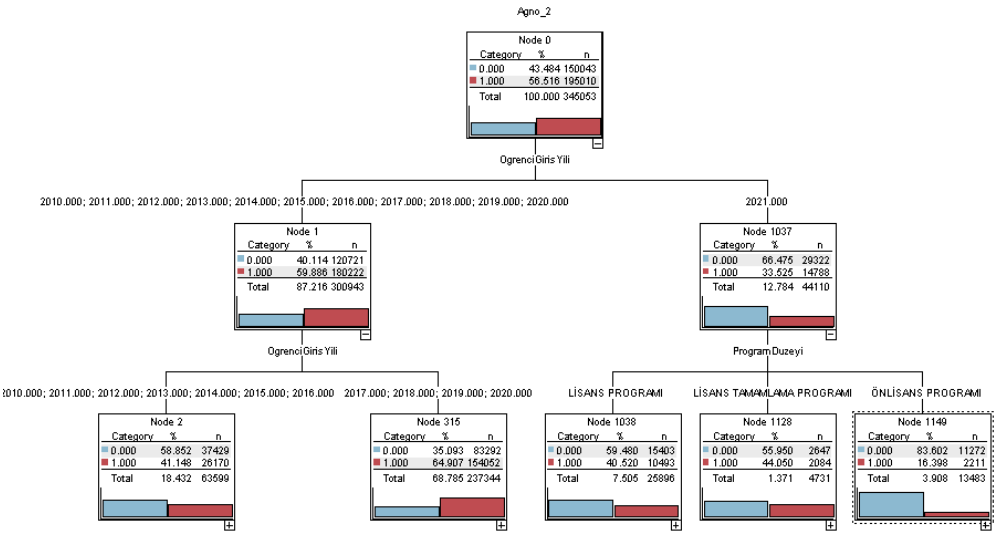
Terk tahmini için tüm veri setlerinde yüksek performans gösteren C5.0 modeli, öğrencilerin başarılı-başarısız olarak sınıflarının öngörülmesi için de 0,696 ile genel sınıflandırmada en yüksek doğruluk oranına erişmiştir. Modelde yer alan faktörler ve etki oranları Tablo 36'da yer almaktadır.

Tablo 36
Öğrenci Başarısının Tahmininde C5.0 Modelindeki Tahmin Değişkenleri ve Önemi

Faktör/Nitelik	Etki katsayısı
Öğrenci Giriş Yılı	0,6069
Yaş	0,2998
İkametgâh ili	0,0635
Cinsiyet	0,0298
Geliş/Kayıt Şekli	0
Program Düzeyi	0
Program	0
Açık-Uzaktan	0

Tablo 36 incelendiğinde öğrencilerin akademik başarıya göre sınıflandırılmasında C5.0 modelde öğrenci giriş yılının en yüksek etkiye sahip olduğu; bunu sırasıyla yaş, ikamet ili, cinsiyet, faktörlerinin takip ettiği görülmektedir. Ayrıca geliş/kayıt şekli, program düzeyi, program ve öğrenim tipi faktörleri modelde hiçbir etkiye sahip değildir.

C5.0 modeli ile öğrenicnin akademik başarısının sınıflandırılmasında 1161 düğüm oluşmaktadır. Bunlardan ağacın 2. düzeyine kadar ki yapısı Şekil 42'de verilmiştir.



Şekil 42. Öğrenci başarısı sınıflandırmasında C5.0 modelinin ağaç yapısı (2. düzey)

Şekil 42 incelendiğinde modelde öğrencilerin %56,5'inin başarılı; %43,5'inin başarısız kategorisinde sınıflandırıldığı görülmekte; bunlardan %87,2'si de giriş yılı faktörüne göre ilk düğümde yer almaktadır. Buna göre giriş yılı "2010-2011-2012-2013-2014-2015-2016-2017-2018-2019-2020" olanlardan %59,9'unun başarılı; %40,1'inin başarısız kategorisinde sınıflandırıldığı ancak bunun da yine alt dallarda giriş yılına göre farklılaştığı gözlenmektedir. Giriş yılı 2021 olanlardan (ki bunlar modeldeki toplam verinin %12,8'i) ise %33,5'inin başarılı; %66,5'inin başarısız olarak sınıflandırıldığı görülmektedir. Giriş yılı 2021 olanlarda program düzeyine göre oluşan bir sonraki dallarda lisans düzeyindeki öğrenicilerden %40,5'inin başarılı; %59,5'inin başarısız; lisans tamamlama düzeyindeki öğrenicilerden %44,5'inin başarılı; %55,5'inin başarısız; önlisans düzeyindeki öğrenicilerden %16,5'inin başarılı; %83,5'inin başarısız olarak tahmin edildiği görülmektedir.

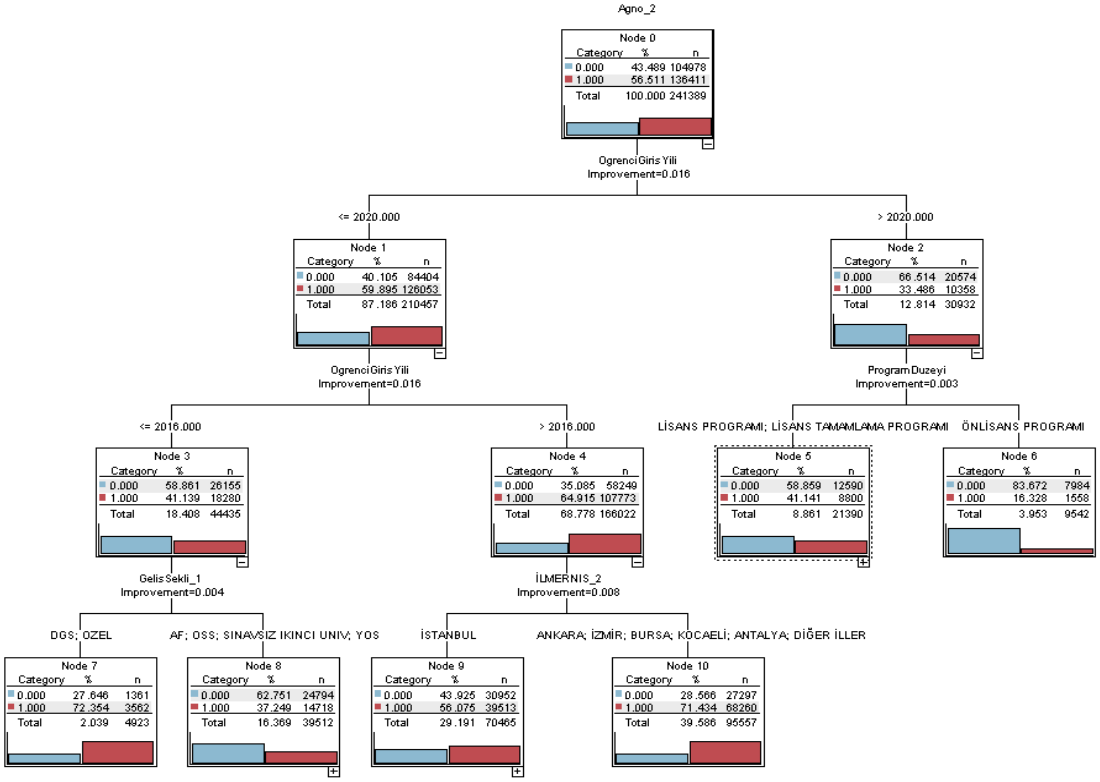
4.4.3. C&RT

Okul terkinin tahmin edilmesi ve önemli değişkenlerin ortaya çıkartılmasına yönelik araştırmanın diğer alt probleminde C&RT modeli ile etkili sonuçlar alındığı belirtilmişti. Benzer şekilde bu aşamada da ağırlıklı genel not ortalamasına göre öğrencilerin başarı durumlarının sınıflandırılmasında C&RT modelinin 0,665 doğruluk; 0,858 duyarlılık; 0,413 hassasiyet; 0,557 F1 puanı; 0,656 Kesinlik; 0,689 NÖD; 0,305 MCC ve 0,687 AUC değerleri ile özellikle model duyarlılığı açısından yüksek performansa sahip olduğu bulunmuştur. Akademik başarının "AGNO2" değişkenine göre tahmininde modelde yer alan değişkenlerin önemlilik dereceleri Tablo 37'deki gibidir.

Tablo 37
Öğrenci Başarısının Tahmininde C&RT Modelindeki Değişkenler ve Önem Dereceleri

Faktör/Nitelik	Etki katsayısı
Öğrenci Giriş Yılı	0,6796
İkametgâh ili	0,1396
Yaş	0,0647
Geliş/Kayıt Şekli	0,0367
Program	0,0313
Cinsiyet	0,0224
Program Düzeyi	0,0187
Açık-Uzaktan	0,007

Tablo 37'de yer alan katsayılar incelendiğinde tahminde "öğrenci giriş yılı" faktörünün diğerlerine göre oldukça yüksek oranda etkili olduğu görülecektir. Bunu sırasıyla ikamet ili, yaş, geliş/kayıt şekli, program, cinsiyet, program düzeyi ve öğrenim tipi (açık-uzaktan) faktörleri izlemektedir. C&RT Modeli C5.0'da yer alan geniş ve karmaşık ağaç yapısına karşın oldukça basit (5 düzeyli) 22 düğümden oluşan bir model ortaya çıkarmıştır. Buna göre düğümlerin yapısı Şekil 43'teki gibidir.



Şekil 43. Öğrenci başarısı sınıflandırmasında C&RT modelinin ağaç yapısı (3. düzey)

Şekil 43 incelendiğinde modelin öğrencilerin %56,5'ini başarılı; %43,5'ini ise başarısız kategorisinde sınıflandırdığı görülmektedir. Bir sonraki düzey öğrenci giriş yılına göre oluşturulmuş ve 2020 yılı öncesinde kaydolan öğrencilerin (ki bu toplamın %87'si kadar) %60'ının başarılı; %40'nın başarısız; 2020 yılı sonrasında kaydolan öğrencilerin ise (ki bu toplamın %13'ü kadar) %33,5'inin başarılı; %66,5'inin başarısız olarak belirlendiği görülmüştür.

Bir alt düzeyde karar ağacı incelendiğinde 2020 sonrasında kaydolan öğrencilerden öğrenim düzeyine göre lisans ve lisans tamamlama programlarında okuyanların %41'i başarılı; %59'u başarısız; ön lisans programlarında okuyanların ise %16'sı başarılı; %84'ü başarısız olarak öngörülmektedir. Öte yandan, 2016 öncesinde "DGS-ÖZEL" kategorisinde programlara

kaydolan öğrencilerin %72'sinin başarılı; %28'inin başarısız; "AF-ÖSS-SINAVSIZ İKİNCİ ÜNİVERSİTE" kategorisinde programlara kaydolan öğrencilerin %37'sinin başarılı; %63'ünün başarısız olduğu belirlenmiştir.

2016 sonrasında (2020 yılı öncesinde) kaydolan öğrencilerden İstanbul ilinde ikamet edenlerin %56'sı başarılı; %44'ü başarısız; Ankara, İzmir, Bursa, Kocaeli, Antalya ve diğer illerde ikamet edenlerin ise %71'i başarılı; %29'u başarısız olarak sınıflandırılmaktadır.

Araştırmanın bu aşamasında ağırlıklı genel not ortalamasına göre öğrenci başarısının sınıflandırılmasında geliştirilen ve yukarda sunulan veri madenciliği modelleri ile incelenen değişkenler için bir özet tablo sunulmuştur. Buna göre C5.0, C&RT ve RA modellerinde gözlenen faktörler Tablo 38'de yer almaktadır.

Tablo 38
Öğrenci Başarısı ile İlişkili Faktörlerin Karşılaştırılması

Öğrenci Başarısının Tahmini					
C5.0		C&RT		RA	
Faktör	*Etki katsayısı*	*Faktör*	*Etki katsayısı*	*Faktör*	*Etki katsayısı*
Öğrenci Giriş Yılı	0,6069	Öğrenci Giriş Yılı	0,6796	Program	...
Yaş	0, 2998	İkametgâh ili	0,1396	İkametgâh ili	...
İkametgâh ili	0, 0635	Yaş	0,0647	Yaş	...
Cinsiyet_1	0, 0298	Geliş/Kayıt Şekli	0,0367	Öğrenci Giriş Yılı	...
Açık-Uzaktan	0	Program	0,0313	Cinsiyet	...
Program	0	Cinsiyet	0,0224	Geliş/Kayıt Şekli	...
Program Düzeyi	0	Program Düzeyi	0,0187	Program Düzeyi	...
Geliş/Kayıt Şekli	0	Açık-Uzaktan	0,007		

Tablo 38 incelendiğinde modellere göre faktörlerin sınıflandırmada etkisinin değiştiği görülmektedir. C&RT algoritması nerdeyse tüm değişkenleri sınıflandırmada kullanmışken; C5.0 algoritması sadece dört faktörün sınıflandırma için etkili olduğu sonucuna ulaşmış; modelde geliş şekline, program düzeyine, program türü ve tipine yer vermemiştir. Geliş şekli ve program düzeyi diğer iki model de düşük etkiye sahiptir. Giriş yılı öğrenci başarısının sınıflandırılmasında C5.0 ve C&RT modelinde en önemli etken olarak ortaya çıkmıştır. Yaş faktörünün ise her üç modelde için de sınıflandırmada etkisi yüksek faktörlerden olduğu söylenebilir. Bunun yanında giriş yılı, ikamet adresi ve cinsiyet faktörleri de değişen oranlarda sınıflandırmada etkili bulunmuştur.

BÖLÜM V

SONUÇ, TARTIŞMA ve ÖNERİLER

5.1. Sonuç ve Tartışma

Öğrenci terki, öğrencileri, ailelerini, iş örgütlerini, üniversiteleri, devlet kurumlarını ve dolayısıyla ülke ekonomisini etkileyen önemli bir konudur. Okul terkinin ve ağırlıklı genel not ortalamasının öngörülmesine yönelik kullanılabilecek veri madenciliği algoritmalarının belirlenmesi ve terk olgusunda etkili değişkenlerin ortaya çıkartılması amacıyla yapılan bu araştırmada bulgular ışığında aşağıdaki sonuçlara ulaşılmıştır.

1. Araştırmanın temelinde açık ve uzaktan öğrenme öğrencilerine ait demografik ve diğer değişkenlerin istatistiki incelemesi sonucunda;

- İncelenen yükseköğretim programlarında en az öğrenci Bilgisayar Programcılığı, Bilgisayar ve Öğretim Teknolojileri Öğretmenliği İle Çağdaş Türk Lehçeleri ve Edebiyatları programlarında; en fazla öğrenci ise Çocuk Gelişimi ile Sosyoloji programlarında bulunmaktadır. Buradaki sonuçlar incelenirken açık ve uzaktan eğitimde bazı programların seneler içinde öğrenci alımlarının YÖK veya ilgili kurum tarafından durdurulduğu veya kontenjan sınırlamalarının bulunduğu göz önüne alınmalıdır.
- Genel olarak en yüksek terk oranı Perakende Satış ve Mağaza Yönetimi (%84,3); Çağdaş Türk Lehçeleri ve Edebiyatları (%85,7) ve İktisat (%74,3) programlarında; en düşük terk ise oranı Sosyal Hizmet (%19,8), İLİTAM (%16,1) ve Acil Yardım ve Afet Yönetimi (%13,4) programındadır.
- En yüksek okul terki ön lisans programlarında (%56,6) gerçekleşmektedir. Lisans programlarında okuyan öğrencilerin %48,8'i okulu terk ederken en düşük oran lisans tamamlama (20,1)programlarındadır.

- Açıköğretim öğrencilerinin %49,6'sı; uzaktan eğitim öğrencilerinin %46,3'ü okulu terk etmektedir.
- Giriş yıllarına göre okul terki incelendiğinde oransal olarak en az okul terkinin 2018 ve 2021 yıllarında olduğu; en yüksek okul terkinin ise 2012 ve 2013 yıllarında gerçekleştiği görülmüştür.
- Bulgular sonucunda ÖSYM tarafında yapılan merkezi sınavlar ile programlara yerleşenlerin %58,7'sinin okulu terk ettiği; daha önce bir üniversite programına kaydolmuş veya tamamlamış olanlardan sınavsız ikinci üniversite kayıt koşulu ile programlara yerleşenlerin ise %49,9'unun terk ettiği görülmektedir. En yüksek devamlılık oranı ise (sadece %24,9'u okulu terk etmekte) DGS ile programlara yerleşen öğrencilere aittir.
- Cinsiyete göre yapılan incelemede erkeklerin %59'unun; kadınların %44,8'inin okulu terk ettiği görülmüştür. Cinsiyet ile okul terki ilişkisi pek çok araştırmada incelenmiş farklı sonuçlar ortaya çıkmıştır. Shyamala (2008) Hindistan'da yürüttüğü araştırmada kız öğrencilerin okulu daha çok terk ettiğini bulmuşken Avrupa'da Eğitim ve Öğretimde Okul Terkiyle Mücadele: Stratejiler, Politikalar ve Önlemler Raporunda (2014) genel olarak Avrupa yükseköğretiminde erkeklerin okul terk oranının kızlara göre daha yüksek olduğu belirtilmiştir.
- Yaş grupları incelendiğinde en yüksek terk oranının %57,5 ile 26-30 yaş grubunda olduğu; 41 yaş ve üstü öğrencilerin ise %38,2 ile okulu en az terk eden yaş grubu olduğu belirlenmiştir. Pierrakeas vd. (2004) ise okulu bırakan öğrencilerin yüzde 57,4'ünün 30-39 yaş grubuna mensup olduğunu ortaya koymuştur.
- Öğrencilerin yaklaşık yarısının İstanbul'da ikamet ettiği; bunu sırasıyla Ankara ve İzmir illerinin takip ettiği görülmüştür. İstanbul'da ikamet eden öğrencilerin %56,1'i, Kocaeli'nde ikamet edenlerin %49,3'ü okulu terk ederken (ki en yüksek iki oran); 6 büyükşehir (İstanbul, Ankara, İzmir, Bursa, Kocaeli, Antalya) dışında yaşayanların %42,3'ü öğrenimi bırakmaktadır.
- Ağırlıklı genel not ortalaması esasında sürekli bir veri iken analizlerde kategorik olarak yer almıştır. Mevzuat gereği AGNO değeri 2 ve üstünde olanlar başarılı; 2'nin altında olanlar başarısız olarak değerlendirilmektedir. AGNO değeri "0" olanların %82,5'inin okulu terk ettiğini daha önce belirtilmişti. Ayrıca AGNO değeri (0,1] aralığında olanların %68'i; (1,2] aralığında olanların %57,1'i; (2,3] aralığında olanların %26,1'i; (3,4] aralığında olanların %13,8'i okulu terk etmektedir.

2. Araştırmanın sorularına istinaden açık ve uzaktan öğrenme öğrencilerinin okul terkine göre sınıflandırılmasında başarılı olan veri madenciliği yöntemlerine dair sonuçlar iki veri seti için ayrı ayrı sunulmuştur. Okul terkinin tahimininde ağırlıklı genel not ortalaması "0" olan ve dolayısı ile kayıt sonrası herhangibir eğitim öğretim faaliyetinde başarılı olmayan öğrencilerin %82.5'inin okul terkettiği bilindiğinden bu grup veriden çıkartılarak yapılacak analizlerden de beklenen sonucun alınıp alınmadığı AGNO>0 veri seti için tekrar incelenmiştir. Araştırmanın ikinci aşamasında hem tüm veri seti hem de AGNO>0 veri seti için yapılan analizlere dayalı sonuçlar yer almaktadır.

- Tüm veri seti için RA, CHAID, C5.0, C&RT, QUEST, BAYES, LR ve YSA algoritmalarında modeller öncelikle eğitim verisi ile eğitilmiş ardında test verisi ile modelin testi gerçekleştirilmiştir. Yapılan analizlerden elde edilen bulgulara göre modellerin genel doğru sınıflandırma oranı 0,754 ile 0,790 arasında değişmektedir.
- AGNO>0 veri seti için eğitilen RA, CHAID, C5.0, C&RT, QUEST, BAYES, LR, YSA, DVM algoritmalarında model eğitim ve test verilerinde sonuçların birbirine yakın olduğu görülmektedir. Ayrıca tüm modellerde genel sınıflandırma doğruluğu da birbirine oldukça yakın ve 0,719 ile 0,777 arasında değişmektedir.
- Bu değerlerden yolara çıkarak genel model doğruluk oranlarının tüm veri seti için daha yüksek; AGNO>0 veri seti için ise sınıflandırma performansının daha düşük olduğu sonucuna ulaşılabilir. Bunun önemli bir sebebi de elde edilen ikinci veri setindeki ranj daralmasıdır.
- Model performanslarının her iki veri seti için bazı kriterlere göre birbirinden ayrıldığı görülmüştür. Buna göre tüm veri setinde genel sınıflandırma doğruluğu en yüksek model C5.0'dır. Ayrıca C5.0 modeli duyarlılık, F1 puanı, kesinlik, negatif öngörü değeri ve Mathews korelasyon katsayısında da en yüksek değere sahiptir. En yüksek hassasiyet C&RT modeli; en yüksek AUC değeri de Yapay Sinir Ağları modeli ile elde edildiği için bu üç modelin okul terkinin belirlenmesinde kullanılabileceği sonucuna ulaşılmış ve terkte etkili olan faktörler bu modeller aracılığı ile ele alınmıştır.
- AGNO>0 veri setinde ise C5.0 genel sınıflandırma doğruluğu ve MCC değeri en yüksek model iken; Duyarlılık, F1puanı, negatif öngörü değeri RA modeli; hassasiyet ve kesinlik CHAID modeli ve AUC değeri YSA modeli ile en yüksek değeri almıştır. Buna bağlı olarak da ağırlıklı genel not ortalaması sıfırdan farklı olan (AGNO>0) öğrencilerin okul terkini belirleyen nitelikler bu dört model aracılığı ile incelenmiştir.

- Modellerin bazı performans değerlerinde istisnalar olsa da genel olarak okul terki tahmininde tüm veri seti ile çalışılmasının daha yüksek doğruluk oranıyla tahmin değerleri verdiğini söyleyebiliriz. Ancak hassasiyet ve duyarlılık arasında dengeyi gözeten kriterlerden MCC ve F1 puanı her iki veri seti için de modellerde yakın değerler almıştır.
- Duyarlılık değeri modellerin terk eden öğrenciyi belirlemedeki başarısını göstermektedir. Dolayısı ile tüm veri seti için en yüksek duyarlılık oranı 0,795 ile C5.0 modelinden elde edilmiş iken AGNO>0 veri seti için 0,767 ile RA modelinden elde edilmiştir. Ayrıca AGNO>0 veri setinde RA modeli dışında diğer tüm modellerin duyarlılık oranı belirgin şekilde düşmektedir. Sonuç olarak terk eden öğrencilerin belirlenmesinde tüm veri setinin kullanılması ile C5.0 modelinden elde edilen sonuçların geçerli en yüksek perfromasnı gösterdiği görülmüştür.

Kesinlik, model tarafından doğru olarak sınıflandırılmış pozitif örnek sayısının, pozitif olarak sınıflandırılmış toplam örnek sayısına oranını ifade ettiği için okul terkinde yüksek kesinlik değerine sahip modellerin tercih edilmesi özellikle terk edebilecek öğrencilere yönelik faaliyetler geliştirecek kurumların maliyet riskini azaltacaktır.

- Tüm veri setinde yapılan analizlerde hem en yüksek doğruluk ve duyarlılık hem de en yüksek kesinlik değeri 0,784 ile C5.0 modelin ait olduğu için özellikle terk eden öğrencilerin öngörülmesinde ile C5.0 modelinin yeterli geçerliğe sahip olduğu sonucuna varılmıştır.
- Tüm veri için yapılan analizlerle kıyaslandığında AGNO>0 veri setinde yapılan sınıflandırma öngörüsünde modellerin hassasiyet değerlerinin yükseldiği görülmektedir. Hassasiyet, modelin başarısının bir göstergesi olduğundan AGNO>0 veri setinde devam eden öğrenci öngörüsü için modellerin daha başarılı sonuçlar ürettiği söylenebilir.
- C&RT modelinin genel sınıflandırma doğruluğu %778 iken %790 ile diğer modellere göre en yüksek hassasiyete sahip olduğu bilindiğine göre modelin devam eden öğrencilerin tahmininde daha etkili olduğu söylenebilir. C&RT Modelinin ağaç yapısı da diğer modellere göre daha yalın ve anlaşılırdır.
- Her iki veri setinde yapılan okul terki sınıflandırması ve üçüncü olarak öğrenci başarısını sınıflandırılması için veri madenciliği modellerinde C5.0 modelinin başarılı olduğu görülmüştür. C5.0 algoritması karar ağacı ile yapılan sınıflandırma çalışmalarında sıkça kullanılan ve oldukça başarılı sonuçlar veren algoritmalardandır.

- Yapay sinir ağlarının klasik doğrusal modelleme anlayışının dışına çıkıp tüm etkenleri ve etkileşimlerini göz önüne alarak yaptığı sınıflandırma pek çok uygulamada başarılı bulunmaktadır. Bu araştırmada YSA ise öğrencilerin terk ve devam durumlarına göre sınıflandırılmasında 0,867 AUC değeri ile en yüksek performansı gösteren modeldir. Modelde AUC değerinin 1'e yakın olması, modelin pozitif (öğrencinin terk edeceği) ve negatif (öğrencinin devam edeceği) sınıflar arasında iyi bir ayrım yapabildiğini; 0,867 gibi yüksek bir AUC değeri, modelin güvenilir tahminlerde bulunduğunu ve karar sınırının doğru bir biçimde belirlendiğini gösterir.

Alanda çok sık kullanılan veri madenciliği sınıflandırma algoritmalarından karar ağaçları ve yapay sinir ağları (Akgün & Özek, 2020) bu çalışmada da başarılı sonuçlar ortaya koymuştur. Yine Pradeep vd. (2015) de bu çalışmanın sonuçları ile tutarlı şekilde karar ağacı gibi sınıflandırma algoritmalarının bir öğrencinin başarısızlığını veya okul terkini tahmin etmek için kullanılabilecek en başarılı yöntemler olduğunu göstermiştir. Quadri ve Kalyankar (2010) da karar ağaçlarının büyük/küçük veri setleri için uygun; kümelerdeki dengesiz veri yapısını ortaya çıkarmak için başarılı ve verimli bir keşif yöntemi olduğunu belirtmiştir. Ağaç tabanlı veri madenciliği yöntemleri, yüksek boyutlu verilerin, yüksek dereceli etkileşimlerin ve tahmin ediciler arasındaki korelasyonların varlığında kullanılmak için çok uygundur. Bu sebeple Behr vd. (2020) de özellikle okulu bırakma tahmini için karar ağaçlarının uygun olduğunu belirtmektedir. Guarín vd. (2015) de Bayes ve karar ağacı kullandıkları çalışmalarında Bayes'in her zaman doğru sonuçlar vermeyebileceğini, karar ağacı sonuçlarının öğrenci performans ve terk tahmininde daha tutarlı ve yeni veriler üzerinde test edilirken daha güvenilir olduğunu belirtmiştir. Ayrıca Yulia (2019) da araştırmasında temel veriler üzerine akademik kayıtlardan elde edilen veriler eklendiğinde, okulu bırakan öğrenci tahmininin iyileştiğini göstermektedir. Kayıt -kabul süreci bilgileriyle, temel veri setinde Naive Bayes algoritmasını iyi sonuç verdiği; eklenen akademik verilerle de karar ağaçlarının %7 oranında bir iyileşme sağladığı bulunmuştur.

Bu araştırmanın sonuçları ile benzer şekilde Yakut (2012) çalışmasında, YSA yönteminin diğer modeller ile karşılaştırıldığında (C5.0 ve DVM) genel anlamda daha iyi sonuçlar verdiğini bulnmuş ancak model karşılaştırmalarında sınıfandırmanın amacına uygunluğu ve diğer kriterler göz önüne alınarak daha anlaşılır karar ağacı kuralları ürettiği için C5.0 algoritmasının kullanılmasını tavsiye etmiştir. Özellikle C5.0 algoritmasında karar ağacı yapısının kolay yorumlanabilmesi bunda önemli bir etkendir. Dutt ve arakadaşları (2017) da e-öğrenmede

öğrenci terki ve mezun öğrencilerin not ortalamalarının tahmini için yine bir karar ağacı olan C5.0 algoritmasının alanyazında çok kullanıldığını göstermiştir.

Bununla birlikte okulu bırakma tahmini için veri madenciliği ve denetimli makine öğrenimi tekniklerini araştıran başka çalışmalarda LR ve DVM algoritmalarının daha başarılı sonuçlar ürettiğini belirten (Lottering vd., 2020; Tamada vd., 2019); Naive Bayes yöntemini uygun gören (Yukselturk, Ozekes & Türel, 2014) ve küçültülmüş veri setinde ise rasgele orman sınıflandırıcısını öneren (Adnan vd., 2021; Lottering vd., 2020) çalışmalar da mevcuttur.

3.Tüm veri seti ve AGNO>0 veri setinde okul terki tahmininde etkili olan değişkenlerle ilgili aşağıdaki sonuçlara ulaşılmıştır.

- Tüm veri setinde incelenen her üç modelin de okul terki için sınıflandırma yaparken ilk iki sırada ağırlıklı genel not ortalamasını ve öğrenci kayıt yılı değişkenini kullandığı görülmektedir. Daha sonra değişkenler farklılaşmakta; C&RT, program düzeyi ve yaş değişkenini daha etkili bir değişken olarak almışken; C5.0, yaş ve program türünü; YSA ise program türü ve geliş şeklini sırasıyla önemli değişkenler olarak değerlendirmişlerdir. Ayrıca her üç modelde de sınıflandırma için öğrenim tipi, ikamet ili ve cinsiyet etkili bulunsa da modele etki düzeyi en düşük değişkenlerdendir. C&RT program türü değişkenini sınıflandırmada hiç kullanmamıştır. Bunun aksine AGNO>0 veri setinde program türü değişkeni daha öncelikli bir faktör olarak çıkmıştır.
- C5. 0 algoritması ile kurulan okul terki sınıflandırma modelinde en önemli değişkenin ağırlıklı genel not ortalaması olduğu; bunu sırasıyla öğrenci giriş yılı, yaş, program türü, ikametgâh, öğrenim tipi, cinsiyet, kayıt/geliş şekli ve program düzeyi değişkenlerinin takip ettiği görülmüştür. Program düzeyi ve geliş şekli değişkenlerinin ise C5.0 sınıflandırma algortimasında öğrencilerin terk veya devam durumlarının tahmininde bir etkisi gözlenmemiştir.
- Modellerin sınıflandırma performanslarında genel doğruluk oranlarının birbirine çok yakın olduğunu da göz önüne alırsak okul terkinin belirlenmesinde en önemli faktörün ağırlıklı genel not ortalaması olduğu; öğrenci kayıt yılı, program türü ve düzeyi ile yaş faktörünün önemli diğer nitelikler olduğu görülmektedir. Bunu yanında ikamet ili de okul terkinde etkili bir faktör olarak ortaya çıkmıştır.
- AGNO>0 veri setinde sınıflandırma yapan modellerin farklı değişkenleri önceliklendirdiği görülmektedir. RA modeli sırasıyla program türü, ikamet ili ve yaş faktörlerini ilk üç sırada tahmin değişkeni olarak ele alırken; C5.0, CHAID ve YSA

modellerinde sıralama değişse de önemli olan ilk üç faktör, AGNO, öğrenci giriş yılı ve program türü değişkenleridir. C5.0 modelinde geliş şekli ve program düzeyi faktörlerinin sınıflandırmada etkisi "0" bulunmuşken; RA ve CHAID modellerinde öğrenim tipi değişkeni hiç yer almamaktadır. Tüm tahmin değişkenlerini model için etkili bulan YSA modelinde ise okul terki sınıflandırmasında en düşük etkiye sahip faktörler öğrenim tipi, ikamet ili ve cinsiyet faktörleridir.

- Tüm bunların yanında tüm veri seti için oluşturulan VM modellerinde etkili olan faktörler ile AGNO> 0 veri seti için oluşturulan modellerdeki faktörler karşılaştırıldığında AGNO değişkeninin etki katsayısının düştüğü; hatta AGNO>0 veri setinde genel sınıflandırma doğruluğu en yüksek model olan RA'da AGNO değişkenin ağacın ilk düğümlerinin oluşmasında etkili olmadığı (ancak 5. sırada yer aldığı) görülecektir. Bunun yerine AGNO>0 veri setinde sınıflandırma için yine bir akademik faktör olan program türünün önemli olduğu görülmektedir.

Bu çalışmanın sonuçları ile benzer şekilde Lottering vd.'nin (2020) aktardığına göre Steward, Lim ve Kim üniversite akademik performansı ve cinsiyet dahil olmak üzere öğrenci demografik özelliklerini; Srilekshmi vd. ise öğrenci kişisel bilgilerinin yanı sıra bazı akademik kayıtları içeren öznitelikleri okul terkinde etkili bulmuştur. Çevrimiçi lisans programlarında kayıtlı yetişkinlerin okul terki faktörlerini araştıran Choi ve Park (2017) da yaş, cinsiyet, eğitim seviyesi gibi kişişsel değişkenler ile genel not ortalaması değişkenlerini; Grau-Valldosera, J. (2019) da özellikle yaş ve önceki üniversite deneyiminin okul terkinde önemli olduğunu göstermiştir. Appavoo (2023) da bu çalışmanın sonuçlarına benzer şekilde olgun öğrencilerin üniversiteyi bırakma olasılıklarının daha düşük olduğunu ortaya koymakta; aksine Araque vd. (2009) ise programa başlangıç yaşı arttıkça okulu bırakma olasılığının arttığını söylemektedir. Yine bu çalışmanın sonuçları ile tutarlı olarak cinsiyet faktörünün okul terki ile ilişkili olduğunu gösteren pek çok araştırma (Appavoo, 2023; Guzman vd., 2021; Günther & Koeszegi, 2012; Tarimo, 2013) mevcuttur. Tarimo (2013), erkek öğrencilerin okulu bırakma oranının daha yüksek olduğunu, benzer şekilde Appavoo (2023) Mauritius Açık Üniversitesi'nde kadın öğrencilerin devam etmede daha ısrarcı olduklarını bildirmiştir.

Başka bir çalışmada ise cinsiyetin diğer demografik değişkenlerin etkisi kontrol altına alındığında okul terkinde önemli bir faktör olduğu ancak yukardaki çalışmaların aksine bırakma riskini tahmin eden modellerde kadınların erkek öğrencilere göre önemli ölçüde daha yüksek oranda okulu bırakma riskiyle karşılaştığı sonucuna ulaşılmıştır (Günther & Koeszegi, 2012). Yine Guzman vd. (2021) de özellikle kırsalda kadınların yükseköğretimi bırakma riskinin

yüksek olduğunu bildirmektedir. Tüm bunların yanında Aguş (2019) ise yaş ve cinsiyet faktörlerinin okulu bırakmada etkili olmadığını; Pierrakeas vd. (2004) de okul terkinde cinsiyetler arasında bir fark oluşmadığını belirtmiştir. Quadri ve Kalyankar (2010) da karar ağacı modelinde diğer değişkenlerin yanında, cinsiyeti terk tahmininde ilgisiz faktörlerden görmüştür.

Bu çalışmanın sonuçları ile uyumlu olarak bir diğer değişken olan program türünün okul terkinde önemli bir faktör olduğu araştırmalarda ortaya konulmuştur (Altıntaş vd., 2021; Araque vd., 2009; Lottering vd., 2020). Benzer şekilde Zawacki-Richter & Qayyum (2019) da program türlerini okul terki için önemli etken olarak değerlendirmiştir. Heublein (2014) Almanya yükseköğretiminde yaptığı okul terki araştırmasında diğer değişkenlerin yanında program türlerine göre okul terki riskinin değiştiğini; Matematik, doğa bilimleri ve işletme gibi alanlarda yüksek okul terki ile karşılaşıldığını göstermiştir. Gregori vd. (2018) de uzaktan eğitimde 2010 yılı analizlerine dayanarak bir dizi lisans programında okul terki riskini incelemiş ve ister Fen bilimleri, ister Beşeri bilimler olsun tümünde oranların yüksek olduğunu ve alanlara göre bu oranların farklılaştığını ortaya koymuştur. Kore Ulusal Açık Üniversitesi'nde okul terki sorununa dikkat çeken çalışmalar yürüten Choi vd. (2013) öğrencilerin diplomanın değerine ilişkin algılarının okulu bırakma olasılığını belirlediğini ortaya koymaktadır. Bu da program türünün başka bir nedenle okul terkindeki etkisini ortaya koymaktadır. Dolayısıyla çeşitli araştırmalar, öğrencilerin okulu bırakma oranının önemli faktörlerinden birinin üniversitede okunan bölüm olduğunu göstermektedir.

Appavoo (2023) Mauritius Açık Üniversitesi'ndeki (OU) okul terki araştırmasında; temel, lisans ve yüksek lisans düzeylerinde öğrenimi bırakma oranlarını sırasıyla %46,08, %39,14 ve %17,31 bulmuş ve daha önce yükseköğrenim görmüş olanların öğrenimi bırakma olasılığının bu çalışmayla da benzer şekilde daha düşük olduğunu göstermiştir. Araque ve diğerleri de (2009) program derecesinin üniversite terkinde önemli bir etken olduğunu (ancak bu çalışmanın aksine temel düzeyde eğitim alanların lisans düzeyine göre 119 kat daha fazla okul terkettiğini) bulmuştur.

Guzman vd.'nin (2021) Güney Afrika'da yürüttükleri araştırmanın sonucuna göre üniversite genel not ortalaması ne kadar yüksekse okulu bırakma olasılığı da o kadar düşüktür. Bu araştırmanın sonucuna da uygun şekilde akademik not okul terkinin en önemli belirleyicisidir. Araque vd. (2009) de akademik performans, başarı ve öğrencilerin ortalama notu gibi değişkenlerin okulu bırakma ile çok ilişkili olduğunu fark etmiştir ve araştırmacılara göre düşük

akademik performans ve başarı seviyesi okulu bırakma olasılığı ile yüksek bağlantılıdır. Yulia (2019) ise akademik veriler eklendikçe sınıflandırıcıların doğruluk oranlarının arttığını; sadece kayıt kabul verilerinin okul terki tahminlerinde bulunmak için yetersiz olduğunu göstermiştir. Özellikle erken okul terki tahminlerinde akademik notların önemli olduğunu ancak çok daha fazla akademik faktör eklemenin de sınıflandırıcıyı geliştirmediğini belirtmektedir. Ribeiro & Canedo (2020) öğrencinin akademik performansı ve lisans dersinin zorluk derecesi gibi faktörlerin hala okulu bırakmanın ana faktörleri olduğunu, ancak sosyal konuların da bir lisans dersinin tamamlanması veya bırakılmasıyla ilgili olduğu sonucuna varmıştır. Ortiz-Lozano vd. (2023) İspanya Açık Üniversitesi'nde okul terkini tahmin etmek için yürüttükleri çalışmalarında üç geniş değişken kategorisinden (kişisel, kurumsal ve hedef başarısı) ikincisi, özellikle de ilk dönemdeki akademik sonuçların, okul terki için en fazla tahmin kapasitesine sahip değişken olduğunu ve bunun yanına demografik değişkenlerin eklenmesinin sadece tahmin başarısını arttırdığını bulmuştur. Adnan vd. (2021) de aşamalı modeller ile yürüttükleri çalışmalarında demografik değişkenlerin (yaş, bölge, cinsiyet vb.) veri madenciliği modelleri ile ilk aşamada okul terkinin belirlenmesinde bir önemli etkisinin olmadığını belirtmekte; ilerleyen deney aşamalarında akademik değişkenlerin eklenmesiyle okul terkinin optimum bir zamanda daha yüksek doğruluk ile belirleneceğini göstermektedir. Çok sayıda değişken üzerinde çalışılmış olsa da, akademik performans sürekli olarak okul terkinin güçlü bir yordayıcısı olarak ortaya çıkmaktadır (Kovacic, 2010). Brezilya'da yükseköğretimde okuldan kaçışla ilgili araştırmalar yapan Pinheiro vd. (2023) akademik başarısızlık ile okul terkini eş anlamlı gören çalışmalara dikkat çekmekte ve bu araştırmanın da bulgularında vurgulandığı gibi akademik başarısızlığın önemli bir okul terk faktörü olduğunu ortaya koymaktadır.

4. Öğrencilerin ağırlıklı genel not ortalamasına göre başarı durumlarının sınıflandırılmasında başarılı olan veri madenciliği modelleri ile etkili faktörlere ilişkin sonuçlar aşağıdaki gibidir.

- Öğrenci başarısının sınıflandırılması için en yüksek genel doğruluk oranı C5.0 ve DVM modellerinden (%69,6 doğru sınıflandırma); en yüksek duyarlılık C&RT modelinde (%85,8) ortaya çıkmıştır. En yüksek hasasiyet değeri ise %73 ile RA modeline aittir. Genel olarak tüm modellerde Mathews korelasyon katsaysının düşük olduğu ve hasasiyet ile duyarlılık değerleri arasında farklar oluştuğu görülmektedir. Dolayısı ile bazı modeller pozitif (yani başarılı) sınıfın tahmininde gösterdiği başarıyı negatif sınıfın belirlenmesinde gösteremeyebilir. Bu sebeple çalışmanın amacına uygun olarak en yüksek doğruluk oranına sahip C5.0 modeli, en yüksek duyarlılık ve negatif öngörü değerine sahip C&RT modeli ile en yüksek hassasiyet ve MCC değerine sahip RA

modellerinin istenilen sonuçları vereceği düşünüldüğünden bu modellerin yapısı ve öğrenci başarının öngörülmesinde etkili değişkenlerin neler olduğu ayrıntılı ele alınmıştır.

- Öğrenci başarısının sınıflandırılmasında başarılı olan VM modellerinde, etkili olan faktörlerin modellere göre sınıflandırmada etkisinin değiştiği görülmektedir. C&RT algoritması nerdeyse modeldeki tüm değişkenleri sınıflandırmada kullanmışken; C5.0 algoritması sadece dört faktörün sınıflandırma için etkili olduğu sonucuna ulaşmış ve modelde geliş şekline, program düzeyine, program türü ve tipine yer vermemiştir. Geliş şekli ve program düzeyi diğer iki model de düşük etkiye sahiptir. Giriş yılı öğrenci başarısının sınıflandırılmasında C5.0 ve C&RT modelinde en önemli etken olarak ortaya çıkmıştır. Yaş faktörünün ise her üç modelde de etkisi yüksek faktörlerden olduğu söylenebilir. Bunun yanında giriş yılı, ikamet adresi ve cinsiyet faktörleri de değişen oranlarda akademik başarının sınıflandırılmasında etkili bulunmuştur.
- Öğrencilerin akademik başarıya göre sınıflandırılmasında C5.0 modeli içinde öğrenci giriş yılının en yüksek etkiye sahip olduğu; bunu sırasıyla yaş, ikamet ili, cinsiyet, öğrenim tipi, program, program düzeyi, kayıt/geliş şekli faktörlerinin takip ettiği görülmektedir. Başarılı-başarısız öğrencilerin sınıflandırılmasında öğrenim tipi değişkeni diğerlerine göre oldukça düşük etkiye sahipken; program düzeyi ve geliş şekli değişkenleri modelde hiçbir etkiye sahip değildir.
- C&RT modelinde akademik başarının tahmininde en etkili değişken öğrenci giriş yılıdır. Bunu sırasıyla ikamet ili, yaş, geliş/kayıt şekli, program, cinsiyet, program düzeyi ve öğrenim tipi (açık-uzaktan) faktörleri izlemektedir. Tahminde "öğrenci giriş yılı" faktörünün diğerlerine göre oldukça yüksek oranda etkili olduğu görülecektir. Ayrıca C&RT Modeli C5.0'da yer alan geniş ve karmaşık ağaç yapısına karşın (ki 1161 düğüm oluşmaktadır) oldukça basit (5 düzeyli) 22 düğümden oluşan bir model ortaya çıkarmıştır.
- Öğrenci başarısının tahmininde RA modelinin tahmin değişkenleri ve önemi incelendiğinde öğrencilerin başarılı-başarısız olarak sınıflandırılmasında sırasıyla, program, ikamet ili, yaş, öğrenci giriş yılı, cinsiyet, geliş şekli ve program düzeyi değişkenleri etkili olmaktadır. Rastgele ağaçlar modeli ayrıca en sık kullanılan kuralları içerecek en yüksek doğruluk oranına sahip kurallar dizisi oluşturmuştur.

Eğitsel veri madenciliği ile yapılan çalışmaların ana odağının öğrenci başarısı ile akademik performansın kestirimine yoğunlaştığı ve sonuç olarak öğrencilerin başarı durumlarına göre

sınıflandırılmasında bu yöntemler ile verimli sonuçlar elde edildiği görülmektedir (Akgün & Özek, 2020; Bonde & Kirange, 2018; Hotaman, 2020). Kovacic (2010) öğrenci başarısının erken tahmini için farklı sınıflandırma ağacı yöntemlerini (CHAID, QUEST ve CART) kullanarak Yeni Zelanda Açık Politeknik'teki öğrencilerin başarısını etkileyebilecek sosyo-demografik değişkenleri (yaş, cinsiyet, etnik köken, eğitim, çalışma durumu ve engellilik) ve çalışma ortamını araştırmıştır. Başarılı -başarısız öğrencileri ayıran en önemli faktörlerin etnik köken ve kurs programı olduğu bulunmuş; en başarılı sınıflandırma yönteminin ise bu çalışmanın sonucuna benzer şekilde CART olduğu sonucuna ulaşılmıştır. Sajadin Sembiring vd.'nin (2011) Malezya UMP'de yaptığı araştırmanın sonucu, veri madenciliği tekniklerinin öğrenci performansı için etkili iyileştirme araçları sağladığını göstermektedir. Veri madenciliğinin yükseköğretimde, özellikle de öğrencinin nihai performansını tahmin etmede ne kadar faydalı olabileceğinin gösterildiği çalışmada özellikle, karar ağacı kullanarak kurulan tahmin modeli başarılı bulunmuştur.

S. Tosun (2007), demografik, eğitim, akademik ve kişisel bilgilerden oluşan değişkenlerle üniversite öğrencilerinin not ortalamalarına göre sınıflandırılması amacıyla yaptığı araştırmasında yapay sinir ağları ve karar ağaçları yöntemlerini karşılaştırmıştır. Karar ağaçları ile oluşturulan modelde %86'lık bir doğruluk oranı elde ederken, aynı veri ile yapay sinir ağlarında %91,77'lik bir doğruluk oranı elde etmiştir. Bu çalışmanın aksine öğrencilerin başarı puanlarına göre sınıflandırılmasında yapay sinir ağının karar ağaçlarına göre daha iyi performans gösterdiğini ifade etmiştir. Ancak üniversite öğrencilerinin başarısını etkileyen faktörler arasında, bu çalışmanın sonuçları ile örtüşen şekilde, eğitim düzeyi ve ikamet yeri gibi değişkenlerin etkisini ortaya koymuştur. Benzer şekilde Şengür ve Tekin (2013) de öğrencilerinin mezuniyet notlarının tahmin edilmesi için yapılan benzetim çalışmalarında YSA'nın, karar ağaçlarına oranla daha iyi tahmin başarımı sağladığını göstermiştir. Eğitsel veri madenciliğinin öğrenci başarısının kestiriminde kullanımını araştıran Ekinci (2022) yapılan analiz sonucunda öğrenci mezuniyet durumu lojistik regresyon modelinin %76,80 ile yüksek düzeyde performans gösterdiği gözlenmiştir. Altun (2019) ise öğrencilerin akademik performansını kestirmek için yaptığı çalışmada alt modellerin her biri için yapay sinir ağları ve çoklu doğrusal regresyon analizi ile farklı modeller geliştirmiş ve performansları karşılaştırmıştır. Geliştirilen modellerin öğrencilerin mezuniyet notunu 1. sınıf 1. dönemden itibaren %94-97 doğrulukla kestirdiği görülmüştür.

Talan (2020), uzaktan eğitim öğrencilerinin akademik başarı durumlarını çeşitli değişkenler açısından incelediği nicel desendeki çalışmasında; uzaktan eğitim öğrencilerinin akademik

başarı durumlarının yaş, cinsiyet ve bilgisayar kullanım süresi değişkenlerine göre istatistiksel olarak farklılaşmadığı sonucuna ulaşılmıştır. Taşdemir(2012) ise öğrenci başarısına etki eden faktörleri regresyon analizi ile incelediği tezinde öğrencinin cinsiyeti ile not ortalaması ve kaldığı ders sayısı arasında pozitif yönlü güçlü bir ilişki bulmuştur. Üniversiteye giriş yaşı ile not ortalaması arasında ise negatif yönlü güçlü bir ilişki olduğunu; yaş arttıkça kaldığı ders sayısının azaldığını belirtmiştir.

5.2. Öneriler

Yükseköğretim kurumları, ailelere ve devletlere olan yüksek ekonomik maliyeti nedeniyle okulu bırakma oranlarını azaltmaya; işgücü piyasasında akademik olarak nitelikli kişilerin sayısını artırmaya ve yanlış yatırımı azaltmaya çalışmaktadır (Bargmann, Thiele & Kauffeld, 2022). Bu çalışmada yer alan metrikler, veri madenciliği modellerinin eğitimde öğrenci terk ve devam durumlarını sınıflandırmada oldukça yetenekli olduğunu, ancak bazı alanlarda iyileştirmeler yapılabileceğini ortaya koymaktadır. Bu tür analizlerin yapılması, eğitim kurumlarının öğrenci davranışlarını daha etkili bir biçimde yönetmelerine ve müdahale stratejilerini optimize etmelerine olanak tanır ve genel performansını artırabilir.

Araştırmada kayıt aşamasında öğrencilerden elde edilen veriler ile 1. dönemin sonunda ortaya çıkan ağırlıklı genel ortalaması değişkenlerine bağlı olarak yapılan sınıflandırmayla erken dönemde ve doğal yollar ile elde edilmiş veriler sonucunda okul terkinin tahmininde başarılı olan VM modelleri ele alınmıştır. Bu veriler ile yola çıkmanın bazı avantajları olduğu gibi dezavantajları da mevcuttur. Dolayısı ile araştırmacıların özellikle akademik değişkenlerin ve psikolojik değişkenlerin okul terkindeki rolü dikkate alındığında başka değişkenler ile amaca uygun değişen zaman dilimlerinde okul terkinin açık ve uzaktan öğrenmede belirlenmesine yönelik VM modellerini çalışması önerilebilir. Böylelikle süreç içinde artan öğrenci verisi ve diğer değişkenlerin okul terki ile anlamlı bir fark yaratıp yaratmadığı gözlenebilir.

Bu çalışmada ve okul terki literatüründe program türlerinin terk için önemli bir faktör olduğu görülmektedir. Dolayısıyla bu çalışmanın çıktılarına uygun şekilde okul terkinin belirlenmesinde başarılı olan VM algoritmaları ile yükseköğretim programlarına özel terk araştırmalarının yapılması önerilmektedir.

Okul terkine yönelik K12 düzeyinde farklı boyutların ele alındığı nitel ve nicel desende çalışmalar bulunsa da erken dönemde okul terkini öngören veri madenciliği araştırmalarına

rastlanmamaktadır. Dolayısı ile benzer şekilde bir çalışmanın orta öğretim düzeyinde okul terkini belirleyecek şekilde veri madenciliği yöntemi ile çalışılması önerilmektedir.

Eğitimde veri madenciliğinde, bu çalışmadaki yordanan değişkenlerde de olduğu gibi çoğunlukla ikili sınıflama modellerinin performanslarının karşılaştırıldığı görülmektedir. Sonraki araştırmalarda, özellikle akademik başarıyı çoklu düzeyde ele alarak, veri madenciliği sınıflandırma modellerinin performanslarının karşılaştırılması ve dolayısı ile 2 kategoride olmayan öğrenci başarısı tahminleri için uygun modellerin araştırılması önerilmektedir.

Araştırmada özellikle donanım yetersizliği sebebi ile DVM algoritmasında verimli sonuçların alınmadığı görülmüştür. Araştırmacıların buna dikkate etmesi önerilmektedir.

Veri Madenciliği, eğitimci, öğrenci, mezunlar ve eğitim süreçleri hakkında daha fazla bilgi vermek ve eğitim faaliyetleri hakkında daha iyi kararlar alınmasına yardımcı olmak için uygun bir teknoloji olarak kabul edilir. Eğitim sistemindeki derin ve kapsamlı bilgi eksikliği, sistem yönetiminin kalite hedeflerine ulaşmasını engelleyebilmektedir. Veri madenciliği teknolojisi, yüksek eğitim sistemindeki bu bilgi açığının kapatılmasına yardımcı olabilir. Sonuç olarak, bu tür araştırmalar yüksek eğitim sistemi verimliliğini en üst düzeye çıkarmak, öğrencilerin okulu bırakma oranını azaltmak, öğrencilerin başarısını ve okulda kalma oranını artırmak ve sistem süreçlerinin maliyetini azaltmak gibi birçok avantaj getirebilir. Dolayısı ile benzer çalışmaların kurumların süreçlerine entegre edilmesi önerilir.

Eğitim kurumlarının bu tür VM çalışmalarını sürekli hale getirirken veri kaynaklarından altyapı imkânlarına kadar bir dizi değişkeni gözden geçirmesi gerekebilir. Bunun yanında kurumların bu çalışmalar için açık veri sunması önerilmektedir. Böylelikle karşılaştırmalı çalışmalar da yapılabilir.

Veri madenciliği çalışmalarının temel kaynağı çoğunlukla kurumların veri ambarları mimarisine dayanır. Dolayısıyla kurumlar kayıt ve eğitim verilerini tutarken bunu göz önünde bulundurmalıdır. Ayrıca bu durum eğitsel veri madenciliği çalışmalarında karşılaşılan sorunların kurum özelinde çözülmesini gerektirir Bu sebeple benzer çalışmaların farklı veri yapılarına sahip diğer açıköğretim üniversitelerinde de tekrarlanması önerilmektedir.

Üniversitelerin, maliyetleri ve çabaları azaltarak okulu bırakma oranını düşürmeye yardımcı olacak belirli bir eylem planı hazırlamak için benzer modelleri göz önünde bulundurabileceği düşünülmektedir. Bununla birlikte farklı zamanlarda, program türlerine ve üniversitelerin genel veri yapısına uygun yapılan; okul terki ve başarının erken öngörülmesine yönelik bu tür

çalışmalar öğrenmede süreç değerlendirme olarak ele alınıp öğrenci başarının arttırılması için ön bir adım olacağından akademisyenlerin bu tür çalışmaları planlı değerlendirme süreçlerine entegre etmesi önerilmektedir.

Başka yerlerde ve daha yakın zamanlarda, araştırmacılar okulu bırakmayı açıklamak için "Bordieusçu" bir mercek kullanma eğiliminde olmuştur. Ancak bu yaklaşımların ortak noktası, öğrencileri 'düzeltmeye' odaklanmaları ve öğrenci başarısı veya devamlılığı için yine öğrencilerin sosyal sermayelerini artıracak, akademinin dilini ve oyunun kurallarını öğrenecek bir araca sahip olmaları gerektiğini savunmalarıdır. Yalnızca yüzyılın başından bu yana ve yalnızca sektörün bazı kesimlerinde, üniversitelerin bugünün öğrencilerinin ve onların gelecekteki yaşamlarının ve dünyalarının ihtiyaçlarını karşılamak için kendilerinin de değişmesi gerekebileceğinin farkına varılmıştır. Bu çalışmanın alana bu yönde bir katkı sunması beklenmektedir.

KAYNAKLAR

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *Ieee Access, 9*, 7519-7539. doi:10.1109/ACCESS.2021.3049446

Agrusti, F., Bonavolonta, G., & Mezzini, M. (2019). University dropout prediction through educational data mining techniques: a systematic review. *Journal of e-learning and Knowledge Society, 15*, 161-182.

Aguş, F. S. (2019). *Açık ve uzaktan eğitimde öğrencilerin okulu bırakma nedenleri: Anadolu Üniversitesi açıköğretim sistemi üzerine bir araştırma.* (Yüksek lisans tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data mining: a prediction for student's performance using classification method. *World Journal of Computer Application and Technology, 2*(2), 43-47. doi:10.13189/wjcat.2014.020203

Akbaş, G., & Apar, A. (2010). *Avrupa 2020 stratejisi: akıllı, sürdürülebilir ve kapsayıcı büyüme için avrupa stratejisi özet bilgi notu*. Ankara.

Akdemir, H. T. (2021). *TIMSS 2019 Türkiye örneklemine göre 4. ve 8. sınıf öğrencilerin matematik başarı durumlarının çeşitli değişkenlere göre CHAID analizi ile incelenmesi.* (Yüksek lisans tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Akgün, K., & Özek, M. B. (2020). Eğitsel veri madenciliği yöntemi ile ilgili yapılmış çalışmaların incelenmesi: içerik analizi. *Uluslararası Eğitim Bilim ve Teknoloji Dergisi*. doi:10.47714/uebt.753526

Akküçük, U. (2011). *Veri madenciliği kümeleme ve sınıflama algoritmaları*. İstanbul: Yalın

Akpınar, H. (2017). *DATA veri madenciliği veri analizi*. İstanbul: Papatya

Aksu, G. (2018). *PISA başarısını tahmin etmede kullanılan veri madenciliği yöntemlerinin incelenmesi.* (Doktora tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Aksu, G., & Doğan, N. (2018). Veri madenciliğinde kullanılan öğrenme yöntemlerinin farklı koşullar altında karşılaştırılması. *Ankara Universitesi Egitim Bilimleri Fakultesi Dergisi*. doi:10.30964/auebfd.464262

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences, 12*(16), 4102-4107.

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences, 11*(9), 552. doi:10.3390/educsci11090552

Altıntaş, Ö., Başer, F., & Babadoğan, M. C. (2021). *Yükseköğretimde akademik riskli öğrencilerin kestiriminde makine öğrenmesi yöntemleri*. A. Apaydın & Ö. Kutlu (Eds.). Ankara: Pegem Akademi.

Altun, M. (2019). *Öğrenci akademik performansının kestirilmesine ilişkin bir model önerisi: Veri madenciliğine dayalı bir çalışma.* (Doktora tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Appavoo, P., Gungea, M., & Sohoraye, M. (2023). Drop-out among ODL learners: A case study at the Open University of Mauritius. *Journal of Educational Technology and Online Learning, 6*(3), 665-682. doi:10.31681/jetol.1273563

Araque, F., Roldán, C., & Salguero, A. (2009). Factors influencing university drop out rates. *Computers & Education, 53*(3), 563-574. doi:10.1016/j.compedu.2009.03.013

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of College Student Personnel, 25*(4), 297-308.

Atif, M. (2022). *Predict student success and performance factors by analyzing educational data using data mining techniques.* (Yüksek Lisans Tezi). https://bspace.buid.ac.ae/handle/1234/2016 ProQuest Dissertations & Theses Global database. (30663493) sayfasından erişilmiştir.

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2017). Predicting student dropout in higher education. *In: 2016 ICML Workshops #Data4Good Machine Learning*, New York, vol. abs/1606.06364, pp. 16–20

Aydın, S. (2007). *Veri madenciliği ve Anadolu Üniversitesi uzaktan eğitim sisteminde bir uygulama.* (Doktora tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Aydın, S., Öztürk, A., Büyükköse, G. T., Er, F., & Sönmez, H. (2019). An investigation of drop-out in open and distance education. *Educational Sciences: Theory and Practice, 19*(2), 40-57.

Bağrıacık Yılmaz, A., & Karataş, S. (2022). Why do open and distance education students drop out? Views from various stakeholders. *International Journal of Educational Technology in Higher Education, 19*(1). doi:10.1186/s41239-022-00333-x

Ballı, F. E., & Kartal, S. (2020). Okul terki araştırmaları: sistematik bir analiz çalışması. *Kırıkkale Üniversitesi Sosyal Bilimler Dergisi (KÜSBD), 10*(2), 257-278. https://dergipark.org.tr/en/download/article-file/1217520 sayfasından erişilmiştir.

Bargmann, C., Thiele, L., & Kauffeld, S. (2022). Motivation matters: predicting students' career decidedness and intention to drop out after the first year in higher education. *Higher Education, 83*(4), 845-861. doi:10.1007/s10734-021-00707-6

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T., & Popelinsky, L. (2012). Predicting drop-out from social behaviour of students. *International Educational Data Mining Society*.

Bean, J. P. (1990). Using retention research in enrollment management. *The strategic management of college enrollments,* 170-185.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55*(4), 485-540. doi:10.2307/1170245

Behr, A., Giese, M., Teguim Kamdjou, H. D., & Theune, K. (2020). Dropping out of university: a literature review. *Review of Education, 8*(2), 614-652. doi:10.1002/rev3.3202

Berge, Z. L., & Huang, Y. (2004). *A model for sustainable student retention: a holistic perspective on the student dropout problem with special attention to e-learning*.

Bhise, R., Thorat, S. S., & Supekar, A. K. (2013). Importance of data mining in higher education system. *Journal Of Humanities And Social Science (IOSR-JHSS), 6*(6), 18-21. doi:10.9790/0837-0661821

Bilici, Z., & Özdemir, D. (2021). Data mining studies in education: literature review for the years 2014-2020. *Bayburt Eğitim Fakültesi Dergisi*. doi:10.35675/befdergi.849973

Bonde, S. N., & Kirange, D. K. (2018). *Educational data mining survey for predicting student's academic performance.* Paper presented at the Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2018), Cham.

Budiman, R. (2018). Factors related to learners' drop-out of a distance language learning programme. *Journal of Curriculum and Teaching, 7*(2), 12-19. doi:10.5430/jct.v7n2p12

Bülbül, T. (2012). Yükseköğretimde okul terki: nedenler ve çözümler. *Eğitim ve Bilim, 37*(166).

Büyükatak, E. (2022). *PISA 2018 Türkiye örnekleminde okuma okuryazarlık düzeylerinin farklı veri madenciliği sınıflandırma yöntemleri ile incelenmesi.* (Doktora tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Cemaloğlu, N., & Duykuluoğlu, A. (2020). *Sosyal bilimlerde veri madenciliği*. N. Cemaloğlu & A. Duykuluoğlu (Eds.). Ankara: Pegem Akademi.

Ceyhan, G. (2020). *Sınıflandırmada kullanılan veri madenciliği yöntemlerinin performanslarının veri seti özelliklerine göre karşılaştırılması.* (Doktora tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Cho, M. H., & Heron, M. L. (2015). Self-regulated learning: the role of motivation, emotion, and use of learning strategies in students' learning experiences in a self-paced online mathematics course. *Distance Education, 36*. doi:10.1080/01587919.2015.1019963

Choi, H. J., & Park, J.-H. (2017). Testing a path-analytic model of adult dropout in online degree programs. *Computers & Education, 116*. doi:10.1016/j.compedu.2017.09.005

Christie, S. T., Jarratt, D. C., & Olson, L. A. (2019). *Machine-learned school dropout early warning at scale.* Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019), Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (Eds.), pp. 726 - 731. https://files.eric.ed.gov/fulltext/ED599217.pdf sayfasından erişilmiştir.

Çırak, G. (2012). *Yükseköğretimde öğrenci başarılarının sınıflandırılmasında yapay sinir ağları ve lojistik regresyon yöntemlerinin kullanılması.* (Yüksek lisans tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Dai Shangping, Z. P. (2008). *A data mining algorithm in distance learning*. 12th International Conference on Computer Supported Cooperative Work in Design, Xi'an, China. pp. 1014-1017. doi: 10.1109/CSCWD.2008.4537118.

Dalipi, F., Imran, A. S., & Kastrati, Z. (2018). *MOOC dropout prediction using machine learning techniques: Review and research challenges.* Paper presented at the 2018 IEEE Global Engineering Education Conference (EDUCON).

Dekker, G., Pechenizkiy, M., & Vleeshouwers, J. (2009). Predicting students drop out: a case study. *Computers, Environment and Urban Systems. pp* 41-50.

Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. In (pp. 129-140): Springer International.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498-506. doi:10.1016/j.dss.2010.06.003

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory and Practice, 13*, 17-35. doi:10.2190/CS.13.1.b

Demir, M. (2021). *Özellik seçim yöntemleri kullanılarak sınıflandırma algoritmalarının performanslarının karşılaştırılması.* (Yüksek lisans tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Dol, S. M., & Jawandhiya, P. M. (2023). Classification technique and its combination with clustering and association rule mining in educational data mining — A survey. *Engineering Applications of Artificial Intelligence, 122*. doi:10.1016/j.engappai.2023.106071

Dominic Ravichandran, S. D. (2021). *Predicting student dropout in self-paced MOOC Course.* (M.S.). Arizona State University, United States -- Arizona. https://www.proquest.com/dissertations-theses/predicting-student-dropout-self-paced-mooc-course/docview/2532530651/se-2?accountid=11637 ProQuest Dissertations & Theses Global database. (28492111) sayfasından erişilmiştir.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *Ieee Access, 5*, 15991-16005. doi:10.1109/access.2017.2654247

Ekinci, H. (2022). *Eğitsel veri madenciliğinin öğrenci başarısının kestirimine yönelik kullanımı.* (Yüksek lisans tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Elibol, S., & Bozkurt, A. (2023). Student dropout as a never-ending evergreen phenomenon of online distance education. *European Journal of Investigation in Health, Psychology and Education, 13*(5), 906. doi:10.3390/ejihpe13050069

Erdoğan, Ş. Z. (2004). *Veri madenciliği ve veri madenciliğinde kullanılan k-means algortimasının öğrenci veri tabanında uygulanması.* (Yüksek lisans tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Ersöz, F. (2019). *Veri madenciliği teknikleri ve uygulamaları kavram – teori – modeller, yöntem*. Ankara: Seçkin.

Evans, M., & Tragant, E. (2020). Demotivation and dropout in adult EFL learners. The *Electronic Journal for English as a Second Language, 23(4).* https://www.tesl-ej.org/pdf/ej92/a8.pdf sayfasından erişilmiştir.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education*. New York: McGraw-Hill.

Resmi Gazete, T. C. (2020). *Yükseköğretim Kanunu ile Bazı Kanunlarda Değişiklik Yapılmasına Dair Kanun*.

Geray, C. (2007). Distance education in Turke*y* . *International Journal of Educational Policies,* 1(1), 33-62. http://ijep.icpres.org/vol1_1/Geray(1)1_33_62.pdf sayfasından erişilmiştir.

Gil, J. S., Jhone, A., Delima, P., & Vilchez, R. N. (2020). Predicting student's dropout ındicators in public school using data mining approaches. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(1), 774-778. doi:10.30534/ijatcse/2020/110912020

Grau-Valldosera, J. (2019). *A dropout definition for continuance intention and effective re-enrolment models in online distance learning.* (Doktora tezi). https://openaccess.uoc.edu/bitstream/10609/112746/1/2019-12%20Tesi%20jgrauv.pdf sayfasından erişilmiştir.

Grau-Valldosera, J., & Minguillón, J. (2014). Rethinking dropout in online higher education: the case of the Universitat Oberta de Catalunya. *International Review of Research in Open and Distributed Learning, 15*(1), 290-308. doi:10.19173/irrodl.v15i1.1628

Grau-Valldosera, J., Minguillón, J., & Blasco-Moreno, A. (2018). Returning after taking a break in online distance higher education: from intention to effective re-enrollment.

Interactive Learning Environments, 27(3), 307-323. doi:10.1080/10494820.2018.1470986

Gregori, P., Martínez, V., & Moyano-Fernández, J. J. (2018). Basic actions to reduce dropout rates in distance learning. *Evaluation and Program Planning, 66*, 48-52. doi:10.1016/j.evalprogplan.2017.10.004

Guarín, C. E. L., Guzmán, E. L., & González, F. A. (2015). A model to predict low academic performance at a specific enrollment using data mining. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 10*(3), 119-125. doi:10.1109/RITA.2015.2452632

Guzmán, A., Barragán, S., & Cala Vitery, F. (2021). Dropout in rural higher education: a systematic review. *Frontiers in Education, 6*. doi:10.3389/feduc.2021.727833

Gülşen, F. U. (2017). *Yükseköğretimde okul terkinin değerlendirilmesi.* (Doktora tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Günther, E., & Koeszegi, S. T. (2012). *Gender counts?! analysis of student dropout at Vienna University of Technology*. A. B´eraud et al. (eds.), GIEE 2011: Gender and Interdisciplinary Education for Engineers, 439–453.

Hagedorn, L. S. (2005). How to define retention: a new look at an old problem. *College student retention formula for student success*, 90-105.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining : concepts and techniques*. https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques sayfasından erişilmiştir.

Hegde, V., & Prageeth, P. P. (2018, 19-20 Jan. 2018). *Higher education student dropout prediction and analysis through educational data mining.* Paper presented at the 2nd International Conference on Inventive Systems and Control (ICISC).

Heublein, U. (2014). Student drop-out from german higher education institutions. *European Journal of Education, 49*(4), 497-513. doi:10.1111/ejed.12097

Holmberg, B. (2008). *The Evolution, Principles and Practices of Distance Education*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1e494afabaf38b9f4ab47639a84b8806ea7c6871 sayfasından erişilmiştir.

Hotaman, D. (2020). Öğrenci başarısının değerlendirilmesinde eğitsel veri madenciliğinin kullanılması. *Ulakbilge Dergisi, 8*(48), 577–587. doi:10.7816/ulakbilge-08-48-08

IBMCorporation. (2022). *IBM SPSS modeler 18.4 algorithms guide*. Retrieved from https://www.ibm.com/docs/en/SS3RA7_18.4.0/pdf/AlgorithmsGuide.pdf

Kabakchieva, D. (2013). Predicting student performanceby using data mining methods for classification. *Cybernetics and Information Technologies, 13(1):6 1* 72.

Kabalcı, E. (2013). Yapay sinir ağları (artificial neural networks). [PowerPoint sunumu]. https://ekblc.wordpress.com/wp-content/uploads/2013/09/ysa.pdf sayfasından erişilmiştir.

Kasap, Y. (2022). *Veri madenciliği yöntemleri ile ülkelerin PISA başarı düzeylerini etkileyen değişkenlerin incelenmesi.* (Doktora tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Kayhan, O. (2019). *Uzaktan eğitim öğrencilerinin mezuniyet durumlarının veri madenciliği yöntemleri ile tahmini: Amasya Üniversitesi örneği.* (Yüksek lisans tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Kayri, M. (2015). An Intelligent approach to educational data: performance comparison of the multilayer perceptron and the radial basis function artificial neural networks. *Educational Sciences: Theory and Practice, 15*, 1247-1255. doi:10.12738/estp.2015.5.0238

Kayri, M. (2007). Using chaid analysis in research and an application pertaining to coping strategies. *Ankara Universitesi Egitim Bilimleri Fakultesi Dergisi*, 001-017. doi:10.1501/egifak_0000000182

Kayri, M., & Şevgin, H. (Eds). (2023). *Eğitimde veri madenciliği ve bilgisayar uygulamaları*. Ankara: Pegem Akademi.

Kember, D. (1995). *Open learning courses for adults: a model of student progress.* https://www.tandfonline.com/doi/pdf/10.1080/0158791950160113 sayfasından erişilmiştir.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2003). *Preventing student dropout in distance learning using machine learning techniques*. In: Palade, V., Howlett, R.J., Jain, L. (eds) Knowledge-Based Intelligent Information and Engineering Systems. KES 2003. Lecture Notes in Computer Science(), vol 2774. Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-540-45226-3_37

Kovacic, Z. (2010, 19–24 June). *Early prediction of student success: Mining students' enrolment data.* Paper presented at the In Proceedings of the Informing Science and Information Technology Education Joint Conference, Cassino, Italy.

Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z., Vaclavek, J., & Wolff, A. (2015). OU Analyse: analysing at-risk students at The Open University. *Learning Analytics Review: LAK15-1*, 1–16.

Lee, Y., & Choi, J. (2011). A review of online course dropout research: implications for practice and future research. *Educational Technology Research and Development, 59*(5), 593-618. doi:10.1007/s11423-010-9177-y

Liang, J., Yang, J., Wu, Y., Li, C., & Zheng, L. (2016, 20-22 April 2016). *Big data application in education: dropout prediction in Edx MOOCs.* Paper presented at the 2016 IEEE Second International Conference on Multimedia Big Data (BigMM).

Lottering, R., Hans, R., & Lall, M. (2020). *A model for the identification of students at risk of dropout at a university of technology*. Paper presented at the 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education, 53*(3), 950-965. doi:10.1016/j.compedu.2009.05.010

Márquez, C., Cano, A., Romero, C., Mohammad, A., Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems, 33*, 107-124. doi:10.1111/exsy.12135

Marr, L., Gribble, J., & Ubachs, G. (2024). *Supporting retention and student services in online and distance education*. EADTU Working Group on Retention & Student Services https://dx.doi.org/10.5281/zenodo.11120440 sayfasından erişilmiştir.

Mathieson, D. E. (1971). Correspondence study: A summary review of the research and development literature. *Syracuse: ERIC Clearinghouse on Adult Education*.

Metlek, S., & Kayaalp, Ü. K. (2020). *Destek vektör makineleri: makine öğrenmesinde, teoriden örnek matlab uygulamalarına kadar*. Ankara: İksad.

Moldovan, A.-N., Ghergulescu, I., & Jha, N. (2019). *OULAD MOOC dropout and result prediction using ensemble, deep learning and regression techniques*. Paper presented at the Proceedings of the 11th International Conference on Computer Supported Education.

Munro, J. (1987). *The discourse of dropout in distance education: a theoretical analysis*. https://files.eric.ed.gov/fulltext/ED310752.pdf sayfasından erişilmiştir.

Nazempour, R. (2023). *Machine learning and statistical analysis to enhance learning outcomes in online learning environments.* (Ph.D.). University of Illinois at Chicago, United States -- Illinois. Retrieved from https://www.proquest.com/dissertations-theses/machine-learning-statistical-analysis-enhance/docview/2830118465/se-2?accountid=11637 ProQuest Dissertations & Theses Global database. (30603429)

Noffsinger, J. S. (1926). *Correspondence schools, lyceums, chatauquas*. New York: Macmillan.

Nouri, J., Larsson, K., & Saqr, M. (2019). Bachelor thesis analytics : using machine learning to predict dropout and identify performance factors. *International Journal of Learning Analytics and Artificial Intelligence for Education, 1*(1), 116-131. doi:10.3991/ijai.v1i1.11065

Okewu, E., & Daramola, O. (2017). *Design of a learning analytics system for academic advising in Nigerian universities.* Paper presented at the 2017 International Conference on Computing Networking and Informatics (ICCNI)(29-31 Ekim 2017).

Okur, M. R., Paşaoğlu Baş, D., & Uça Güneş, E. P. (2019). Açık ve uzaktan öğrenmede öğrenimi bırakma sebeplerinin incelenmesi [Examination of dropout causes in open and distance learning]. *Yükseköğretim ve Bilim Dergisi/journal of Higher Education and Science, 9*(2), 225–235. doi:10.5961/jhes.2019.324

Ortiz-Lozano, J. M., Aparicio-Chueca, P., Triadó-Ivern, X. M., & Arroyo-Barrigüete, J. L. (2023). Early dropout predictors in social sciences and management degree students. *Studies in Higher Education*, 1-14. doi:10.1080/03075079.2023.2264343

Özcan, T. (2013). *Veri madenciliği*. Retrieved from https://cdn-acikogretim.istanbul.edu.tr/auzefcontent/21_22_Bahar/veri_madenciligi/1/index.html

Öztürk, A. (2015). Açık ve uzaktan öğrenme sistemlerinde kümeleme analizi yöntemiyle öğrenen gruplarının belirlenmesi. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi,*

1(3), 4-7. https://earsiv.anadolu.edu.tr/xmlui/handle/11421/3051 sayfasından erişilmiştir.

Park, J.-H. (2007). Factors related to learner dropout in online learning. *Online Submission*. https://files.eric.ed.gov/fulltext/ED504556.pdf sayfasından erişilmiştir.

Paura, L., & Arhipova, I. (2014). Cause Analysis ofstudents' dropout rate in higher education study program. *Procedia - Social and Behavioral Sciences, 109(2014) :6*.

Perchinunno, P., Bilancia, M., & Vitale, D. (2019). A statistical analysis of factors affecting higher education dropouts. *Social Indicators Research, 156*, 341 - 362.

Pierrakeas, C., Xenos, M., Panagiotakopoulos, C., & Vergidis, D. (2004). A comparative study of dropout rates and causes for two different distance education courses. *International Review of Research in Open and Distance Learning, 5(2)*, 1-15. doi:https://doi.org/10.19173/irrodl.v5i2.183

Pinheiro, C. B., Ribeiro, J. L. L. d. S., & Fernandes, S. A. F. (2023). Theoretical models of dropout in higher education and notes on the national context. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*. doi:10.1590/S1414-40772023000100023

Poudyal, S. (2022). *Three case studies of using hybrid model machine learning techniques in educational data mining to ımprove the classification accuracies.* (Ph.D.). Mississippi State University, United States -- Mississippi. Retrieved from https://www.proquest.com/dissertations-theses/three-case-studies-using-hybrid-model-machine/docview/2704060802/se-2?accountid=11637 ProQuest Dissertations & Theses Global database. (29261802)

Pradeep, A., Das, S., & Kizhekkethottam, J. J. (2015). *Students dropout factor prediction using EDM techniques*. Paper presented at the 2015 International Conference on Soft-Computing and Networks Security (ICSNS).

Pulikottil, S. C., & Gupta, M. (2020). ONet – A temporal meta embedding network for mooc dropout prediction. Retrieved from https://ieeexplore.ieee.org/document/9378001

Qiu, L., Liu, Y., & Liu, Y. (2018). An integrated framework with feature selection for dropout prediction in massive open online courses. *Ieee Access, 6*, 71474-71484. doi:10.1109/ACCESS.2018.2881275

Quadri, M. M. N., & Kalyankar, D. N. V. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology, 10*.

Ramesh, V., P.Parkavi, & Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. *International Journal Of Computer Applıcations, 63*, 975-8887.

Reed, K., Wise, N., Tynan, B., & Bossu, C. (2013). Attrition in the digital age: reminders from the literature. In B. Tynan, J. Willems, & R. James (Eds.), *Outlooks and Opportunities in Blended and Distance Learning* (pp. 279-293). Hershey, PA, USA: IGI Global.

Ribeiro, R. C., & Canedo, E. D. (2020). Using data mining techniques to perform school dropout prediction: A Case Study. In (pp. 211-217): Springer International.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state-of-the-art.

Rovai, A. P. (2003). In search of higher persistence rates in distance education online programs. *The Internet and Higher Education, 6*(1), 1-16. doi:10.1016/s1096-7516(02)00158-6

Savaş, S., Topaloğlu, N., & Yılmaz, M. (2012). Veri madenciliği ve Türkiye'deki uygulama örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi,* 11(21), 1-23.

Schumann, J. A. (2005). *Data mining methodologies in educational organizations.* (Ph.D.). University of Connecticut, United States Connecticut. Retrieved from https://www.proquest.com/dissertations-theses/data-mining-methodologies-educational/docview/305010498/se-2?accountid=11637 ProQuest Dissertations & Theses Global database. (3180252)

Segura, M., Mello, J., & Hernández, A. (2022). Machine learning prediction of university student dropout: does preference play a key role? *mathematics*.

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011). *Prediction of student academic performance by an application of data mining techniques*. Paper presented at the International Conference on Management and Artificial Intelligence IPEDR, Indonesia.

Serra, A., Perchinunno, P., & Bilancia, M. (2018). Predicting student dropouts in higher education using supervised classification algorithms. In (pp. 18-33).

Shyamala, K. (2008). *A study on data mining techniques using higher educational system for efficient prediction.* (Ph.D.). Mother Teresa Women's University (India), India.

Retrieved from https://www.proquest.com/dissertations-theses/study-on-data-mining-techniques-using-higher/docview/2344721400/se-2?accountid=11637 ProQuest Dissertations & Theses Global database. (27732605)

Simpson, O. (2003). *Student retention in online, open and distance learning*. https://www.taylorfrancis.com/books/mono/10.4324/9780203416563 sayfasından erişilmiştir.

Simpson, O. (2013). Student retention in distance education: are we failing our students? *Open Learning, 28*, 105-119. doi:10.1080/02680513.2013.847363

Singer, C. G. (2023). *Educational data mining: an application of a predictive model of online student enrollment decisions.* (Ed.D.). Arizona State University, United States -- Arizona. Retrieved from https://www.proquest.com/dissertations-theses/educational-data-mining-application-predictive/docview/2901425184/se-2?accountid=11637 ProQuest Dissertations & Theses Global database. (30695078)

Souza, J., Matwin, S., & Japkowicz, N. (2002, 8–12 September 2002). *Evaluating data mining models: a pattern language.* Paper presented at the In: 9th Conference on Pattern Language of Programs (PLOP'02), Monticello, Illinois.

SPSSModeler, I. (2020). IBM Cloud Pak for Data/SPSS Modeler/Nodes palette/Modeling. https://dataplatform.cloud.ibm.com/docs/content/wsd/nodes/_nodes_modeling.html?context=cpdaas&audience=wdp sayfasından erişilmiştir.

Şahin, M. (2018). *E-Öğrenme ortamlarına yönelik öğrenme analitiklerine dayalı müdahale motoru tasarımı ve geliştirilmesi.* (Doktora tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Şengür, D., & Tekin, A. (2013). Öğrencilerin mezuniyet notlarının veri madenciliği metotları ile tahmini. *Bilişim Teknolojileri Dersgisi, 6*(3), 7-16.

Şevgin, H. (2020). ABİDE 2016 fen başarısının yordanmasında mars ve brt veri madenciliği yöntemlerinin karşılaştırılması. (Doktora tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*: Pearson Education.

Talan, T. (2020). *Uzaktan eğitim öğrencilerinin akademik başarı durumlarının çeşitli değişkenler açısından incelenmesi.* paper presented at the ejercongress 2020.

tamada, m. m., netto, j. f. d. m., & de Lima, D. P. R. (2019). *Predicting and reducing dropout in virtual learning using machine learning techniques: a systematic review*. Paper presented at the 2019 IEEE Frontiers in Education Conference (FIE). https://dx.doi.org/10.1109/fie43999.2019.9028545

Tarimo, A. (2013). Factors affecting students enrolment and dropout at the Open University of Tanzania, Lindi Region. Huria. *Journal of the Open University of Tanzania, 14(1)*, 111-133.

Tasnim, N., Paul, M., & Sattar, A. H. M. (2019). *Identification of drop out students using educational data mining*. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). 7-9 February, 2019. DOI: 10.1109/ECACE.2019.8679385

Taşdemir, M. (2012). *Veri madenciliği (Öğrenci başarısına etki eden faktörlerin regresyon analizi ile tespiti).* (Yüksek lisans tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Tekin, A., & Öztekin, Z. (2018). Eğitsel veri madenciliği ile ilgili 2006-2016 yılları arasında yapılan çalışmaların incelenmesi. *Eğitim Teknolojisi Kuram ve Uygulama, 8*(2), 108-124. doi:10.17943/etku.351473

Thalhammer, V., Hoffmann, S., Von Hippel, A., & Schmidt-Hertha, B. (2022). Dropout in adult education as a phenomenon of fit – an integrative model proposal for the genesis of dropout in adult education based on dropout experiences. *European Journal for Research on the Education and Learning of Adults*.

Theune, K., Behr, A., Giese, M., & Teguim Kamdjou, H. D. (2020). Early prediction of university dropouts – a random forest approach. *Jahrbücher für Nationalökonomie und Statistik, 240*(6), 743-789. doi:10.1515/jbnst-2019-0006

Thomas, L., & Unger, M. (2015). *Dropout and completion in higher education in Europe*. Luxembourg: Office of the European Union.

Tinto, V. (1975). Dropout from higher education: a theoretical synthesis of recent research. *Review of Educational Research, 45*, 125 - 189. doi:10.3102/00346543045001089

Tinto, V. (1982). Defining dropout: a matter of perspective. *New Directions for Institutional Research, 1982*(36), 3-15. doi:10.1002/ir.37019823603

Tinto, V., & John Cullen. (1973). *Dropout in Higher Education: A Review and Theoretical Synthesis of Recent Research*. https://eric.ed.gov/?id=ED078802 sayfasından erişilmiştir.

Topuz, S. (2021). *Eğitsel verilerde Weka ve Orange veri madenciliği yazılımlarından elde edilen analiz sonuçlarının karşılaştırılması.* (Yüksek lisans tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Tosun, M. (2016). *Açık öğretim öğrencilerinin akademik başarı düzeylerinin karşılaştırılması.* (Yüksek lisans tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Tosun, S. (2007). *Sınıflandırmada yapay sinir ağları ve karar ağaçları karşılaştırması: öğrenci başarıları üzerine bir uygulama.* (Doktora tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Véliz Palomino, J. C., & Ortega, A. M. (2023). Dropout intentions in higher education: systematic literature review. *Journal on Efficiency and Responsibility in Education and Science, 16*(2), 149-158. doi:10.7160/eriesj.2023.160206

Vossensteyn, H., Kottmann, A., Jongbloed, B., Kaiser, F., Cremonini, L., Stensaker, B., . . . Wollscheid, S. (2015). *Dropout and completion in higher education in Europe main report*. Retrieved from https://research.utwente.nl/files/5137444/dropout-completion-he-summary_en.pdf

Xavier, M., & Meneses, J. (2020). A literature review on the definitions of dropout in online higher education. *EDEN Conference Proceedings*(1), 73-80. doi:10.38069/edenconf-2020-ac0004

Yakut, E. (2012). *Veri madenciliği tekniklerinden c5.0 algoritması ve destek vektör makineleri ile yapay sinir ağlarının sınıflandırma başarılarının karşılaştırılması: imalat sektöründe bir uygulama..* (Yüksek lisans tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Yılmaz, A. (2020). *Açık ve uzaktan eğitim öğrencilerinin öğrenimi bırakma ve öğrenime devam nedenlerinin incelenmesi.* (Doktora tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Koçak, K. & Günay, A. (2021). *Üniversite-izleme-ve-degerlendirme-genel-raporu*. http//www.yok.gov.tr sayfasından erişilmiştir.

YÖKSİS. (2023). Higher education information management system; 2022-2023 academic year higher education statistics. https://istatistik.yok.gov.tr/ sayfasından erişilmiştir.

Yu, R., Lee, H., & Kizilcec, R. F. (2021). *Should college dropout prediction models include protected attributes?* Paper presented at the Proceedings of the Eighth ACM Conference on Learning @ Scale.

Yukselturk, E., Ozekes, S., & Türel, Y. (2014). Predicting dropout student: an application of data mining methods in an online education program. *European Journal of Open, Distance and E-Learning, 17*, 118-133. doi:10.2478/eurodl-2014-0008

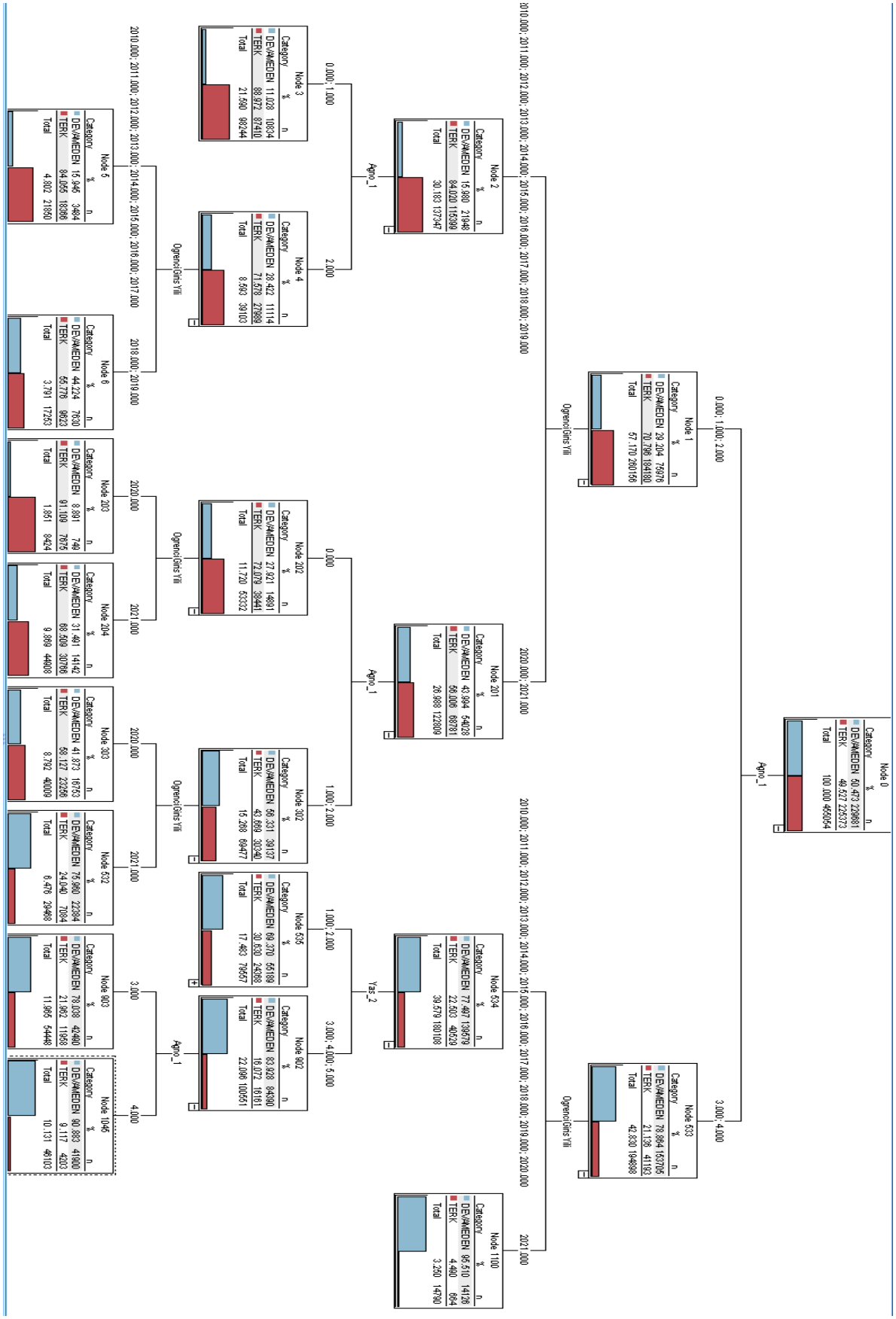
Yulia, L. W. S. (2019). The analysis of student performance using data mining. 559-573. doi:10.1007/978-981-13-6861-5_48

Zawacki-Richter, O., & Qayyum, A. (2019). Open and distance education in Asia, Africa and the Middle East national perspectives in a digital age. https://link.springer.com/book/10.1007/978-981-13-5787-9 sayfasından erişilmiştir.

Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). *Use data mining to improve student retention in higher education - A case study.* Paper presented at the International Conference on Enterprise Information Systems.

EKLER

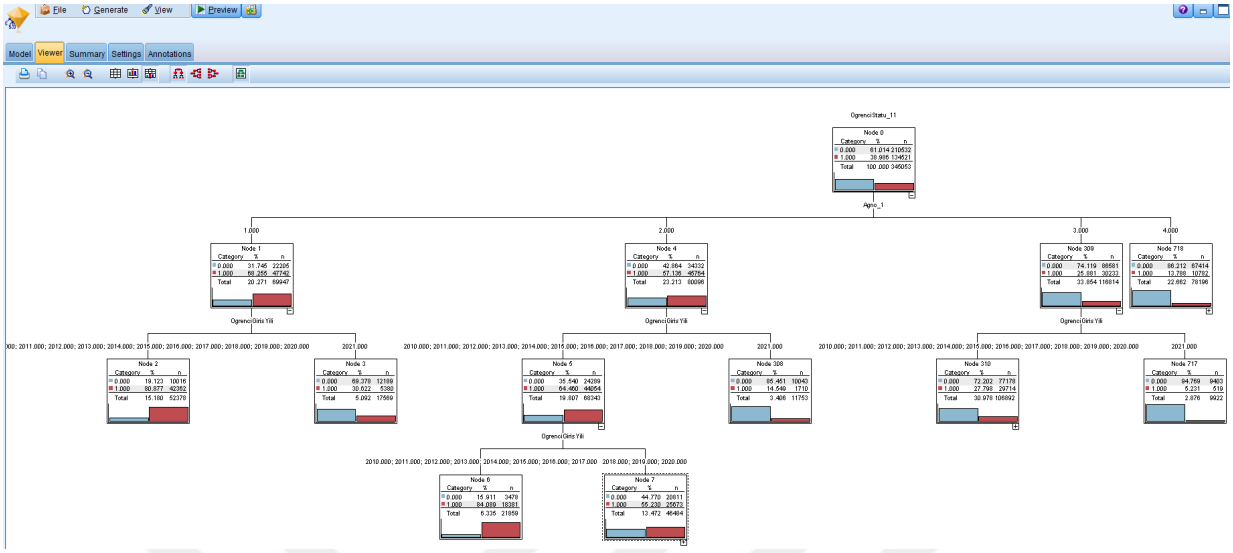
EK 1. Tüm veri için C5.0 algoritmasının ilk 4 düzeyi



EK 2. Okul terki öngörüsü için rastgele ağaçlar modelinde en iyi (TOP RULES) karar verme kuralları

Karar Kuralı	En Sık Kullanılan Kategori	Kural Doğruluğu	Interestingness (İlginçlik) indeksi
(OgrenciGirisYili <= 2014.0) and (OgrenciGirisYili <= 2017.0) and (OgrenciGirisYili <= 2020.0) and (Program_1 = [ACİL DURUM VE AFET YÖNETİMİ,BANKACILIK VE SİGORTACILIK,BİLGİSAYAR VE ÖĞRETİM TEKNOLOJİLERİ ÖĞRETMENLİĞİ,COĞRAFYA,COĞRAFİ BİLGİ SİSTEMLERİ,DIŞ TİCARET,EKONOMETRİ,ENDÜSTRİ MÜHENDİSLİĞİ,FELSEFE,GAZETECİLİK,HALKLA İLİŞKİLER VE TANITIM,HUKUK BÜRO YÖNETİMİ VE SEKRETERLİĞİ,KAMU YÖNETİMİ,KÜLTÜREL MİRAS VE TURİZM,MALİYE,MEDYA VE İLETİŞİM,PERAKENDE SATIŞ VE MAĞAZA YÖNETİMİ,RADYO, TELEVİZYON VE SİNEMA,SAĞLIK KURUMLARI İŞLETMECİLİĞİ,SOSYAL HİZMETLER,SOSYOLOJİ,SİVİL HAVA ULAŞTIRMA İŞLETMECİLİĞİ,TARİH,TÜRK DİLİ VE EDEBİYATI,ULUSLARARASI TİCARET VE LOJİSTİK YÖNETİMİ,ÇALIŞMA EKONOMİSİ VE ENDÜSTRİ İLİŞKİLERİ,ÇAĞDAŞ TÜRK LEHÇELERİ VE EDEBİYATLARI,İKTİSAT,İŞLETME]) and (Agno_1 = [1.0,2.0])	1.0	0.894	0.799
(AGNO_1 = [1.0,2.0]) and (Program_1 = [ADALET,BANKACILIK VE SİGORTACILIK,COĞRAFYA,COĞRAFİ BİLGİ SİSTEMLERİ,DIŞ TİCARET,EKONOMETRİ,FELSEFE,HUKUK BÜRO YÖNETİMİ VE SEKRETERLİĞİ,KÜLTÜREL MİRAS VE TURİZM,MALİYE,MEDYA VE İLETİŞİM,PERAKENDE SATIŞ VE MAĞAZA YÖNETİMİ,SAĞLIK KURUMLARI İŞLETMECİLİĞİ,SOSYAL HİZMETLER,SOSYOLOJİ,SİVİL HAVA ULAŞTIRMA İŞLETMECİLİĞİ,TARİH,ÇALIŞMA EKONOMİSİ VE ENDÜSTRİ İLİŞKİLERİ,ÇAĞDAŞ TÜRK LEHÇELERİ VE EDEBİYATLARI,İKTİSAT,İŞ SAĞLIĞI VE GÜVENLİĞİ,İŞLETME]) and (OgrenciGirisYili <= 2016.0)	1.0	0.874	0.764
(OgrenciGirisYili <= 2017.0) and (Program_1 = [ACİL DURUM VE AFET YÖNETİMİ,BANKACILIK VE SİGORTACILIK,BİLGİSAYAR VE ÖĞRETİM TEKNOLOJİLERİ ÖĞRETMENLİĞİ,COĞRAFYA,COĞRAFİ BİLGİ SİSTEMLERİ,DIŞ TİCARET,EGZERSİZ VE SPOR BİLİMLERİ,EKONOMETRİ,FELSEFE,GAZETECİLİK,GRAFİK TASARIMI,HALKLA İLİŞKİLER VE TANITIM,HEMŞİRELİK,HUKUK BÜRO YÖNETİMİ VE SEKRETERLİĞİ,KÜLTÜREL MİRAS VE TURİZM,LABORANT VE VETERİNER SAĞLIK,MALİYE,MARKA İLETİŞİMİ,MEDYA VE İLETİŞİM,PERAKENDE SATIŞ VE MAĞAZA YÖNETİMİ,RADYO, TELEVİZYON VE SİNEMA,REKREASYON,SAĞLIK KURUMLARI İŞLETMECİLİĞİ,SAĞLIK YÖNETİMİ,SOSYAL HİZMET,SOSYAL HİZMETLER,SOSYOLOJİ,SİVİL HAVA ULAŞTIRMA İŞLETMECİLİĞİ,SİYASET BİLİMİ VE KAMU YÖNETİMİ,SİYASET BİLİMİ VE ULUSLARARASI İLİŞKİLER,TARİH,TIBBİ DOKÜMANTASYON VE SEKRETERLİK,TÜRK DİLİ VE EDEBİYATI,ULUSLARARASI TİCARET VE LOJİSTİK,YAŞLI BAKIMI,YÖNETİM BİLİŞİM SİSTEMLERİ,ÇALIŞMA EKONOMİSİ VE ENDÜSTRİ İLİŞKİLERİ,ÇAĞDAŞ TÜRK LEHÇELERİ VE EDEBİYATLARI,İKTİSAT,İLİTAM,İŞ SAĞLIĞI VE GÜVENLİĞİ,İŞLETME]) and (OgrenciGirisYili <= 2018.0) and (OgrenciGirisYili <= 2020.0) and (AGNO_1 = [1.0,2.0])	1.0	0.863	0.745
(OgrenciGirisYili <= 2017.0) and (Yas_2 > 1.0) and (AGNO_1 = [2.0]) and (AGNO_1 = [1.0,2.0]) and (Program_1 = [ACİL DURUM VE AFET YÖNETİMİ,ADALET,BANKACILIK VE SİGORTACILIK,BİLGİSAYAR VE ÖĞRETİM TEKNOLOJİLERİ ÖĞRETMENLİĞİ,COĞRAFYA,COĞRAFİ BİLGİ SİSTEMLERİ,DIŞ TİCARET,EGZERSİZ VE SPOR BİLİMLERİ,EKONOMETRİ,FELSEFE,GAZETECİLİK,HALKLA İLİŞKİLER VE TANITIM,HUKUK BÜRO YÖNETİMİ VE SEKRETERLİĞİ,KAMU YÖNETİMİ,KÜLTÜREL MİRAS VE TURİZM,LABORANT VE VETERİNER SAĞLIK,MALİYE,MEDYA VE İLETİŞİM,PERAKENDE SATIŞ VE MAĞAZA YÖNETİMİ,RADYO, TELEVİZYON VE SİNEMA,REKREASYON,SAĞLIK KURUMLARI İŞLETMECİLİĞİ,SOSYAL HİZMETLER,SOSYOLOJİ,SİVİL HAVA ULAŞTIRMA İŞLETMECİLİĞİ,SİYASET BİLİMİ VE KAMU YÖNETİMİ,SİYASET BİLİMİ VE ULUSLARARASI İLİŞKİLER,TARİH,TIBBİ DOKÜMANTASYON VE SEKRETERLİK,TÜRK DİLİ VE EDEBİYATI,ULUSLARARASI TİCARET VE LOJİSTİK,ULUSLARARASI TİCARET VE LOJİSTİK YÖNETİMİ,YÖNETİM BİLİŞİM SİSTEMLERİ,ÇALIŞMA EKONOMİSİ VE ENDÜSTRİ İLİŞKİLERİ,ÇAĞDAŞ TÜRK LEHÇELERİ VE EDEBİYATLARI,İKTİSAT,İNSAN KAYNAKLARI YÖNETİMİ,İŞ SAĞLIĞI VE GÜVENLİĞİ,İŞLETME])	1.0	0.849	0.720

EK 3. AGNO> veri seti için C5.0 modelinin ağaç yapısı (3. düzey) ile 2. düzeyde düğümler ve değişkenlerin önem katsayıları



OgrenciStatu_11

File Generate View Preview

Model Viewer Summary Settings Annotations

1 2 3 4 5 6 7 8 All

- Agno_1 = 1.000 [Mode: 1] (69,947)
  - OgrenciGirisYili in [2010.000 2011.000 2012.000 2013.000 2014.000 2015.000 2016.000 2017.000 2018.000 2019.000 2020.000] [Mode: 1] ⇨ **1.0** (52,378; 0.809)
  - OgrenciGirisYili in [2021.000] [Mode: 0] ⇨ **0.0** (17,569; 0.694)
- Agno_1 = 2.000 [Mode: 1] (80,096)
  - OgrenciGirisYili in [2010.000 2011.000 2012.000 2013.000 2014.000 2015.000 2016.000 2017.000 2018.000 2019.000 2020.000] [Mode: 1] (68,343)
    - OgrenciGirisYili in [2010.000 2011.000 2012.000 2013.000 2014.000 2015.000 2016.000 2017.000] [Mode: 1] ⇨ **1.0** (21,859; 0.841)
    - OgrenciGirisYili in [2018.000 2019.000 2020.000] [Mode: 1] (46,484)
  - OgrenciGirisYili in [2021.000] [Mode: 0] ⇨ **0.0** (11,753; 0.855)
- Agno_1 = 3.000 [Mode: 0] (116,814)
  - OgrenciGirisYili in [2010.000 2011.000 2012.000 2013.000 2014.000 2015.000 2016.000 2017.000 2018.000 2019.000 2020.000] [Mode: 0] (106,892)
    - Program_1 in ["ACİL DURUM VE AFET YÖNETİMİ" "ACİL YARDIM VE AFET YÖNETİMİ" "ADALET" "BANKACILIK VE SİGORTACILIK" "BİLGİSAYAR VE ÖĞRETİM TEKNOLOJİLERİ ÖĞRE
    - Program_1 in ["COĞRAFYA"] [Mode: 0] (3,600)
    - Program_1 in ["FELSEFE"] [Mode: 0] (1,395)
    - Program_1 in ["GRAFİK TASARIMI" "MARKA İLETİŞİMİ"] [Mode: 0] ⇨ **0.0** (0)
    - Program_1 in ["HUKUK BÜRO YÖNETİMİ VE SEKRETERLİĞİ"] [Mode: 0] (916)
    - Program_1 in ["KÜLTÜREL MİRAS VE TURİZM"] [Mode: 0] (1,250)
    - Program_1 in ["LABORANT VE VETERİNER SAĞLIK"] [Mode: 0] (1,054)
    - Program_1 in ["MALİYE"] [Mode: 0] (153)
    - Program_1 in ["PERAKENDE SATIŞ VE MAĞAZA YÖNETİMİ"] [Mode: 1] ⇨ **1.0** (34; 0.559)
    - Program_1 in ["SAĞLIK KURUMLARI İŞLETMECİLİĞİ"] [Mode: 0] (2,116)
    - Program_1 in ["SOSYAL HİZMETLER"] [Mode: 0] (4,405)
    - Program_1 in ["SOSYOLOJİ"] [Mode: 0] (6,637)
    - Program_1 in ["SİVİL HAVA ULAŞTIRMA İŞLETMECİLİĞİ"] [Mode: 0] (3,174)
    - Program_1 in ["SİYASET BİLİMİ VE KAMU YÖNETİMİ"] [Mode: 0] (796)
    - Program_1 in ["SİYASET BİLİMİ VE ULUSLARARASI İLİŞKİLER"] [Mode: 1] (1,518)
    - Program_1 in ["TARİH"] [Mode: 0] (2,626)
    - Program_1 in ["TIBBİ DOKÜMANTASYON VE SEKRETERLİK"] [Mode: 0] (4,526)
    - Program_1 in ["ULUSLARARASI TİCARET VE LOJİSTİK"] [Mode: 1] (848)
    - Program_1 in ["YÖNETİM BİLİŞİM SİSTEMLERİ"] [Mode: 1] (2,528)
    - Program_1 in ["İKTİSAT"] [Mode: 1] (1,168)
    - Program_1 in ["İNSAN KAYNAKLARI YÖNETİMİ"] [Mode: 0] (1,282)
    - Program_1 in ["İŞ SAĞLIĞI VE GÜVENLİĞİ"] [Mode: 0] (4,454)
    - Program_1 in ["İŞLETME"] [Mode: 0] (3,784)
  - OgrenciGirisYili in [2021.000] [Mode: 0] ⇨ **0.0** (9,922; 0.948)
- Agno_1 = 4.000 [Mode: 0] (78,196)
  - OgrenciGirisYili in [2010.000 2011.000 2012.000 2013.000 2014.000 2015.000 2016.000 2017.000 2018.000 2019.000] [Mode: 0] ⇨ **0.0** (46,494; 0.923)
  - OgrenciGirisYili in [2020.000 2021.000] [Mode: 0] (31,702)
    - OgrenciGirisYili in [2020.000] [Mode: 0] (26,836)
    - OgrenciGirisYili in [2021.000] [Mode: 0] ⇨ **0.0** (4,866; 0.974)

EK 4. 'AGNO_2' için rastgele ağaçlar modelinde en iyi (TOP RULES) karar verme kuralları

Karar Kuralı	En Sık Kullanılan Kategori	Kural Doğruluğu	Orman Doğruluğu	Interestingness (İlginçlik) indeksi
(GelisSekli_1 =[AF,DGS,OZEL]) and (Program_1 =[ACİL DURUM VE AFET YÖNETİMİ,ACİL YARDIM VE AFET YÖNETİMİ,EGZERSİZ VE SPOR BİLİMLERİ,HEMŞİRELİK,MEDYA VE İLETİŞİM,REKREASYON,SAĞLIK YÖNETİMİ,SOSYAL HİZMET,ÇOCUK GELİŞİMİ,İLİTAM]) and (OgrenciGirisYili > 2016.0) and (OgrenciGirisYili <= 2020.0)	1.0	0.899	0.899	0.808
(OgrenciGirisYili > 2016.0) and (GelisSekli_1 =[AF,DGS,OZEL]) and (OgrenciGirisYili <= 2020.0) and (Program_1 =[ACİL DURUM VE AFET YÖNETİMİ,ACİL YARDIM VE AFET YÖNETİMİ,BİLGİSAYAR VE ÖĞRETİM TEKNOLOJİLERİ ÖĞRETMENLİĞİ,EGZERSİZ VE SPOR BİLİMLERİ,ENDÜSTRİ MÜHENDİSLİĞİ,HEMŞİRELİK,RADYO, TELEVİZYON VE SİNEMA,REKREASYON,SAĞLIK YÖNETİMİ,SOSYAL HİZMET,ÇOCUK GELİŞİMİ,İLİTAM])	1.0	0.890	0.890	0.793
(GelisSekli_1 =[AF,DGS,OZEL]) and (İLMERNIS_2 =[2.0,4.0,5.0,6.0,83.0]) and (ProgramDuzeyi =[LİSANS PROGRAMI,LİSANS TAMAMLAMA PROGRAMI]) and (OgrenciGirisYili <= 2020.0) and (Program_1 =[ACİL DURUM VE AFET YÖNETİMİ,ACİL YARDIM VE AFET YÖNETİMİ,BİLGİSAYAR VE ÖĞRETİM TEKNOLOJİLERİ ÖĞRETMENLİĞİ,EGZERSİZ VE SPOR BİLİMLERİ,ENDÜSTRİ MÜHENDİSLİĞİ,HEMŞİRELİK,MEDYA VE İLETİŞİM,RADYO, TELEVİZYON VE SİNEMA,REKREASYON,SAĞLIK YÖNETİMİ,SOSYAL HİZMET,ÇOCUK GELİŞİMİ,İLİTAM])	1.0	0.884	0.884	0.779
(GelisSekli_1 =[AF,DGS,OZEL]) and (ProgramDuzeyi =[LİSANS PROGRAMI,LİSANS TAMAMLAMA PROGRAMI]) and (Program_1 =[ACİL DURUM VE AFET YÖNETİMİ,ACİL YARDIM VE AFET YÖNETİMİ,BİLGİSAYAR VE ÖĞRETİM TEKNOLOJİLERİ ÖĞRETMENLİĞİ,EGZERSİZ VE SPOR BİLİMLERİ,ENDÜSTRİ MÜHENDİSLİĞİ,HEMŞİRELİK,MEDYA VE İLETİŞİM,RADYO, TELEVİZYON VE SİNEMA,REKREASYON,SAĞLIK YÖNETİMİ,SOSYAL HİZMET,YAŞLI BAKIMI,ÇOCUK GELİŞİMİ,İLİTAM]) and (OgrenciGirisYili <= 2020.0) and (İLMERNIS_2 =[2.0,3.0,4.0,5.0,6.0,83.0])	1.0	0.878	0.880	0.771
(GelisSekli_1 =[AF,DGS,OZEL]) and (OgrenciGirisYili <= 2020.0) and (Program_1 =[ACİL DURUM VE AFET YÖNETİMİ,ACİL YARDIM VE AFET YÖNETİMİ,EGZERSİZ VE SPOR BİLİMLERİ,ENDÜSTRİ MÜHENDİSLİĞİ,HEMŞİRELİK,RADYO, TELEVİZYON VE SİNEMA,REKREASYON,SAĞLIK YÖNETİMİ,SOSYAL HİZMET,TÜRK DİLİ VE EDEBİYATI,ÇOCUK GELİŞİMİ,İLİTAM])	1.0	0.862	0.863	0.740

Ek Bilgi: Grafik çizim ve referans destek

Bu çalışmanın temel araştırma süreçlerinde referans yönetimi için "EndNote20" paket programından ve özellikle kuramsal çerçevede yer alan modellerin şema çizimleri için " tldraw" (https://www.tldraw.com/) aracından yararlanılmıştır.



GAZİLİ OLMAK AYRICALIKTIR...