HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı
Eğitimde Ölçme ve Değerlendirme Programı

PISA BAŞARISINI TAHMİN ETMEDE KULLANILAN VERİ MADENCİLİĞİ YÖNTEMLERİNİN İNCELENMESİ

Gökhan AKSU

Doktora Tezi

Ankara, 2018

Liderlik, araştırma, inovasyon, kaliteli eğitim ve değişim ile

Daha ileriye... En İyiye...

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı
Eğitimde Ölçme ve Değerlendirme Programı

PISA BAŞARISINI TAHMİN ETMEDE KULLANILAN VERİ MADENCİLİĞİ YÖNTEMLERİNİN İNCELENMESİ

INVESTIGATION OF DATA MINING METHODS USED FOR ESTIMATING PISA SUCCESS

Gökhan AKSU

Doktora Tezi

Ankara, 2018

Kabul ve Onay

Eğitim Bilimleri Enstitüsü Müdürlüğüne,

Gökhan AKSU'nun hazırladığı "PISA Başarısını Tahmin Etmede Kullanılan Veri Madenciliği Yöntemlerinin İncelenmesi" başlıklı bu çalışma jürimiz tarafından **Eğitim Bilimleri Ana Bilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalında Doktora Tezi** olarak kabul edilmiştir.

Jüri Başkanı	Prof. Dr. Selahattin GELBAL	İmza
Jüri Üyesi (Danışman)	Doç.Dr. Nuri DOĞAN	İmza
Jüri Üyesi	Prof. Dr. Zekeriya NARTGÜN	İmza
Jüri Üyesi	Prof.Dr. Cem Oktay GÜZELLER	İmza
Jüri Üyesi	Dr. Öğr. Üyesi Kübra ATALAY KABASAKAL	İmza

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim, Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca yukarıdaki jüri üyeleri tarafından 13 / 04 / 2018 tarihinde uygun görülmüş ve Enstitü Yönetim Kurulunca ..... / ..... / ..... tarihinde kabul edilmiştir.

Prof. Dr. Ali Ekber ŞAHİN
Eğitim Bilimleri Enstitüsü Müdürü

Öz

Bu çalışmada veri madenciliği ve makine öğrenme yaklaşımının eğitim alanında kullanılması ve bu algoritmalara dayalı olarak elde edilen sonuçların güvenirlik ve geçerlik değerlerinin ne düzeyde olduğu belirlenmeye çalışılmıştır. PISA 2015 Türkiye ortalamasına göre öğrencilerin başarılı ve başarısız olarak sınıflandığı çalışmada farklı öğrenme yöntemleri kullanılarak fen okuryazarlığı bakımından öğrencilerin hangi sınıfta yer alacağı tahmin edilmiş ve bu aşamada elde edilen sonuçların güvenirlik ve geçerlik ölçütleri incelenmiştir. Bunun yanında WEKA programının sahip olduğu tüm algoritma ve yöntemler farklı koşullar altında karşılaştırılarak hangi makine öğrenme yönteminin hangi durumlarda avantajlı veya dezavantajlı olduğu belirlenmiştir. Araştırmada elde edilen sonuçlar doğrultusunda PISA Fen okuryazarlığını yordamak amacıyla kullanılacak değişken sayısının 29 olması sebebiyle ilk olarak farklı algoritmalar yardımıyla en iyi özellikler belirlenmeye çalışılmıştır ve 10 katlı çapraz geçerleme yöntemiyle her katmanda başarılı olan 12 değişken yardımıyla PISA 2015 fen okuryazarlığı başarıları tahmin edilmiştir. Sonrasında çalışma kapsamında ele alınan 8 farklı öğrenme yönteminden doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği, karekök hata ve göreceli karekök hata değerleri bakımından en iyi sonuçların Random Forest yöntemiyle elde edilirken Ridge lojistik regresyon, Lojistik model ve Hoefding tree yöntemlerinin en başarılı diğer yöntemler olduğu belirlenmiştir. Çapraz geçerleme yöntemi kullanılmadan tüm veri setinin eğitim ve test veri seti olarak ayrılması durumunda Lojistik model, Random Forest ve Ridge Regresyon yöntemlerinin farklı büyüklükteki test verilerinde en düşük hata değerlerini verirken Random Tree ve J.48 yönteminlerinin en yüksek hata değerlerine sahip olduğu belirlenmiştir. Ridge regresyon, Random forest ve Lojistik model tarafından elde edilen hata değerlerinin de farklı yüzdelikteki test verilerinde oldukça tutarlı olduğu sonucuna ulaşılmıştır. Farklı yöntemler yardımıyla elde edilen ölçme sonuçlarının veri setini test ve eğitim verisi olarak ayırmayıp aynı veri seti üzerinden hem öğrenme yöntemini eğitip hem de test ettiğimiz taktirde özellikle Random tree ve J.48 öğrenme yöntemlerinin gerçek performanslarından daha yüksek doğru sınıflama oranına sahip oldukları belirlenmiştir.

Anahtar sözcükler: Veri madenciliği, sınıflama, tahmin, öğrenme, PISA, WEKA.

Abstract

In this study, it was tried to determine the use of data mining and machine learning approach in the field of education and the level of reliability and validity of the results obtained based on these algorithms. In the study that students were classified as successful and unsuccessful according to the Turkey's PISA average, it was predicted that in which class the students will take place in terms of science literacy using different learning methods, and the reliability and validity criteria of the results obtained at this stage were examined. Besides, all the algorithms and methods of Weka program were compared under different conditions, and it was determined that which learning method is advantageous or disadvantageous in which situations. In the study, the best results in terms of correct classification number, correct classification ratio, kappa statistic, square root error and relative square root error were obtained from Random Forest method, and It was identified that Ridge logistic regression, logistic model and Hoefding tree methods are the most successful other methods It was also determined that in case the whole data set is separated as training and test data set without using the cross validation method, the Logistic model, Random Forest and Ridge Regression methods gave the lowest error values in test data with different size, and Random Tree and J.48 methods have the highest error values. It was concluded that the error values obtained by the Ridge regression, Random forest and Logistic model were quite consistent in test data.in different percentile. It was determined that if we do not allocate the data set of the measurement results obtained by different methods as test and training data and we train and test the learning method through the same data set, especially Random tree and J.48 learning methods have a higher correct classification rate than real performances.

Keywords: Data mining, classification, prediction, learning, PISA, WEKA

Teşekkür

Ölçme ve Değerlendirme bilim dalında özel öğrenci olarak eğitime başladığım ilk günden bu yana bilgisini ve emeğini esirgemeden önerileri ve fikirleriyle beni yönlendiren, tez sürecinin tüm aşamalarında yanımda olan ve kafama takılan her soru için çekinmeden kapısını çalabildiğim, her koşulda anlayışlı tavrı ve yardımseverliği için değerli danışmanım Doç. Dr. Nuri Doğan'a içten teşekkürlerimi sunarım.

Doktora öğrenimim boyunca her zaman desteklerini hissettiğim ayrıca görüş ve önerileriyle de bu çalışmaya katkıda bulunan değerli hocalarım Prof. Dr. Cem Oktay GÜZELLER'e ve Prof. Dr. Zekeriya NARTGÜN'e sonsuz teşekkürlerimi sunarım.

Ölçme ve değerlendirmeye alanında eğitime başladığım ilk günden bu yana bizlerden bilgi, destek ve emeklerini esirgemeyen doktora eğitimim boyunca aldığım dersler aracılığıyla bilgisinden ve deneyiminden faydalandığım değerli hocalarım Prof. Dr. Selahattin GELBAL'a, Prof. Dr. Hülya KELECİOĞLU'na ve Prof. Dr. Halil YURDUGÜL'e teşekkür ederim.

Tezimi tamamlayabilmem için her zaman bana destek olan değerli eşim Nurşah AKSU'ya sonsuz teşekkürlerimi sunarım.

İçindekiler

Öz ii

Abstract iii

Teşekkür ivv

Tablolar Dizini vii

Şekiller Dizini x

Simgeler ve Kısaltmalar Dizini xii

Bölüm 1 Giriş 1

Problem Durumu 2

Araştırmanın Amacı ve Önemi 3

Araştırma Problemi 5

Sınırlılıklar 5

Tanımlar 6

Bölüm 2 Araştırmanın Kuramsal Temeli ve İlgili Araştırmalar 8

Veri Madenciliği ve Makine Öğrenme 8

Temel Yöntemler ve Algoritmalar 19

Elde Edilen Sonuçların Değerlendirilmesi 32

Kestirimin Gücü: Hatanın Büyüklüğünü Dikkate Alma 47

Veri Madenciliği ile İlgili Araştırmalar 58

Bölüm 3 Yöntem 67

Araştırmanın Türü 67

Araştırmanın Evreni ve Örneklemi 68

Veri Toplama Süreci 68

Veri Toplama Araçları 68

Verilerin Analizi 69

Verilerin Analizinde Kullanılan Yazılım ve Algoritmalar 70

Bölüm 4 Bulgular ve Yorumlar 75

Birinci Alt Probleme İlişkin Bulgular 82
İkinci Alt Probleme İlişkin Bulgular 94
Üçüncü Alt Probleme İlişkin Bulgular 114
Bölüm 5 Sonuç, Tartışma ve Öneriler 117
Sonuçlar ve Tartışma 117
Öneriler 125
Kaynaklar 129
EK-A: Matlab'tan Elde Edilen Karışıklık Matrisi 140
EK-B: Matlab'tan Elde Edilen ROC Eğrisi 141
EK-C: Etik Komisyonu Onay Bildirimi 142
EK-Ç: Etik Beyanı 143
EK-D: Doktora Tez Çalışması Orijinallik Raporu 144
EK-E: Dissertation Originality Report 145
EK-F: Yayımlama ve Fikrî Mülkiyet Hakları Beyanı 146

Tablolar Dizini

Tablo 1 *Hava Durumu ile ilgili Karar Tablosu* 9
Tablo 2 *Karar Tablosu Örneği* 13
Tablo 3 *Hava Durumu ile İlgili Veri dosyasındaki Özellikleri Değerlendirme* 22
Tablo 4 *Hava Durumu Verisinin Sayısal Olarak Betimlenmesi* 23
Tablo 5 *Yeni Bir Güne İlişkin Veri* 24
Tablo 6 *Başka Yeni Bir Gün* 27
Tablo 7 *Normal Dağılım için Güven Aralıkları* 45
Tablo 8 *İki Sınıflı Bir Tahmine İlişkin Farklı Sonuçlar* 49
Tablo 9 *Üç Sınıflı Bir Tahmine İlişkin Farklı Sonuçlar* 50
Tablo 10 *Hayvanların Türlerine İlişkin Gerçek Durum ve Tahmin Sonuçları* 51
Tablo 11 *Sayısal Tahminin Değerlendirmesinde Kullanılan Performans Ölçütleri* 55
Tablo 12 *Dört farklı sayısal tahmin modelleri için performans ölçütleri* 56
Tablo 13 *Çalışma Kapsamında Kullanılan Değişkenlere İlişkin Tanımlayıcı Bilgiler* 69
Tablo 14 *Weka Programındaki Algoritmalar ve İşlevleri* 71
Tablo 15 *Çalışma Kapsamında Kullanılacak Değişkenlere İlişkin İstatistikler* 75
Tablo 16 *Kayıp veri atama yöntemine ilişkin sonuçlar* 78
Tablo 17 *PISA Okuryazarlığını En İyi Yordayan Değişkenler* 79
Tablo 18 *PISA Okuryazarlığını En İyi Yordayan Değişkenler için Çapraz Geçerlik Sonuçları* 79
Tablo 19 *Veri Madenciliğinde Kullanılacak Değişkenlere İlişkin Özet Bilgiler* 81
Tablo 20 *Decision Stump Yöntemiyle Elde Edilen Güvenirlik Değerleri* 82
Tablo 21 *Decision Stump Yöntemiyle Elde Edilen Geçerlik Ölçütleri* 83
Tablo 22 *Hoeffding Tree Yöntemiyle Elde Edilen Güvenirlik Değerleri* 83
Tablo 23 *Hoeffding Tree Yöntemiyle Elde Edilen Geçerlik Ölçütleri* 84
Tablo 24 *J4.8 Yöntemiyle Elde Edilen Güvenirlik Değerleri* 85
Tablo 25 *J4.8 Yöntemiyle Elde Edilen Geçerlik Ölçütleri* 85
Tablo 26 *Lojistik Model Yöntemiyle Elde Edilen Güvenirlik Değerleri* 86
Tablo 27 *Lojistik Model Yöntemiyle Elde Edilen Geçerlik Ölçütleri* 87
Tablo 28 *REPTree Yöntemiyle Elde Edilen Güvenirlik Değerleri* 88
Tablo 29 *REPTree Yöntemiyle Elde Edilen Geçerlik Ölçütleri* 88
Tablo 30 *Random Forest Yöntemiyle Elde Edilen Güvenirlik Değerleri* 89

Tablo 31 *Random Forest Yöntemiyle Elde Edilen Geçerlik Ölçütleri* 90
Tablo 32 *Random Tree Yöntemiyle Elde Edilen Güvenirlik Değerleri* 90
Tablo 33 *Random Tree Yöntemiyle Elde Edilen Geçerlik Ölçütleri*..... 91
Tablo 34 *Ridge Regresyon Yöntemiyle Elde Edilen Güvenirlik Değerleri*..... 92
Tablo 35 *Ridge Regresyon Yöntemiyle Elde Edilen Geçerlik Ölçütleri*..... 92
Tablo 36 *Farklı Algoritmalar Kullanılarak Elde Edilen güvenirlik Ölçütleri* 93
Tablo 37 *Farklı Algoritmalar Kullanılarak Elde Edilen Geçerlik Ölçütleri* 94
Tablo 38 *Decision Stump Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri*..... 95
Tablo 39 *Decision Stump Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri*.... 96
Tablo 40 *Hoeffding Tree Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri*..... 97
Tablo 41 *Hoeffding Tree Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri*..... 98
Tablo 42 *J.48 Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri* . 99
Tablo 43 *J.48 Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri*..... 101
Tablo 44 *Lojistik Model Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri*..... 102
Tablo 45 *LMT Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri* 103
Tablo 46 *RepTree Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri*..... 104
Tablo 47 *RepTree Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri*..... 105
Tablo 48 *Random Forest Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri*..... 106
Tablo 49 *Random Forest Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri* .. 107
Tablo 50 *Random Tree Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri*..... 108
Tablo 51 *Random Tree Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri*..... 109
Tablo 52 *Ridge Regresyon Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri*..... 110
Tablo 53 *Ridge Regresyon Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri* 111
Tablo 54 *Farklı Algoritmalar Kullanılarak Elde Edilen Doğru Pozitif Oranları* 112
Tablo 55 *Farklı Algoritmalar Kullanılarak Elde Edilen Hataların Ortalama Karekökü* 113

Tablo 56 *Farklı Öğrenme Yöntemlerine İlişkin Karışıklık Matrisi ve Doğruluk Oranları* 115
Tablo 57 Veri madenciliği yöntemlerini karşılaştırmada kullanılabilecek standart ölçütler..... 124

Şekiller Dizini

Şekil 1. Sınıflama İçin Genel Bir Karar Ağacı Örneği. 10
Şekil 2. Problem Durumu - Özel Durum. 14
Şekil 3. Tekrarlı Alt Ağaçları Olan Karar ağacı 15
Şekil 4. Hava Durumu Verisi İçin Ağaç Kütükleri. 29
Şekil 5. Eğitim ve test verilerileri üzerinden öğrenme süreci..... 34
Şekil 6. Çapraz Geçerlik Süreci..... 37
Şekil 7. Bootstrap Yönteminde Örnekleme Süreci..... 41
Şekil 8. Tek Kuyruk Testi İçin Olasılık Dağılımları. 44
Şekil 9. Güven Aralıkları Ve Olasılık Dağılımı. 46
Şekil 10. Araştırma Süreci 67
Şekil 11. Kayıp Verilere İlişkin Özetleyici Bilgiler. 76
Şekil 12. Kayıp Veri Örüntüsü Grafiği. 77
Şekil 13. PISA 2015 Verilerinin Weka Programına Aktarılması 78
Şekil 14. Sosyo Ekonomik Durum İndeksi Özelliği İçin Dağılım Grafiği. 81
Şekil 15. Hoeffding Tree Yöntemiyle Elde Edilen Ağaç Yapısı 84
Şekil 16. J.48 yöntemiyle Elde Edilen ağaç Yapısı 86
Şekil 17. Lojistik Model Yöntemiyle Elde Edilen Ağaç Yapısı. 87
Şekil 18. REPTree Yöntemiyle Elde Edilen Ağaç Yapısı. 89
Şekil 19. Random Tree Yöntemiyle Elde Edilen Ağaç Yapısı. 91
Şekil 20. Decision Stump Yöntemiyle Elde Edilen Hata Ölçütlerinin Değişimi..... 96
Şekil 21. Decision Stump Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri.... 97
Şekil 22. Hoeffding Tree Yöntemiyle Elde Edilen Hata Ölçütlerinin Değişimi 98
Şekil 23. Hoeffding Tree Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütlerinin Değişimi..... 99
Şekil 24. J.48 Yöntemiyle Elde Edilen Hata Ölçütlerinin Değişimi 100
Şekil 25. J.48 Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütlerinin Değişimi. 101
Şekil 26. Lojistik Model (LMT) Yöntemiyle Elde Edilen Hata Ölçütlerinin Değişimi 102
Şekil 27. Lojistik Model Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri. 103
Şekil 28. RepTree Yöntemiyle Elde Edilen Hata Ölçütlerinin Değişimi..... 104
Şekil 29. RepTree Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütlerinin Değişimi 105

Şekil 30. Randon Forest Yöntemiyle Elde Edilen Hata Ölçütlerinin Değişimi. 106
Şekil 31. Random Forest Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütlerinin Değişimi..... 107
Şekil 32. Randon Tree Yöntemiyle Elde Edilen Hata Ölçütlerinin Değişimi. 108
Şekil 33. Random Tree Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütlerinin Değişimi..... 109
Şekil 34. Ridge Regresyon Yöntemiyle Elde Edilen Hata Ölçütlerinin Değişimi . 110
Şekil 35. Ridge Regresyon Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütlerinin Değişimi..... 111
Şekil 36. Farklı Algoritmalar Kullanılarak Elde Edilen Doğru Pozitif Oranlarının Değişimi..... 113
Şekil 37. Farklı Algoritmalar Kullanılarak Elde Edilen Hataların Değişimi..... 114
Şekil 38. Farklı Öğrenme Yöntemlerine İlişkin Doğruluk Oranları..... 116

Simgeler ve Kısaltmalar Dizini

PISA: Uluslararası Öğrenci Değerlendirme Programı

WEKA: Waikato Environment for Knowledge Analysis

MEB: Milli Eğitim Bakanlığı

OECD: The Organization for Economic Cooperation and Development

VM: Veri Madenciliği

Bölüm 1
Giriş

Teknoloji çağının yaşandığı günümüzde her geçen gün bilgi miktarı sürekli artış göstermektedir (Larose, 2005). Bu sebeple sahip olunan bilginin depolanması ve bu işlenmemiş ham verilerden anlamlı bilgiler elde etmek oldukça zorlaşmaktadır (Fayyad, Piatetsky-Shapiro ve Smyth, 1996). Büyük miktarlardaki verilerin büyük veri tabanlarında bilgisayar programları vasıtasıyla aranarak, elde edilen sonuçlar yardımıyla gelecekle ilgili tahmin yapılması işlemlerine veri madenciliği (VM) denilmektedir (Thuarisingham, 2003). Geleceğe dair tahmin yapılabilmesi için geçmişe dönüp, geçmişte bu konularla ilgili ne gibi bilgiler olduğunu ve ne gibi uygulamalar yapıldığını görmek gerekir. Günümüzde bu amaçla birçok algoritma ve yazılım geliştirilmiştir. Bu algoritma ve yazılımlar sayesinde, araştırmacıların işleri oldukça kolaylaşmıştır (Imielinski ve Mannila, 1996).

Bazen bilgi keşfi olarak adlandırılan VM veriyi yeni bir perspektiften analiz etmek ve bu veriler arasındaki yararlı yeni bilgileri özetleme sürecini tanımlamak için kullanılmaktadır (Sieber, 2008). VM'de büyük miktardaki verinin içinde saklı olan ve araştırmacılar için oldukça faydalı olan bilgilerin bir dizi işlemlerin ardından ortaya çıkarılması hedeflendiğinden cevher elde etmek için yapılan madenciliğe benzetilmektedir (Aydın, 2007). Bu sistemlerin algoritmaları genel olarak tahminleme veya sınıflama teknikleri üzerine kuruludur ve eldeki verilerden henüz bilinmeyen nesnelerin davranışlarını tahmin etmek için kullanılabilen ampirik verilerin sınıflandırma şemalarının oluşturulmasını hedeflemektedir (Weiss ve Kulikowski, 1991).

Veri tabanlarından elde edilen bilginin keşfi sürecinde verilerin ve yapılan işin özelliklerinin bilinmemesi, VM algoritması ne kadar iyi olursa olsun faydalı sonuçlar vermesi mümkün değildir. Dolayısıyla ilk önce verilerin ve işin özelliklerinin öğrenilmesi gerekir. Daha sonra aşağıdaki aşamalar gerçekleştirilir.

- Problemin Tanımlanması,
- Verilerin Hazırlanması,
- Modelin Kurulması ve Değerlendirilmesi,

- Modelin Kullanılması,
- Modelin İzlenmesi

veri tabanlarında bilgi keşfi sürecinde izlenmesi gereken temel aşamalardır (Chen, Han ve Yu, 1996).

Son yıllarda WEKA (Waikato Environment for Knowledge Analysis), Enterprise Mining ve Clementine gibi veri madenciliği programları, jeoloji, tıp, pazarlama, bankacılık ve diğer ticari alanlarda elde edilen büyük veri gruplarına başarılı bir şekilde uygulanmaktadır (Mannila, 1997). İlgili alan yazın incelendiğinde VM yöntemlerinin eğitim alanında çok fazla uygulaması olmadığı görülmektedir. Bunun yanında eğitim alanında çok az sayıda yapılan makale ve tezlerde ise veri madenciliği uygulamalarının sadece sınıflama amaçlı kullanıldığı belirlenmiştir. Bu çalışmada uluslararası geniş ölçekli sınavlardan biri olan PISA (**P**rogramme for **İ**nternational **S**tudent **A**ssessment) sınavı ile öğrenci ve okula ilişkin oldukça fazla verinin içerisinden anlamlı sonuçlar çıkarmak ve eğitim alanında VM yöntemlerini kullanılmak amaçlanmaktadır. Bunun yanında sınıflama ve tahmin yöntemlerinin farklı algoritmalar kullanılarak elde edilen güvenirlik ve geçerlik değerlerinin farklı koşullar altında karşılaştırılması amaçlanmaktadır.

Problem Durumu

Günümüzde bireyler hakkında farklı ortamlarda ve farklı amaçlarla çok fazla veri toplanmaktadır. Bu veriler yardımıyla bireylerin birçok farklı özelliğine dayalı olarak onların satın alma, yeme-içme ve benzeri davranışları hakkında tahminde bulunulmaktadır. Bu aşamada elde edilen çok fazla veri arasından hangilerinin karar verme sürecinde anlamlı hangilerinin değersiz olduğunun belirlenmesi büyük önem taşımaktadır. Bireylerin özelliklerine ilişkin değişkenlerin yapılan tahminleme ve karar alma işlemlerinde elde edilen sonuçların doğruluğu üzerinde büyük etkisi bulunduğundan hangi değişkenlerin bizim için önemli olduğunun belirlenmesi gerekmektedir. Bilgisayar yazılımlarının gelişimiyle çok fazla veri arasında var olan gizli örüntüleri ortaya çıkarmak ve geleceğe ilişkin tahminde bulunmak günümüzde araştırmacıların ve karar alıcı mercilerin üzerinde çalıştıkları bir konu haline gelmiştir. Büyük miktardaki verilerden anlamlı sonuçlar çıkarabilmek için son yıllarda büyük önem kazanan veri madenciliği yöntemleri ile bireylerin davranış örüntüleri analiz edilerek gelecekteki davranışları hakkında tahminde

bulunulmaktadır. Her ne kadar farklı alanlarda son yıllarda kullanımı yaygınlaşmaya başlasa da veri madenciliğinde teorik temelleri farklı istatistiksel modellere dayalı olan birçok farklı yöntem bulunmaktadır. Söz konusu yöntemlerin fazla olması sebebiyle bu yöntemlerden hangilerinin daha etkili kestirimlerde bulunduğu ve hangilerinin daha az hatalı hesaplamalar yaptığının belirlenmesi elde edilen sonuçların güvenirliği ve geçerliği bakımından büyük önem taşımaktadır. Öte yandan VM'nin eğitim alanına uyarlanması için yeterli çalışma yapılmadığı ancak VM'nin disiplinler arası bir alan olan eğitim alanında da öğrenciler, öğretmenler ve eğitim alanında söz sahibi olan yetkililer için etkili analizler ve bu analizlere dayalı olarak önemli sonuçlar ortaya çıkaracağı düşünülmektedir. Özellikle eğitim alanında ölçme sonuçlarına dayalı olarak seçme ve sınıflama işlevi için VM'nin işe koşulabileceği düşünülmektedir. Bu sayede hangi sınıfta hangi değişkenlerin etkili olduğu belirlenerek öğrencilerin öğrenme davranışları daha iyi anlaşılabilecektir. Bu konuda eğitim alanında çok fazla çalışma olmaması sebebiyle geniş ölçekli sınavlardan biri olan PISA verileri yardımıyla VM yöntemlerinin eğitim alanında uygulanması ve sonuçlarının karşılaştırılması amaçlanmıştır. PISA, TIMMS (Trends in International Mathematics and Science Study) vb. geniş ölçekli sınavlarda çok sayıda değişkene ilişkin bilgi toplandığından bireylerin başarılarının sadece test sonuçlarına dayanarak karara varılıp varılamayacağının belirlenmesi ve bu aşamada çok sayıda değişken kullanarak yapılacak tahminlerin ne düzeyde doğru olduğunun belirlenmesi gerekmektedir. Özellikle de VM'de bulunan çok fazla sayıdaki öğrenme ve değişken indirgeme yöntemi arasından hangilerinin diğerlerine göre daha etkili olduğunun belirlenmesi daha sonra yapılacak uygulamalar için yol gösterici olacaktır. Bu sayede hem çok fazla öğrenme yöntemi arasından hangilerinin daha tutarlı ve kararlı sonuçlar elde ettiği belirlenecek hem de farklı koşullarda öğrenme yöntemlerinin nasıl bir performans gösterdiği ortaya çıkarılmış olacaktır. Bunun yanında öğrencileri belli bir kesme puanına göre geçti-kaldı şeklinde sınıflama etkili olan girdi değişkenlerinin önem dereceleri belirlenerek başarı üzerinde etkili olan değişkenlerin neler olduğu belirlenmiş olacaktır.

Araştırmanın Amacı ve Önemi

Araştırmanın temel amacı 15 yaş grubundaki öğrencilerinin PISA 2015 verilerinden disiplin ortamı, öğretmen desteği, sorgulamaya dayalı fen eğitimi, öğretmen merkezli eğitim, çevreye duyarlık, fenden keyif alma, fen konularına ilgi, motivasyon, özyeterlik, bilimsel inançlar, fen yaşantısı, anne-baba eğitim düzeyi, fen öğrenme süresi, okul dışı ders çalışma süresi, okula ait hissetme, kaygı, tutum, tercih ve inançlar gibi derse ve okula ilişkin bilgiler yardımıyla PISA başarılarını tahmin etmek ve süreçte kullanılan veri madenciliği sınıflama algoritmalarının güvenirlik ve geçerlik değerlerinin ne düzeyde olduğunun belirlenmesidir. Bu amaç doğrultusunda;

- Öğrencilerin PISA Fen Okuryazarlığı başarılarını tahmin etmek amacıyla başarı üzerinde anlamlı bir etkiye sahip olan değişkenlerin belirlenmesi,
- PISA başarılarını tahmin etmede kullanılan değişkenlerin önem sırası ve doğru sınıflama düzeylerinin belirlenmesi,
- Öğrencilerin PISA başarı durumlarının nasıl ve ne düzeyde tahmin edildiği,
- PISA başarılarını tahmin etmede kullanılan veri madenciliği sınıflama algoritmalarının güvenirlik ve geçerlik değerlerinin belirlenmesi amaçlanmıştır.

Bunun yanında elektronik sistemlerden hangi tür verilerin toplanacağı, ne kadar verinin ne kadar süre ile toplanacağı, verilerin nasıl depolanacağı, ne tür ön işleme süreçlerinden geçirileceği de araştırılması gereken önemli problemler olarak görülmektedir (Bienkowski ve diğerleri, 2012). Çalışma kapsamında belirtilen problemler göz önüne alınarak OECD (The Organisation for Economic Co-operation and Development) tarafından yayınlanan veri tabanlarından elde edilecek değişkenler belirlenmiştir. Bunun yanında öğrencilerden elde edilen bilgilerin büyük veri (*bigdata*) olması sebebiyle veri madenciliği yöntemleri kapsamında kullanılabileceği belirlenmiştir (Nisbet, Elder ve Miner, 2009). Büyük veri aslında elde edilen bilginin farklı ortamlardan ve farklı formatlardan elde edilmesi anlamına gelmektedir. Her ne kadar büyük veri tanımı uygulama alanına göre farklılık gösterse de eldeki verinin boyutu ve elde edildiği kaynağı büyük veri için önemli bir belirleyicidir (Vaitsis, Hervatis ve Zary, 2016). PISA 2015 sınavında öğrencilere ilişkin farklı formatlarda ve farklı kaynaklarda bilgilerin olması ve çok

sayıda öğrenciye ulaşılması sebebiyle çalışma kapsamında kullanılacak verilerin büyük veri olarak tanımlanabileceği düşünülmüştür.

Araştırma Problemi

PISA sınavına katılan öğrencilerinin Fen okuryazarlığı bakımından başarılarını tahmin etmede kullanılan veri madenciliği sınıflama algoritmalarının güvenirlik ve geçerlik değerleri ne düzeydedir?

Alt problemler. Araştırmanın ana problem cümlesine bağlı olarak oluşturulan alt problemler aşağıda listelenmiştir.

1. PISA öğrenci anketiyle ölçülen değişkenlerden yararlanarak başarıyı tahmin etmede kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge lojistik Regresyon yöntemleri yardımıyla edilen ölçme sonuçlarının 10 katlı çapraz geçerleme yöntemi yardımıyla elde edilen güvenirlik ve geçerlik değerleri ne düzeydedir?

2. PISA öğrenci anketiyle ölçülen değişkenlerden yararlanarak başarıyı tahmin etmede kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge lojistik Regresyon yöntemleri yardımıyla edilen ölçme sonuçlarının test edilen verilerin %5, %10, %15, %20, %25, %30 ve %35 olması durumunda güvenirlik ve geçerlik sonuçları ne düzeydedir?

3. PISA öğrenci anketiyle ölçülen değişkenlerden yararlanarak başarıyı tahmin etmede kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge lojistik Regresyon yöntemleri yardımıyla edilen ölçme sonuçlarının karışıklık matrisi yardımıyla elde edilen doğruluk dereceleri ne düzeydedir?

Sınırlılıklar

Bu araştırmanın sınırlılıkları aşağıda maddeler halinde sıralanmıştır.

1. Çalışma kapsamında kullanılan sınıflama ve tahmin modelleri WEKA programında var olan algoritmalar ile sınırlıdır.

2. Başarının tahmin edilmesinde kullanılacak değişkenler PISA öğrenci anketinde yer alan alt ölçekler ve sosyodemografik özellikler ile sınırlıdır.
3. Çalışmada PISA Fen okuryazarlığı başarısı olarak ortalama puan hesaplanmış ve bu puanın üstündeki öğrenciler başarılı diğerleri ise başarısız olarak sınıflandırılmıştır. PISA 2015 sınav uygulamasında fen okuryazarlığı alanında katılımcı tüm ülkelere ilişkin ortalama puan 465,00 iken ülkemiz için ortalama puan 425,00 olarak belirlenmiştir (MEB, 2016). Araştırma kapsamına test edilen modellerin dış geçerliği için farklı bir veri seti ele alınmadığından modele ilişkin sonuçlar 10k-çapraz-geçerlik yöntemi kullanılarak genelleştirilmiştir. Çapraz geçerlilik yöntemi veri madenciliği çalışmalarında oluşturulan modellerin sonuçlarının genelleştirilmesi (yeni veri setinde test edilmesinin olanaklı olmadığı durumlarda) ve dolaylı olarak örneklem sayısından kaynaklı sınırlılıkların aşılması konusunda başvurulan bir yöntemdir (Kohavi, 1995).

Tanımlar

PISA başarısı: Çalışma kapsamında ele alınan PISA 2015 fen okuryazarlığı puanlarının ortalaması hesaplanarak bu değerin altındaki öğrenciler başarısız (0), bu kesme değerine eşit ve büyük olanlar ise başarılı (1) olarak kabul edilecektir.

Okuryazarlık: Okuryazarlık kavramı, öğrencilerin temel konu alanlarındaki çeşitli durumlarda karşılaştıkları problemleri tanımlarken, yorumlarken ve çözerken, bilgi ve becerilerini kullanma, analiz etme, mantıksal çıkarımlar yapma ve etkili iletişim kurma yeterlilikleri olarak ifade edilmektedir (OECD, 2016).

Ham veri: OECD tarafından resmi internet sayfasında paylaşılan ve 72 ülkeden yaklaşık 29 milyon öğrenciyi temsilen 540.000'e yakın öğrencinin katılımıyla gerçekleşen sınavdan elde edilen işlenmemiş verilerdir.

Eğitim veri seti: Veri madenciliği (VM) çalışmalarında model oluşturma sürecinde kullanılan veri setidir.

Test veri seti: VM çalışmalarında oluşturulan modelin performansının test edildiği veri setidir.

Değişken (metrik): Toplanan ham verilerinin bir takım matematiksel dönüşümler sonucu analizlerde kullanılacak verilere dönüştürülmüş halidir.

Decision stump: Tek düzeyli karar ağacı oluşturma yöntemi.

Hoeffding tree: Örneklerin dağılımının zaman içinde değişmediğini varsayarak, büyük veri setlerinden ağaç oluşturma yapısına sahip öğrenme yöntemi.

J.48: C4.5 algoritmasına dayalı karar ağacı öğrenmesi.

LMT (Logistic Model Trees): Lojistik modele dayalı ağaç oluşturma yöntemi.

Rep tree: Azaltılmış hatayı kullanan hızlı ağaç öğrenmesi.

Random forest: Rastgele ormanlar oluşturan ağaç öğrenmesi.

Random tree: Her düğümde rastgele özelliklerin belirli bir sayısına göre ağaç oluşturma yöntemi.

Ridge lojistik regresyon: Bağımsız değişkenler arasında çoklu doğrusal bağlantı sorununu ortadan kaldıran Ridge temelli Lojistik Regresyon yöntemiyle ağaç oluşturma.

Bölüm 2
Araştırmanın Kuramsal Temeli ve İlgili Araştırmalar

Veri Madenciliği ve Makine Öğrenme

Bilgi çağının yaşandığı günümüzde modern bilgisayar sistemleri neredeyse hiç tahmin edilemeyecek bir oranda ve çok çeşitli kaynaklardan veri toplamaktadır. Ticari veya bilimsel amaçlı olarak giderek daha büyük miktarda verinin nispeten düşük maliyetle depolanmasını mümkün kılan depolama teknolojisindeki ilerlemelerin yanı sıra, bu tür verilerin içinde bulabileceğiniz bir bilginin doğru bir şekilde kullanılması büyük önem taşımaktadır (Bramer, 2013). Günlük hayatta internet üzerinden bir alış veriş yapıldığında, internette bir web sitesi ziyaret edildiğinde veya kredi kartı ile sanal bir işlem yapıldığında ister istemez veri oluşturulur. Bu veriler farklı şirketlerin sahip olduğu güçlü bilgisayarlarda depolanmakta ve bu veri setleri içinde yatan gizli örüntüler, çıkarların, alışkanlıkların ve davranışların göstergeleri olabilmektedir. VM, kullanıcıların bu kalıpları bulmalarını ve yorumlamalarını sağlayarak şirketlerin daha bilinçli kararlar almalarına ve müşterilerine daha iyi hizmet vermelerine yardımcı olmaktadır (North, 2012). Makine öğrenmesinde ilişkilerin keşfedileceği desenleri içeren birçok farklı yol bulunmaktadır ve bunların her biri veriden elde edilecek çıktıdan anlam çıkarmak için kullanılabilecek tekniğin türünü belirlemektedir. Elde edilen çıktının nasıl yorumlanacağını anladıktan sonra, bunun nasıl elde edildiğini anlamak için uzun bir yol kat etmek gerekmektedir.

Büyük veri tabanlarında depolanan veriler iş, devlet ve kar amacı gütmeyen kuruluşlar için değerli bir kaynak olabilir; ancak bu verilerin niceliğinden ziyade niteliği de oldukça önemlidir. Veri madenciliği ile bir işletmenin farklı bileşenlerinin birbirleriyle nasıl ilişkili olduğunu anlaşılabilmektedir. Bu sayede işletmenin daha düzgün çalışması ve daha fazla gelir elde etmesi için yapabilecek eylemler hakkında ipuçları sağlanmaktadır. Bu durumda veri madenciliği kuruma zarar vermeden nerede maliyetlerin düşürülebileceğinin ve en iyi getirinin nereden bulabileceğini belirlemeye yardımcı olabilir (Brown, 2014). Bilgi keşfi daha önce önemsiz olarak görülen verilerden daha önce bilinmeyen ve potansiyel olarak yararlı bilginin çıkarımı olarak tanımlanmıştır.

Karar Tabloları. Makine Öğrenmesinde elde edilen çıktının sunumunda basit ve en ilkel yöntem veri girişinde olduğu gibi birer karar tablosu elde etmektir (Witten ve Frank, 2007). Örnek olarak Tablo 1 hava durumu ile ilgili birer karar tablosudur.

Tablo 1

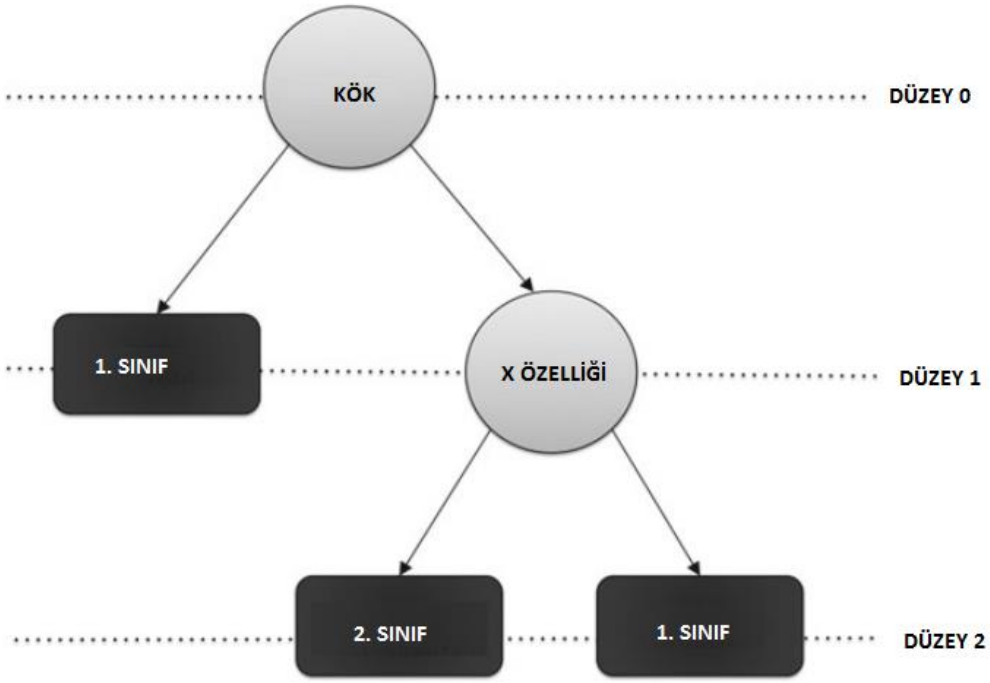
*Hava Durumu ile ilgili Karar Tablosu**

Genel Görünüş	Sıcaklık	Nem	Rüzgar	Oyun Oynama
Güneşli	Yüksek	Yüksek	Yok	Hayır
Güneşli	Yüksek	Yüksek	Var	Hayır
Bulutlu	Yüksek	Yüksek	Yok	Evet
Yağmurlu	Ilık	Yüksek	Yok	Evet
Yağmurlu	Serin	Normal	Yok	Evet
Bulutlu	Serin	Normal	Var	Evet
Güneşli	Serin	Normal	Var	Evet
Güneşli	Ilık	Yüksek	Yok	Hayır
Yağmurlu	Serin	Normal	Yok	Evet
Güneşli	Ilık	Normal	Yok	Evet
Yağmurlu	Ilık	Normal	Var	Evet
Bulutlu	Ilık	Yüksek	Var	Evet
Bulutlu	Yüksek	Normal	Yok	Evet
Yağmurlu	Ilık	Yüksek	Var	Hayır

* Witten ve Frank (2005)'dan uyarlanmıştır.

Bu tablodaki uygun koşullara bakarak oyun oynayıp oynanmayacağına karar verebilir. Daha önemsiz bir şekilde karar tablosu oluşturma bazı özelliklerin seçimini gerektirebilir. Eğer tabloda verildiği üzere sıcaklık değişkeni aldığınız karar ile ilişkili değilse ya da başka bir ifadeyle alınan kararda herhangi bir etkiye sahip değilse bu özelliklerin yer almadığı yoğunlaştırılmış bir karar tablosudaha yönlendirici olabilir. Bu aşamada esas problem nihai kararı etkilemeden hangi özelliğin çıkarılacağına karar verilmesidir.

Karar Ağaçları. Bağımsız örnek durumlar kümesinden "böl ve yönet" yaklaşımı ile ilgili bilgi elde etme çabası araştırmacıları doğal olarak "karar ağacı" adı verilen bilgi sunma yöntemine götürmektedir. Şekil 1'de X özelliğine göre kök düğümden (node) başlayarak nasıl bir sınıflama yapılacağı ile ilgili yapısal bir gösterim olan karar ağacı verilmektedir. Bu ağaç sayesinde daha özet ve daha çekici bilgi sunumu yapılabilmektedir (Barros vd., 2015).



Şekil 1. Sınıflama için genel bir karar ağacı örneği.

Ağaçta ilk olarak düzey 0 olarak verilen başlangıç noktasında kök düğüm ile başlanmaktadır. Ağaçta yer alan çemberler kök ve iç düğümleri, kareler ise yaprak düğümlerini belirtir. Bu özel örnekte, karar ağacı sınıflandırma için tasarlanmıştır ve bu nedenle yaprak düğümleri sınıf etiketleri barındırmaktadır. Düzey 1'de herhangi bir X özelliği için bu özelliğin belli bir eşik değerine göre örnekler iki sınıfa ayrılmaktadır. Ağacın sol tarafında görüldüğü 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. Benzer şekilde bu testler devam ederek ağacın dallarının oluştuğu görülmektedir. Karar ağacında yer alan düğümler belirli bir özelliği test etmeyi amaçlamaktadır. Genellikle düğümler üzerinde gerçekleştirilen testler, bir özelliğe ilişkin gözlenen değerin sabit bir değer ile karşılaştırması şeklinde yapılmaktadır. Buna rağmen bazı ağaçlar iki özelliği birbiriyle karşılaştırır veya bir ya da daha fazla özellik için bazı fonksiyonlar kullanırlar. Yaprak düğümler o yaprağa ulaşıncaya kadarki tüm örnek durumlar için bir sınıflama veya sınıflama kümesi ve olası tüm sınıflamalara ilişkin olasılık dağılımlarını vermektedir. Bilinmeyen bir örneği sınıflamak için, ardışık düğümlerde test edilen özelliklerin değerlerine göre ağacın alt dallarına yönlendirme yapılır ve bir yaprağa ulaşıldığında, örnekler yaprağa atanan sınıfa göre sınıflandırılmış olacaktır (Witten ve Frank, 2007).

Eğer düğümde test edilen özellik sınıflamalı ölçek düzeyinde ise, alt düğümlerin (child node) sayısı genellikle özelliğin pozitif değerlerinin sayısına eşit olacaktır. Bu durumda, olası her bir değer için bir dallanma olacaktır ve aynı özellik ağacın alt dallarında tekrardan test edilmeyecektir. Bazen özellik değerleri iki alt kümeye ayrılmakta ve ağacın dalları özellik değerlerinin hangi alt kümeye bağlı olma durumuna göre sadece iki şekilde dallanmaktadır. Bu durumda, belirtilen özellik dallanma boyunca birden fazla sayıda test edilebilmektedir (Maimon ve Rokach, 2010).

Eğer özellik aralık veya oran ölçeği düzeyinde ise düğümde yapılan test genellikle önceden belirlenmiş tahmini değerden büyük ya da küçük olma durumuna göre iki yönlü bir ayrım yapmaktadır. Alternatif olarak farklı olasılıkların bulunduğu üç yönlü bir bölünme kullanılabilmektedir. Eğer kayıp veriler kendi başına bir özellik olarak ele alınırsa, bu durumda karar ağacında üçüncü bir dal oluşacaktır. Bir özelliğin tam sayı değerine ilişkin o değerden küçük, büyük veya eşit olacak şekilde üç yönlü bir bölme de (dallanma, sınıflanma) yapılabilmektedir. Buna ek olarak özelliğin gerçek değeri yerine o değeri içine alacak şekilde belirlenen bir aralık için; aralığın altında, aralığın içinde ve aralığın üzerinde olacak şekilde üç yönlü bir sınıflama yapılabilmektedir. Sayısal bir nitelik genellikle, verilen bir dal boyunca kökten yaprağa doğru birkaç kez test edilmektedir ve her test farklı bir sabit içermektedir (Witten ve Frank, 2007).

Çalışmalarda kayıp veriler bariz bir problem oluşturmaktadır. Bir düğüm test edilirken hangi dallanmada hangi değerin kayıp veriye ilişkin olduğunu belirlemek kolay değildir. Bazen de eksik verilerin kendisi tek başına bir özellik olarak kabul edilebilmektedir. Eğer durum böyle değilse, eksik veriler özelliğin olası bir değeri olarak görmek yerine özel bir şekilde ele alınması gerekmektedir. Eksik veri için basit bir çözüm eğitim setindeki (training set) dallara inen örneklerin sayısını kaydetmek ve test edilen durum eksik veriye sahipse en popüler dalı kullanmaktır. Daha karmaşık bir çözüm ise örneği kavramsal olarak parçalara ayırmak ve bir kısmını da altına ve oradan da sağdan (0) aşağı doğru (1) var olan alt ağaçlara göndermektir. Bölme işlemi sıfır ile bir arasında değişen sayısal olarak ağırlıklandırılmış bir değer kullanılarak gerçekleştirilir ve her biri için ağırlık o daldan aşağı doğru inen eğitim setindeki örneklerin sayısıyla orantılı olacak şekilde seçilir. Tüm ağırlıklar toplam bir tanedir. Ağırlıklı bir örnek, belki daha alt

düğümlerde tekrardan düğüme ulaşacak ve bu yaprak düğümündeki kararlar yapraklardan aşağı inmiş ağırlıklar kullanılarak yeniden birleşebilmektedir.

Bir veri seti için elle bir karar ağacı oluşturmak eğitici hatta eğlenceli olabilmektedir. Bunu etkili bir şekilde yapabilmek için, elinizdeki veriyi görselleştirebilmek amacıyla test edilecek en iyi özelliklerin hangisi olabileceğine karar vermek gerekmektedir. Bu tez kapsamında kullanılan WEKA Programı kullanıcılara etkileşimli bir şekilde karar ağacı elde etmelerini sağlayan farklı algoritmalara sahiptir (Witten, Frank ve Hall, 2016).

Sınıflama Kuralları. Sınıflama kuralları karar ağaçlarının popüler olan alternatifleridir. Sınıflandırma, yeni verinin sınıf etiketini tahmin etmek amacıyla ileride kullanılabilecek bir model oluşturmak için kullanılan bir veri analizi formudur. Çeşitli sınıflandırma uygulamaları sayesinde sahteciliği algılama, hedef pazar belirleme, performans tahmini, imalat sayısını belirleme ve tıbbi erken tanı gibi birçok uygulama da kullanılabilmektedir. Bu süreçte ilk adım öğrenme adımıdır ve bir sınıflandırma algoritması, veritabanında yer alan örneklerden ve bunlarla ilişkili sınıf etiketlerinden oluşan bir eğitim setini analiz ederek sınıflandırıcıyı oluşturur. İkinci adımda ise sınıflandırıcının doğruluğu tahmin edilmektedir. Bunun için, eğitim setleri dışında başka bir veri seti “test seti” olarak adlandırılmaktadır ve daha sonra bu test seti üzerinden sınıflandırıcı test edilmektedir (Chadha ve Singh, 2012). Bir karar ağacından doğrudan bir dizi kural okumak oldukça basittir. Her yaprak için bir kural oluşturulur. Kuralın öncüsü kökten yaprağa kadarki yol üzerinde yer alan her bir düğüm için bir koşul içerir ve kuralın sonucu yaprak tarafından atanan sınıftır. Uygulanan bu işlem ilişkisiz durumların çıkarılabilmesi için belirsizliğe yer vermeyen kesin kurallar oluşturmaktadır. Buna rağmen, genel olarak karar ağacından elde edilen kurallar ihtiyacınızdan daha karmaşık bir yapıdadır ve ağaçlardan elde edilen kurallar genellikle gereksiz testleri ortadan kaldırmak için budanmış (*pruned*) olarak karşımıza gelirler.

Tablo 2'de bir deponun müşterilerinin kredi durumu, müşteri durumu ve depodaki stok durumuna göre taleplerini kabul, beklemede ve ret şeklinde sınıflandırmasına ilişkin veri dosyasındaki müşteri durumları incelenerek elde edilen beş farklı karar gösterilmektedir.

Tablo 2

Karar Tablosu Örneği

K1. Eğer kredi durumu=iyi ve stok= yeterli	ise	sipariş=kabul
K2. Eğer kredi durumu=iyi ve stok= yetersiz	ise	sipariş=beklemede
K3. Eğer kredi durumu=kötü ve müşteri durumu = iyi ve stok=yeterli	ise	sipariş=kabul
K4. Eğer kredi durumu=kötü ve müşteri durumu = iyi ve stok=yetersiz	ise	sipariş=beklemede
K5. Eğer kredi durumu=kötü ve müşteri durumu = kötü	ise	sipariş=ret

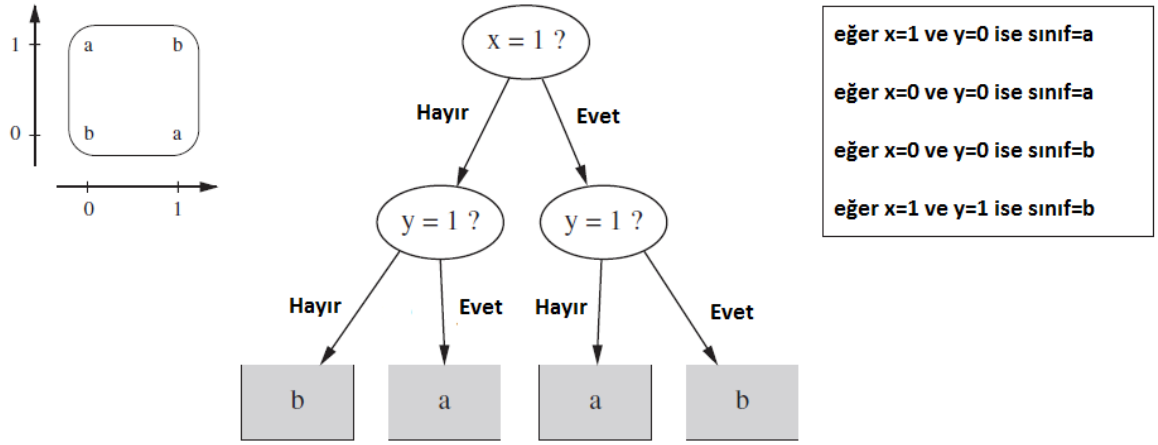
* Metzner ve Barnes (2014)'den uyarlanmıştır.

Verilen kurallar ilk önce birinci satır, sonra ikinci ve sırasıyla son satır olacak şekilde yorumlanmaktadır. Kuralların yer aldığı bu satırlar genelde "Karar Listesi olarak" isimlendirilmektedir. Bu listeler, karar tablosunda yer alan durumların tamamını doğru bir şekilde sınıflaması için gerekli kuralları göstermektedir (Metzner ve Barnes, 2014).

Bu kuralların öncülü veya önkoşulu karar ağacındaki düğümlerde yapılan testler gibi bir dizi testten elde edilen kuraldır ve sonuçta kuralın (örnek veri) uygun olması sonucunda elde edilen sınıfı veya sınıfları verir, belki de sınıflar üzerindeki olasılık dağılımları hakkında bilgi verir. Genel olarak, koşullar mantıksal olarak birlikte sonlandırılmakta ve eğer kural uygulamaya konulacaksa tüm testlerin başarılı bir şekilde sonuçlanması gerekmektedir. Buna rağmen, bazı kural formülasyonların da ön şartlar basit bağlayıcılar yerine genel mantıksal ifadelerdir. Genellikle, bireysel kuralların mantıksal olarak "OR" bağlacıyla daha etkili olduğu düşünülmektedir. Eğer bu kuralı herhangi biri uygularsa sonuçta verilen sınıf (ya da olasılık dağılımı) örneğe uygulanmaktadır. Bununla birlikte, farklı sonuçlara sahip farklı kurallar uygulandığında çelişkiler ortaya çıkabilmektedir (Nalepa, 2008). Karar ağaçları bir kümede uygulanan farklı kurallar arasındaki ayrılığı kolaylıkla ifade edemediği için, genel kurallar kümesini bir ağaca dönüştürmek o kadar da basit değildir. Bu durumun ortaya çıkmasına güzel bir örnek, aynı yapının farklı özelliklere sahip bir kural uygulandığında oluşmasıdır.

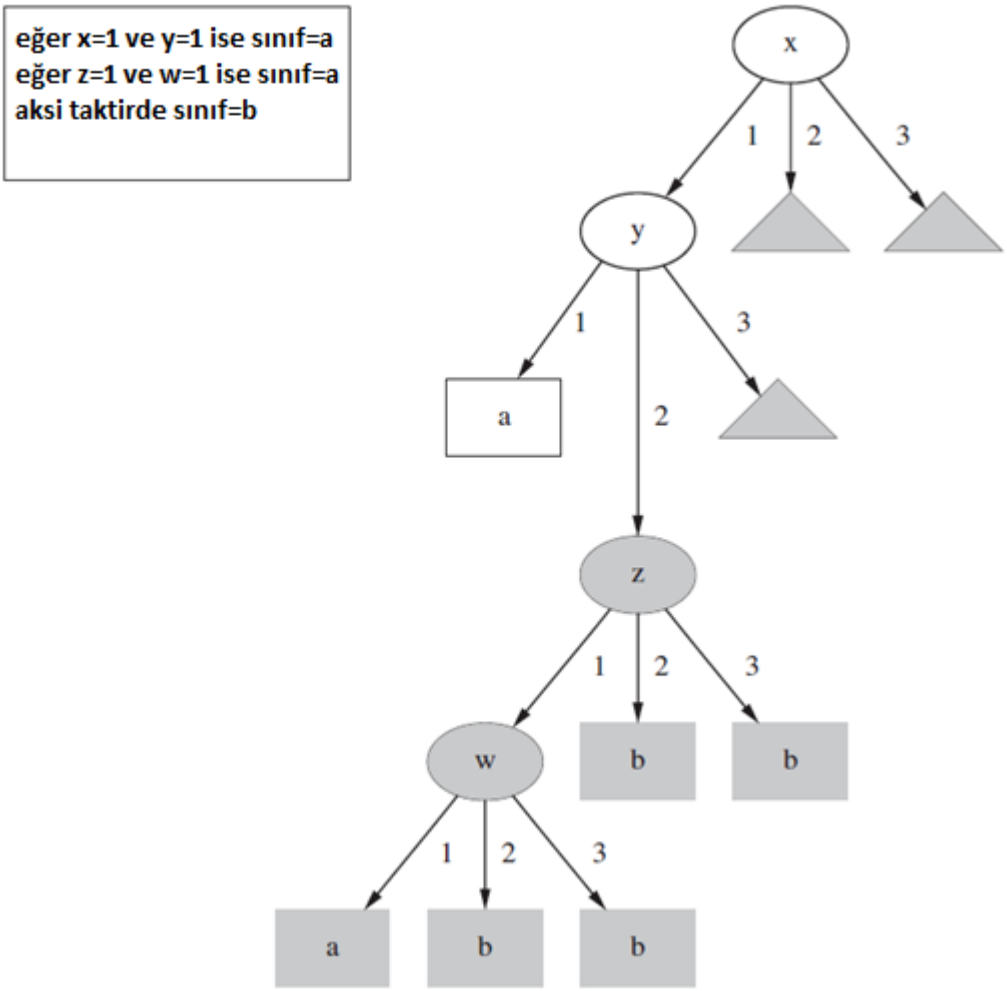
Şekil 2'de verilen diyagramın sol tarafı çıktı sınıfı a ise x=1 veya y=1 ancak ikisi aynı anda olmamak üzere "or" bağlacı ile yapılmış özel bir fonksiyonu göstermektedir. Bunu bir ağaç haline getirmek için, önce bir özelliği şeklin ortasında gösterildiği gibi ikiye bölmek gerekmektedir. Buna karşın, kurallar şeklin

sağ tarafında görüldüğü gibi incelenen özellikle ilişkili problemin gerçek simetrisini gerçekçi bir şekilde yansıtmaktadır (Witten ve Frank, 2007).



Şekil 2. Problem durumu - özel durum.

Bu örnekte verilen kurallar ağaçtakilerden daha anlaşılır değildir. Aslında bunlar ağaç olmadan kuralları okuyarak ne elde ettiğinizi göstermektedir. Fakat başka durumlarda, kurallar ağaç gösteriminden daha karmaşıktır ve özellikle de diğer kurallar tarafından belirtilmeyen durumları kapsayan "varsayılan bir kural" olmadığı durumlarda kurallar ağaçlardan daha karmaşık olacaklardır. Örneğin, Şekil 3'de gösterilen kuralların etkisini görebilmek için x,y,z ve w olarak verilen 4 özellik ile 1,2,3 olarak gösterilen alt özellikleri inceleyelim. Şeklin sol tarafında x,y,z,w özellikleri için birimlerin hangi sınıflara dahil olabileceklerine ilişkin kurallar verilmiştir. Gri renkte gösterilen ve şeklin sağ üst tarafında verilen üçgenler, tekrarlanan alt ağaç problemine birer örnek durumundadır. Bu durum aslında basit bir kavramın ne yazık ki daha karmaşık bir şekilde gösterimine örnektir.



Şekil 3. Tekrarlı alt ağaçları olan karar ağacı.

Kuralların popüler olma sebeplerinden bir tanesi her kuralın birer bağımsız bilgi külçesi olarak görülmesidir. Külçe ifadesi kuralın eldeki veri hakkında oldukça değerli bilgiler içermesi sebebiyle kullanılmaktadır. Daha önceden var olan kurallara herhangi bir etkide bulunmadan yeni kurallar eklenebilmektedir. Buna rağmen, kurallar için belirtilen bu bağımsızlık durumu bir çeşit yanılsamadır; çünkü eklenen yeni kural "kural kümesinin" aslında nasıl yürütüldüğü sorusunu göz ardı etmektir. Ana mantığın dışında bağlamdan bağımsız olarak eklenen bir kural yanlış olabilmektedir. Öte yandan eğer yorum sırası önemsiz olarak görülürse, aynı örnek için farklı sonuçlara sebep olan farklı kurallarla karşılaşıldığında ne yapılacağı açık olmayacaktır. Bu durum doğrudan bir karar ağacından okunan kurallar için ortaya çıkmaz; çünkü kuralların yapısında yer alan bu gereksiz fazlalık yorumlamada yapılabilecek belirsizlikleri ortadan kaldırmaktadır. Fakat belirtilen bu

durum kuralların başka yollardan elde edildiği zaman ortaya çıkmaktadır (Elomaa ve Kaariainen, 2001).

Eğer bir "kural kümesi" belirli bir örnek durum için çoklu sınıflama sonuçları üretirse, problemin çözüm yöntemlerinden biri hiçbir sonuç vermemektedir. Bir diğer çözüm yöntemi ise her bir kuralın eğitim verisi üzerinde ne kadar kayıp verdiğini belirlemek ve bu hatalardan en popüler olan bir tanesine odaklanmaktır. Belirtilen bu yöntemler tamamiyle farklı sonuçlar doğurmaktadır.

Tekrardan vurgulamak gerekirse bu sorunlar karar ağaçlarında veya onlardan doğrudan okunan kurallarda ortaya çıkmazlar, fakat bu sorunlar genel kural kümelerinde kolaylıkla ortaya çıkabilirler. Bu sorunlarla başa çıkabilmenin bir yolu böyle bir örneği sınıflandırmamaktır. Diğer yolu ise en çok tekrarlanan sınıfa örneği varsayılan sınıf olarak atamaktır. Belirtilen bu stratejiler için tamamiyle farklı sonuçlar elde edilebilir. Bireysel ya da tek bir örneğe dayalı kurallar basittir ve buna bağlı olarak geliştirilen kural kümeleri de aldatıcı bir şekilde basit görünebilir fakat ek bilgi verilmeden doğrudan bir kural kümesi verilmiş ise bunun nasıl yorumlanacağı açık değildir. Özellikle mantıksal (Boolean) bir sınıflamada (ör.Evet-Hayır) kullanılan kurallar tüm örnekleri sadece tek bir çıktıya (ör.Evet) yönlendiriyorsa daha önceden belirtilen sorunlar ortaya çıkmaktadır. Eğer belirli bir örnek "Evet" sınıfında değilse, belirtilen örneğin bu kümenin evrendeki tümleyeni anlamına gelen "Hayır" sınıfında olması gerekmektedir. Eğer durum böyleyse, kurallar çatışmaz ve kuralın yorumunda herhangi bir belirsizlik olmayacaktır: Böyle bir kural kümesi "Ayırıcı Normal Form" olarak isimlendirilen veya “OR” şeklinde uyarıcı ya da “AND” şeklinde birleştirici fonksiyonlar yardımıyla yazılabilir (Tan, Stainbech ve Kumar, 2005).

Bu verilen mantıksal ifadeler kurallarla başa çıkmayı oldukça kolay olarak gösteren özel durumlardır. Burada her kural her bir bağımsız bilgi parçasını evet veya hayır şeklinde ayrıştırmak için doğrudan izlenen yolları göstermektedir. Ne yazık ki bunlar sadece kategorilere ayrılmış çıktılar için geçerlidir ve evrenin sınırlarının belli olmasını yani belirtilen kümenin tümleyeninin var olması varsayımını sağlanması gerekir ki pratikte bu kısıtlamalar çoğu durumda gerçekçi değildir (Quinlan, 1993).

Birliktelik Kuralları. Birliktelik kuralları evrende yer alan örnekleri sınıflamanın yanısıra araştırmacılara farklı özellik kombinasyonlarını da tahmin edebilme özgürlüğü sunmaktadır. Bunun yanında, birliktelik kuralları sınıflama kurallarında olduğu gibi bir dizi kural kümesinin birlikte kullanılmasını da gerektirmezler. Farklı ilişkilendirme kuralları veri kümesinde yer alan farklı düzenlemeleri ifade eder ve genellikle farklı şeyleri tahmin etmektedirler (Jain, Tiwari ve Ramaiya, 2013).

İlişkilendirmeli sınıflama olarak da bilinen birliktelik kuralları VM'nin bir koludur. Bu yöntemde tahmin amacıyla bir model oluşturmak için sınıflama ve ilişki kuralı belirleme yaklaşımları bir arada kullanılır (Abdeljaber, 2007). Her ne kadar sınıflandırma ve ilişkilendirme kuralı keşfi veri madenciliğinde benzer görevler olarak görülse de sınıflandırmanın temel amacının sınıf etiketlerinin tahmini iken, ilişki kuralı keşfinde temel amaç bir veri dosyasındaki öğeler arasındaki ilişkileri ortaya çıkarmaktır (Yin & Han, 2003).

Çok sayıda farklı birliktelik kuralları küçük bir veri kümesinden bile türetilebildiği için bunların oldukça fazla sayıda örneğe uygulanabilmesi konusundaki ilgi kısıtlıdır ve uygulandıkları örneklerde yüksek doğruluk (accuracy) değerlerine sahiptirler. Bir birliktelik kuralının kapsamı tahmin ettiği örneklerin sayısıdır ve buna genellikle destek (support) adı verilir. Birliktelik kurallarının "doğruluğu" genellikle güvenirliği (confidence) olarak isimlendirilir ve bu doğru olarak sınıflanan örneklerin sayısına eşittir. Daha açık bir ifadeyle uygulama kapsamındaki tüm örneklerin içinde doğru olarak sınıflananların oranı olarak ifade edilebilir. Örneğin daha önce verilen oyun oynama ile ilgili veri dosyasındaki kural*:*

eğer hava sıcaklığı= serin sonrasında nem= normal ise

Yukarıda verilen bu kural için kapsamı hem serin hem de normal nem değerine sahip günlerin sayısıdır. Daha önce verilen Tablo 1 incelendiğinde bu değer tablonun 5,6,7 ve 9. satırlarında gösterilen toplam güne karşılık gelmektedir. Doğruluk ise serin hava sıcaklığına sahip günlerin normal nem değerine sahip oranıdır ki bu oran verilen örnekler için %100 olarak hesaplanır. Hava durumuyla ilgili verilen örnekte kapsamı en az ve doğruluğu %95 olan 58 kural bulunmaktadır.

Birden fazla sonucu öngören birliktelik kurallarının oldukça dikkatli bir şekilde yorumlanması gerekmektedir. Örneğin Tablo 1'de verilen hava durumu ile ilgili örnekteki kural olan

eğer rüzgar=yok ve oyun=oynanmamış ise genel görünüş=güneşli ve nem=yüksek

kuralı aşağıda verilen iki kuralın kısaltılmış bir ifadesi değildir.

eğer rüzgar=yok ve oyun=oynanmamış ise genel görünüş=güneşli

eğer rüzgar=yok ve oyun=oynanmamış ise nem=yüksek

Bu aslında verilen kuralın minimum kapsam ve doğruluk rakamlarını aştığını ima etmektedir. Fakat aslında burada daha farklı anlamlarda bulunmaktadır. İlk verilen birleşik kuralın anlamı rüzgarlı olmayan, oyun oynanmayan günlerdeki havanın güneşli ve nemin yüksek olduğu örnek sayısıdır. Bu sayı belirtilen en az kapsama sayısı ile aynı büyüklüktedir. Bu ayrıca rüzgarlı olmayan, oyun oynanmayan günlerin oranına ifade edilen bu şekilde günlerin en azından belirtilen minimum doğruluk sayısı olacağını göstermektedir. Bu ayrıca,

eğer nem=yüksek ve rüzgar=yok ve oyun=oynanmamış ise genel görünüş=güneşli

şeklinde verilen nem yüksek, rüzgar yok ve oyun oynanmamış ise havanın güneşli olacağı şeklindeki verilen kuralla aynı anlama gelmektedir. Çünkü orijinal kuraldaki ile aynı kapsama sahiptir ve onun doğruluğuyla orijinal kural ile aynı olacaktır çünkü kuralda verilen yüksek nem, rüzgarlı olmama ve oyun oynanmayan gün sayıları rüzgarlı olmama ve oyun oynanmayan gün sayılarından daha azdır ve bu durum doğruluk değerinin daha da artmasına sebep olur (Freitas, 2003).

Gördüğünüz gibi, belirli birliktelik kuralları arasında bazı kuralların diğerlerini ima etmesi gibi bağlantılar olabilir. Üretilen kuralların sayısını azaltmak için, birkaç kuralın ilişkili olduğu durumlarda kullanıcıya yalnızca en güçlü olanı sunmak mantıklıdır. Buna göre iki farklı kural yerine daha fazla kapsamı olan tek bir kuralın kullanılması daha anlamlı olacaktır (Stefanowski, 2008).

Yeni verinin sınıfını tahmin etme. Eğer karar tablosunda tanımlanan örneklerden veri dosyasında olmayan yeni bir durum ile karşılaşıldığı zaman bazı ilave kuralların oluşturulması gerekebilir. Verilen bu kuralların yeni örneği doğru bir

şekilde tanımlayabilmesi için yeniden düzenlenmesi gerekmektedir. Bununla birlikte kural kümesi içinde özellik değerinin sınırlarında yapacağınız basit değişiklikler kural kümesi oluşturmak amacıyla kullanılan örneklerin yanlış sınıflanmasına sebep olacağı için bu değişiklikler bazen yeterli olmayabilmektedir. Bu sebeple bir kural kümesinde değişiklik yapmak sanıldığı kadar basit değildir (Witten ve Frank, 2005). Hâlihazırda kurallarda var olan testleri değiştirmek yerine, uzmanlardan eklenen bu yeni kuralı ana kuralı neden ihlal ettiğine ilişkin görüş alınabilir ve sadece ilgili kuralları genişletmek için kullanılabilecek ilave açıklamalar tanımlanabilir.

Temel Yöntemler ve Algoritmalar

Veri madenciliğinde kullanılan algoritma ve öğrenme yöntemlerinin tamamı sağlam istatistiksel temelleri olan çoklu mantıksal sınama modellerine dayanmaktadır (Mease ve Wyner, 2008). Veri madenciliği yöntemlerinin denetimsiz (unsupervised), yarı denetimli (semi- supervised) ve denetimli (supervised) öğrenme yaklaşımı olmak üzere üç türü bulunmaktadır.

Denetimli öğrenmede algoritma, etiketleri bilinen bir dizi örnekle çalışır. Etiketler, sınıflandırma görevi için nominal değerler veya regresyon durumunda ise sayısal değerler olabilir. Denetimsiz öğrenmede, aksine, veri kümesindeki örneklerin etiketleri bilinmemektedir ve algoritma, tipik olarak örnekleri, kümeleme görevini karakterize eden öznitelik değerlerinin benzerliğine göre gruplandırmayı amaçlamaktadır. Son olarak, yarı denetimli öğrenmede ise etiketli örneklerin küçük bir alt kümesi mevcutken bunları çok sayıda etiketlenmemiş örnekle birlikte kullanmayı esas almaktadır (Neelamegam ve Ramaraj, 2013). Karar ağacı, en yakın komşuluk, destek vektör makinesi, Naive Bayes sınıflandırıcısı ve yapay sinir ağları temel sınıflandırma yöntemleri arasında yer almaktadır ve bunlar denetimli öğrenme yaklaşımlarıdır. Ancak küme analizi algoritmaları, bir sınıftaki nesne sayısının veya sınıf sayısının belirtilmesini gerektirmediği için denetimsiz öğrenme yaklaşımı olarak kabul edilmektedir (Kusiak, 2001). Bunun yanında Wu ve diğerleri (2008) VM'de en popüler 10 algortimanın C4.5, k-ortalama yöntemi, destek vektör makineleri, önsel dağılımlar (*apriori*), maksimum beklenti (*EM*), sayfa derecesi (*PageRank*), k-en yakın komşuluk, Naive Bayes ile Sınıflama ve Resgresyon Ağacı (*CART*) olduğunu belirtmektedir. Bu yöntemlerden C4.5

algortiması ID3 yönteminin sınırlılıklarını ortadan kaldırmak için geliştirilmiş bir versiyonu olarak düşünülebilir (Hssina, Merbouha, Ezzikouri ve Erritali, 2014).

Veri seti üzerinde kullanılan algoritma ve yöntemler farklı şekillerde ortaya çıkabilmektedir. Örneğin bir veri setinde, tek bir özelliğin çok iyi çalıştığı diğerlerinin ilişkisiz veya gereksiz olduğu örnekler olabilir. Başka bir veri setinde, özelliklerin birbirinden bağımsız ve çıktı değişkenine eşit bir şekilde katkı sağladığı durumlar olabilir. Üçüncü bir veri setinde basit bir mantıksal yapıya sahip ve karar ağacı tarafından elde edilen birkaç özelliğin yer aldığı durumlar olabilir. Dördüncü bir veri setinde, örnekleri farklı sınıflara atayan birkaç bağımsız kural kümesi olabilir. Beşinci bir veri setinde özelliklerin farklı alt kümeleri arasındaki birbirine bağımlı olma durumları incelenebilir. Altıncı bir veri setinde uygun ağırlıklandırma yönetimiyle belirlenmiş sayısal özelliklerin ağırlıklı toplamları arasındaki doğrusal bağımlılık incelenebilir. Yedinci bir veri setinde örneklerin birbirleri arasındaki mesafeleri esas alarak örnek uzayın belirli bölgelerine bu örnekleri yerleştirmek olabilir. Sekizinci bir veri setinde, öğrenmenin denetimsiz olduğu algoritmalar da görüldüğü gibi hiçbir sınıf değeri olmayabilir (Witten ve Frank, 2007). Olası veri setlerinin sonsuz çeşitliliğinde farklı yapıların ve farklı veri madenciliği araçlarının kullanıldığı farklı yapıda örnekler bulunmaktadır. Bu örneklerde farklı bir türün uygunluğunu tamamen gözden kaçırabilen ve sadece tek bir sınıfa yönelik doğru sınıflamalar yapan algoritmalar da bulunabilmektedir. Sonuç basit, şık ve hemen anlaşılabilir bir yapının yerine gösterişli (baraque) ve monoton (opaque) bir sınıflama olacaktır. Bu durum bahsedilen sekiz farklı veri seti bünyesinde yer alan bilginin keşfedilmesi için farklı bir makine öğrenme yönteminin kullanılmasına sebep olmaktadır.

Basit kurallardan anlam çıkarma. Makine öğrenme yaklaşımında kullanılacak önemli yöntemlerden biri de kural kümelerinden anlam çıkarma yaklaşımıdır. Kural temelli modellerin öğrenilmesi, 1960'ların başında başından bugüne kadar makine öğrenme alanında ana araştırma konularından biri olmuştur (Friedman & Fisher, 1999). Bu yaklaşımın temelinde “if-then” kuralı şeklinde ifade edilebilen verilerde bir düzenlilik bulma amacı bulunmaktadır. (Fürnkranz ve Klieger, 2015). Her kuralda kuralın koşulu kısmında bir öznitelik değeri (özellik olarak adlandırılacaktır) ve sonuçta ise kuraldaki bir sınıf etiketi bulunmaktadır. Bu türdeki mantıksal sınama kurallarına alternatif olarak, olasılığa dayalı kurallar da

oluşturulabilmektedir. Bu kurallarda öngörülen sınıf etiketine ek olarak, “then” ifadesinden sonra bu kuralın sonucu olarak eğitim veri setindeki örneklerin bu sınıfta bulunma olasılıkları yer alabilir (Clark & Boswell, 1991). Örnek bir kuralda if durumu=bekar ve cinsiyet=kadın ise sonuç=evet şeklinde bir ifade medeni durumu bekar ve cinsiyeti kadın ise sonucun evet şeklinde sınıflanacağını göstermektedir.

Kurallar birbirlerinden bağımsız oldukları için, herhangi bir kuralda değişiklik yapıldığında diğer kuralları etkilemeksizin ilgili kuralın kapsamı kolaylıkla genişletilebilmektedir (Kurgan, Cios ve Dick, 2006). Bir karar ağacı tarafından oluşturulan kurallar ile bir kural öğrenme algoritması tarafından üretilen kurallar arasındaki ana farklılık, eski kural kümesinin veri setindeki tüm örnekleri kapsayan ve birbiriyle çakışmayan kurallardan oluşmasıdır (Fürnkranz, 1999).

Bir örnekler kümesinden oldukça basit sınıflama kuralları bulmanın basit bir yolu bulunmaktadır. Bu kurallar 1R (1Rule) olarak simgelenmekte ve sadece belirlenmiş tek bir özelliği test eden kuralların kümesi şeklinde ifade edilen bir düzeyli karar ağacı oluşturmaktadır. 1R, veri yapısını karakterize etmek için oldukça iyi kurallar üreten basit ve oldukça ucuz bir yöntemdir. Bu basit kurallar sık sık şaşırtıcı şekilde yüksek doğruluk derecesinde sonuçlar vermektedir. Bunun sebebi belki de gerçek dünyadan elde edilen veri setlerinin altında yatan yapının oldukça basit ya da ilkel olması ve bir örneğin sınıfını doğru olarak belirleyebilmek için sadece bir özelliğin yeterli olmasıdır. Her durumda en basit olan şeyi ilk önce denemek her zaman iyi bir yöntem olarak düşünülmektedir. Buradaki temel fikir tek bir özelliği test eden kurallar oluşturup buna göre örnekleri sınıflara (şube, branş, dallara) ayırmaktır. Daha sonra elde edilen dallar özelliğin farklı bir değerine karşılık gelir. Her bir dala ayrılacak örnekler için en iyi sınıflamanın hangisi olduğu açıktır. Bunun için eğitim verisinde en çok ortaya çıkan sınıfı kullanmanız yeterli olacaktır. Sonrasında kuralların hata oranları kolaylıkla belirlenebilmektedir. Sadece eğitim verisinde ortaya çıkan hata değeri çoğunluğun girdiği sınıfa girmeyen örneklerin sayısı olarak belirlenmektedir (Witten, Frank ve Hall, 2016).

1R Kuralının nasıl çalıştığını anlamak için daha önce verilen hava durumu ile ilgili veri dosyasını göz önüne alalım. Tablo 1’in son sütununda sınıflamaya ilişkin sonuç yer almakta ve bu sütun "Oyun Oynanma" (Play) olarak isimlendirilmiştir. 1R kuralı her bir özellik için bir tane olmak üzere toplam 4 tane kural kümesi olduğunu kabul etmektir. Tablo 3'ün birinci sütunun da bu 4 kural

kümesi gösterilmektedir. Oyun oynama durumu genel görünüş, sıcaklık, nem ve rüzgar olmak üzere 4 farklı özellik yardımıyla tahmin edilmeye çalışıldığından ortaya dört farklı kural kümesi çıktığına dikkat ediniz.

Tablo 3

*Hava Durumu ile İlgili Veri dosyasındaki Özellikleri Değerlendirme**

ÖZELLİKLER		KURALLAR	HATA ORANI	TOPLAM ORAN
1	Genel Görünüş	Güneşli→Hayır (Oynanmamış)	2/5	4/14
		Bulutlu→Evet (Oynanmış)	0/4	
		Yağmurlu→Evet (Oynanmış)	2/5	
2	Sıcaklık	Yüksek→Hayır* (Oynanmamış)	2/4	5/14
		Ilık→Evet (Oynanmış)	2/6	
		Serin→Evet (Oynanmış)	1/6	
3	Nem	Yüksek→Hayır (Oynanmamış)	3/7	4/14
		Normal→Evet (Oynanmamış)	1/7	
4	Rüzgar	Yok→Evet (Oynanmış)	2/8	5/14
		Var→Hayır* (Oynanmamış)	3/6	

* Witten ve Frank (2005)'dan uyarlanmıştır.

1R kuralının çalıştığını görmek için, daha önceki bölümlerde verilen Tablo 1'i göz önüne alalım. Oyun oynama durumuna ilişkin sonuç çıktı değişkenini Evet=Oynandı ve Hayır=Oynanmadı şeklinde sınıflandırmak amacıyla her bir özellik (genel görünüşü, sıcaklık, nem ve rüzgar) için 4 tane kural kümesi dikkate alınacaktır. Tablonun alt kısmında verilen yıldız işareti sonuç (çıktı) değişkeni için olaylara ilişkin olasılık değerlerinin eşit olması durumunda bu olaylardan bir tanesinin rastgele seçildiğini göstermektedir. Tek bir kural için verilen hata oranı bu kural kümesi için tüm kurallar göz önüne alındığında elde edilecek toplam hata oranı ile birlikte verilmektedir. 1R kuralında en küçük hata oranına sahip özellikler seçilmektedir. Tablo 3'deki Genel Görünüş ve Nem özellikleri en düşük hata oranına (4/14) sahiptirler. Şaşırtıcı bir şekilde, görünürde hava bulutlu ve yağmurlu olduğunda oyun oynandığı ancak hava güneşli olduğunda oynanmadığı belirlenmiştir. Belki de bu durumun oluşmasında doğrudan veri dosyasında göremediğiniz başka sebepler bulunmaktadır (Witten ve Frank, 2005).

İstatistiksel modelleme. 1R Kuralı kararlarında ve seçimlerinde en iyi sonucu veren bir özelliği temel özellik olarak kullanmaktadır. Diğer basit bir teknik

ise tüm özellikleri kullanmak ve bunların belirlenen sınıf göz önüne alındığında eşit derecede önemli ve birbirinden bağımsız şekilde karar alma sürecine katkı sağlamalarına izin vermektir. Elbette bu yaklaşım gerçekçi değildir çünkü günlük hayatta elde ettiğiniz veriler için belirlenen özelliklerin eşit derecede önemli veya birbirinden bağımsız olması pek mümkün değildir. Fakat bu düşünce bizi pratikte şaşırtıcı şekilde iyi işleyen bir şemaya götürür. Tablo 4 daha önce verilen hava durumu örneğiyle ilgili olarak oyun oynama durumunun kaç defa gerçekleştiğine ilişkin (Evet ve Hayır) özetleyici bilgiler sunmaktadır.

Tablo 4

Hava Durumu Verisinin Sayısal Olarak Betimlenmesi

Genel Görünüş			Sıcaklık			Nem			Rüzgar			Oyun Oynama	
	Evet	Hayır		Evet	Hayır		Evet	Hayır		Evet	Hayır	Evet	Hayır
Güneşli	2	3	Yüksek	2	2	Yüksek	3	4	Yok	6	2	9	5
Bulutlu	4	0	Ilık	4	2	Normal	6	1	Var	3	3		
Yağmurlu	3	2	Serin	3	1								
Güneşli	2/9	3/5	Yüksek	2/9	2/5	Yüksek	3/9	4/5	Yok	6/9	2/5	9/14	5/14
Bulutlu	4/9	0/5	Ilık	4/9	2/5	Normal	6/9	1/5	Var	3/9	3/5		
Yağmurlu	3/9	2/5	Serin	3/9	1/5								

Tablo 4 her bir özellik değerinin Evet ve Hayır şeklinde sınıflanan oyun oynama durumu için kaç defa evet yani oyun oynanmış, kaç defa Hayır yani oyun oynanmamış şeklinde gözlem sayılarını göstermektedir. Örneğin daha önce Tablo 1'de havanın toplamda 5 gün güneşli olduğu ve diğer değişkenlere bakılmaksızın 5 günün 3'ünde oyun oynanmadığı (hayır) 2'sinde oyun oynandığı (evet) görülmektedir. Tablonun satırlarında yer alan özelliklerin her biri için olası değişkenlerin (genel görünüş, sıcaklık, nem ve rüzgar) kaçında oyun oynandığı kaçında ise oyun oynanmadığı görülmekte, tablonun son sütunun da ise toplamda kaç defa oyun oynandığı ve oyun oynanmadığı verilmektedir. Tablo 4'ün alt kısmında aynı özellikler için bu defa frekans sayıları yerine oranlar veya gözlemlerin olasılık değerleri verilmektedir. Örneğin, toplamda 9 gün oyun oynanmış ve bunlardan 2 günde hava güneşli ise güneşli havada oyun oynanma olasılığı 2/9 olacaktır. Ancak oyun oynama için elde edilen oranlar (kesirli ifadeler) farklıdır. Burada toplamda Evet ve Hayır için oyun oynanma ve oynanmama oranları verilmektedir. Şimdi Tablo 5'de gösterildiği gibi yeni bir veri ile karşılaştığımızı düşünelim. Daha açık bir ifadeyle havanın güneşli, sıcaklığın serin,

nemin yüksek ve rüzgarın olması durumunda oyun oynama olasılığını hesaplayalım.

Tablo 5

Yeni Bir Güne İlişkin Veri

Genel Görünüş	Sıcaklık	Nem	Rüzgar	Oyun oynama
Güneşli	Serin	Yüksek	Var	**?**

Tablo 5'de beş tane özellik gösterilmektedir. Bunlar genel görünüş, sıcaklık, nem, rüzgar ve oyun oynama durumudur. Verilen yeni durum için oyun oynama olasılığı;

$$P(\text{oyun oynama}) = \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0{,}0053$$

Yukarıdaki eşitlikte verilen oranlar sırasıyla güneşli, serin, yüksek nemli ve rüzgarlı günlerde oyun oynama olasılıkları ile toplamda oyun oynama olasılığını göstermektedir. Benzer şekilde oyun oynamama durumu için elde edilecek olasılıklar çarpımı aşağıda verilmiştir.

$$P\ (\text{oyun oynamama}) = \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0{,}0206$$

Elde edilen bu sonuçlar veri dosyasına giriş yapan bu yeni durum için oyun oynamama olasılığı (0,0206) oyun oynama olasılığından (0,0053) yaklaşık 4 kat daha fazladır. Elde edilen bu oranlar toplamı 1 olacak şekilde normalleştirildiği taktirde; yani yeni gerçekleşme olasılığını gerçekleşme ve gerçekleşmeme olasılıklarının toplamlarına böldüğümüzde;

$$\text{Oyun oynama olasılığı} = \frac{0{,}0053}{0{,}0053+0{,}0206} = 0{,}2046\ (\%20{,}5)$$

$$\text{Oyun oynamama olasılığı} = \frac{0{,}0206}{0{,}0053+0{,}0206} = 0{,}7953\ (\%79{,}5)$$

Bayesci koşullu olasılık kuralına dayanan bu sonuca göre Tablo 5’te verilen yeni bir gün için oyun oynama olasılığı %20,50 iken oyun oynamama olasılığı %79,50 olarak belirlenmiştir. Buna göre veri dosyasında yer almayan bu yeni gün için oyun oynamama olasılığının daha yüksek olduğu görülmektedir.

Naive Bayes yönteminde sınıflandırma yapılırken özellik değerleri birbirinden bağımsız olarak hesaba katılmaktadır. 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). Verilen x gözlemi altında C sınıfa ilişkin koşullu olasılık değeri aşağıdaki denklem yardımıyla hesaplanabilmektedir.

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

Verilen eşitliklikte $P(C)$ ifadesi C sınıfına ait önsel olasılık değerlerini, $P(x/C)$ ifadesi C sınıfından bir örneğin x olma olasılığı, $P(x)$ 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). Naive varsayımına göre tüm özellikler istatistiksel olarak bağımsız ve her bir özelliğin önsel olasılık değerinin birbirine eşit olduğu kabul edilir. Günlük hayattaki pratik uygulamalar göz önüne alındığında verilen sınıf değeri için özelliklerin nadiren bağımsız olduğu görülmektedir. Bu Naive Bayes yönteminin bir sınırlılığı olarak görülse de Naive Bayes yöntemi daha karmaşık öğrenme yöntemleri olarak görülen karar ağacı, kural öğrenme ve örnek temelli öğrenme algoritmalarından daha doğru kestirimler yapmaktadır (Ceci, 2005).

Tablo 4'ün üst tarafında oyun oynanan günlerden 2’sinde havanın güneşli, 4'ünde bulutlu, 3'ünde ise yağmurlu olduğu görülmekte ve tablonun alt tarafında ise belirlenen toplam 9 durum için olasılık değerleri sırasıyla 2/9, 4/9 ve 3/9 olarak gösterilmektedir. Bunun yerine verilen olasılık değerlerinin payına 1 ve eşitliği bozmamak için paydasına 3 ekleyerek sırasıyla 3/12, 5/12 ve 4/12 değerlerini elde ederiz. Böylelikle veri setinde hiç gözlenmeyen bir özelliğin sıfıra eşit olmayan bir olasılık değeri alması sağlanmış olur. Bunun sebebi veri setinde hiç gözlenmeyen bir özellik için olasılık değerinin sıfır olması durumunda incelenen duruma ilişkin sonsal olasılık değerinin 0 olmasını engellemektir. Her bir sayıma 1 ekleme mantığı 18. YY’ da yaşamış olan Fransız matematikçi Pierre Laplace'ın "Laplace Tahmini" olarak adlandırılan standart bir tekniğe dayanmaktadır. Uygulamada iyi çalışsa da sayma sonuçlarına (olasılık değerlerinin payı) 1 eklemek için özel bir

sebep bulunmamaktadır, hatta bunun yerine Mü (μ) olarak isimlendirilen küçük bir sabit sayı belirleyip bunu da kullanabilir. Böylelikle daha önce verilen 2/9, 4/9 ve 3/9 olasılık değerleri;

$$\frac{2+\mu\backslash 3}{9+\mu}, \frac{4+\mu\backslash 3}{9+\mu} \text{ ve } \frac{3+\mu\backslash 3}{9+\mu}$$

Sabit olan ve öncesinde 3'e eşitlenen μ değeri daha önceden 3 özelliklerin her biri için $\frac{1}{3}$, $\frac{1}{3}$ ve $\frac{1}{3}$ olarak belirlenen önsel değerlerin ne kadar etkili olduğunu belirleyen bir ağırlık değeri sağlanmaktadır. Daha büyük μ değeri bu önsel değerleri eğitim veri setindeki yeni kanıt ile karşılaştırıldığında, öncekinden daha etkili olduğunu göstermesi bakımından daha büyük olan μ değerinin daha önemli olduğunu belirtmektedir. Son olarak μ sabitini 3'e bölmenin herhangi bir belirgin sebebi yoktur, bunun yerine;

$$\frac{2+\mu\,.\mathrm{P1}}{9+\mu}, \frac{4+\mu\,.\mathrm{P2}}{9+\mu} \text{ ve } \frac{3+\mu\,.\mathrm{P3}}{9+\mu}$$

şeklinde P1, P2 ve P3 değerlerinin toplamı 1 olacak şekilde başka sabit sayılar da kullanılabilir. Etkili bir biçimde bu üç sayı sırasıyla; güneşli, bulutlu ve yağmurlu olarak belirlenen genel görünüş özelliğinin önsel P olasılık (priori probability) değerleridir. Bu verilen formül ise önsel olasılıklara değer atanan Tam Bayesci bir formüldür. Daha kesin sonuçlar vermesi bir avantaj iken önsel olasılık değerlerinin nasıl atanacağının her zaman açık olmaması bu yöntemin bir dezavantajıdır. Uygulamada, eğitim veri setinde uygun sayıda örnek olması durumunda önsel olasılıklar küçük farkı ortaya çıkarır ve hesaplamalarda Laplace Tahmin yöntemini kullanmak frekans hesaplamalarında tüm sayımları sıfıra değil 1'e eşitlemektedir. Tablo 6'da yeni bir gün için oyun oynama olasılığının ne olduğu gösterilmektedir.

Tablo 6

Başka Yeni Bir Gün

Genel Görünüş	Sıcaklık	Nem	Rüzgar	Oyun Oynanma
Güneşli	66	90	Var	?

Tabloda verilen yeni bir gün için oyun oynama ve oyun oynamama durumlarına ilişkin olasılık değerleri aşağıda verilmiştir.

Oyun oynama olasılığı = 2/9 x 0,0340 x 0,0221 x 3/9 x 9/14 = 0,000036

Oyun oynamama olasılığı = 3/5 x 0,0221 x 0,0381 x 3/5 x 5/14 = 0,000108

Verilen bu değerler için Bayesci yöntem ile oranlar toplamı 1'e eşit olacak şekilde normalleştirildiğinde elde edilen olasılıklar aşağıdaki gibidir.

P(Oyun oynama) = $\frac{0,000036}{0,000036+0,000108}$ = 0,25 (%25)

P(Oyun oynamama) = $\frac{0,000108}{0,000036+0,000108}$ = 0,75 (%75)

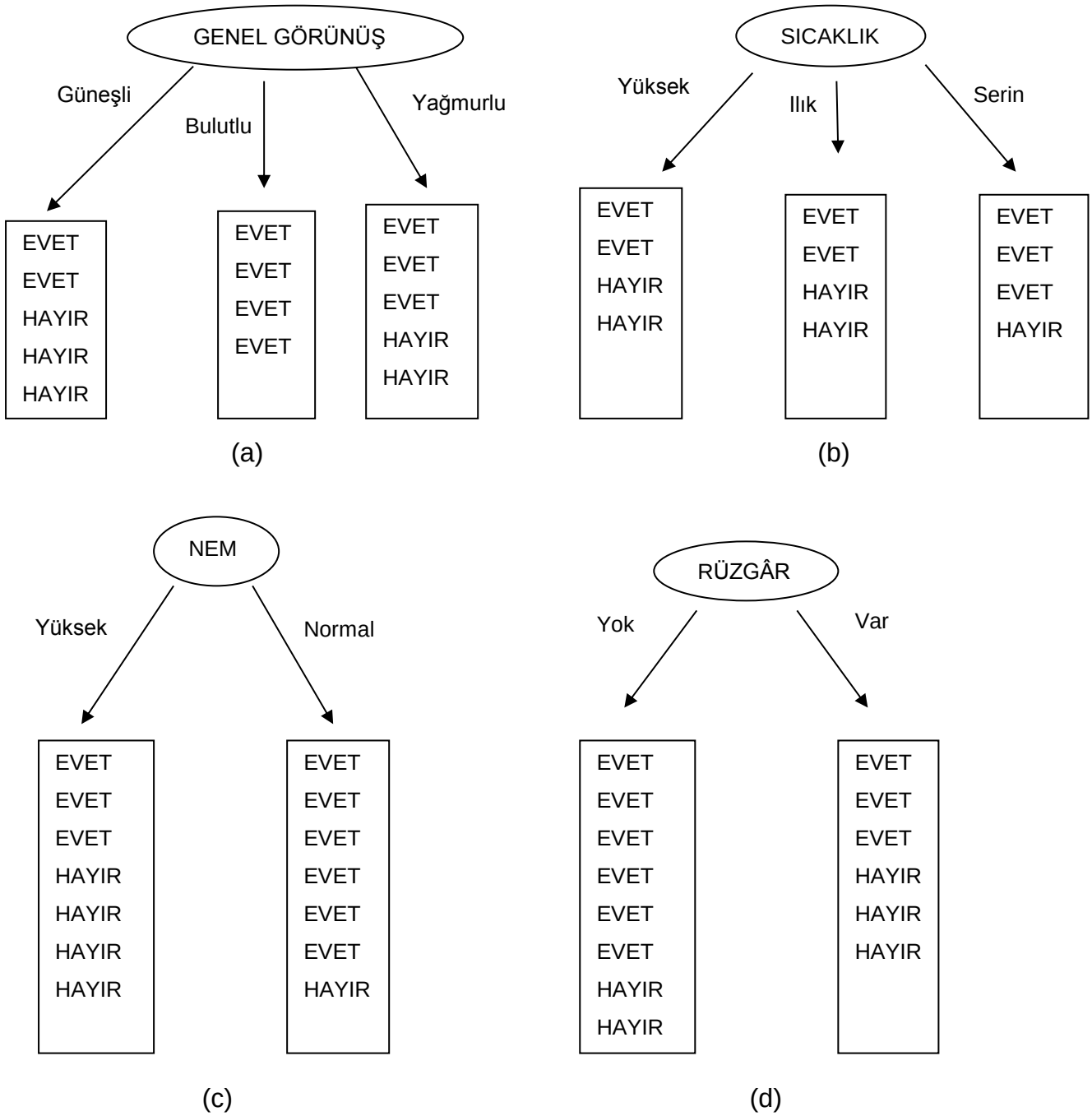
Elde edilen bu olasılık değerleri daha önce Tablo 5'te verilen yeni gün için elde edilen olasılık değerine oldukça yakındır ve bunun sebebi sıcaklığın 66ºF ve nemin 90 olarak ölçülmesi sırasıyla sıcaklığın serin ve nemin yüksek olma durumuyla benzer olasılıklara sahip olmasıdır. Normal dağılım varsayımı Naif Bayes sınıflandırıcının sayısal özelliklerle başa çıkabilmek için genişletilmesini kolaylaştırmaktadır. Eğer herhangi bir özelliğe ilişkin kayıp veri sorunu varsa ortalama ve standart sapma sadece veri setindeki mevcut verilere dayalı olarak hesaplanmaktadır.

Böl ve yönet: Karar ağaçlarının oluşturulması. İlk olarak 1986 yılında Quinlan tarafından ID3 algoritması altında karar ağacı oluşturulmuş ve sonrasında bu algoritma güncellenerek C4.5 adını almıştır ve halen geleneksel istatistiksel yöntemlerle elde edilen karar ağaçlarına temel teşkil etmektedir. ID3 ve C4.5 algoritmalarının her biri karar ağacı oluşturmak için tek bir öznitelikten gelen bilgi kazancının istatistiksel olarak hesaplanmasına dayalıdır. Bu şekilde eğitim veri setindeki örneklere dayalı olarak alınacak kararla ilgili en fazla bilgiyi veren bir özellik seçilir ve kalan özelliklerden en fazla bilgi veren bir başka özellik daha seçilerek işleme devam edilir (Podgorelec, Kokol ve Rozman, 2002).

Bir karar ağacı oluşturma problemi kendi kendini tekrarlar şekilde ifade edilmektedir. Karar ağacı en altta kökü ve daha sonra aşağı doğru dal ve düğümleri olan bir yapıya sahiptir. Kök haricindeki diğer düğümlere içsel düğüm veya test edilen düğüm adı verilmektedir. Bir karar ağacında her bir düğüm, veri setindeki örneklerin girdi özellik değerlerinin belli bir ayırma fonksiyonuna göre iki veya daha fazla alt alana bölmektedir. Düğümlerde gerçekleştirilen testlerde tek bir özelliğe sahip olup olmama durumuna göre sınıflama (ayrışma) yapılırken sayısal bir özellik için özelliğin belli bir kesme puanına göre farklı aralıklardaki değerleri esas alınarak sınıflama yapılır (Maimon ve Rokach, 2005). Her bir örnek için kök düğüme yerleşecek şekilde bir özellik belirlendikten ve olası her bir değer için bir şube (dal) oluşturulduktan sonra bu işlem her bir dala ulaşan örnekleri kullanarak her bir dal için ardışık olarak tekrar edilir. Eğer herhangi bir durumda bir düğümdeki tüm örnekler aynı sınıfa dahil olurlarsa ağacın o parçasının (düğümü) geliştirmesi durdurulur. Çünkü artık farklı sınıflara ayrışma olmayacaktır (Tan, Steinbach ve Kumar, 2006).

Bir karar ağacı kullanarak yeni verileri sınıflandırmak için, kök düğümden başlayarak, bir yaprağa ulaşılana kadar ardışık iç düğümler ziyaret edilir. Her iç düğümde, düğüme ilişkin testler yapılarak bunlar kayıt altına alınır. Bir iç düğümdeki testin sonucu, geçilen dalı ve ziyaret edilecek sonraki düğümü belirler. Kayıt altına alınan sınıf sadece son yaprak düğümünün sınıfıdır. Böylece, kökten bir yaprağa kadar olan dallar için tüm koşulların birleşimi, yaprak ile ilişkili sınıfın koşullarından birini oluşturur (Rastogi ve Shim, 2000).

Karar ağacında düğüme bağlı olan her bir yaprak bir sınıf değerini göstermektedir. Karar vermek için geriye kalan tek şey, farklı sınıflarda bir dizi örnek verildiğinde, hangi özelliğin nasıl bölüneceğini belirlemektedir. Tekrardan daha önceki bölümlerde ele aldığımız farklı hava koşulları ve oyun oynama durumuna ilişkin örnekleri göz önüne aldığımızda her bölünme için olası 4 ihtimal bulunmaktadır ve Şekil 4'de gösterildiği gibi en üstte özellik olacak şekilde örnekler iligli sınıflara ayrılmaktadırlar.



Şekil 4. Hava durumu verisi için ağaç kütükleri (Witten, Frank ve Hall, 2016)

Şekilde verilen dallanmalardan hangisinin en iyi seçim olacağına karar vermek için yapraklardaki evet ve hayır sınıflarının sayılarına bakmak yeterli olacaktır. Sadece Evet veya Hayır şeklinde yaprakta tek bir sınıf oluştuğunda tekrardan ayrılmaya gerek kalmayacak ve aşağıya doğru yinelenen dallanma

süreci sona erecektir. Bu işlemin mümkün olduğunca kısa bir sürede gerçekleşmesini istediğimiz için genellikle küçük ağaçları inceleriz. Eğer her bir düğümün saflığını (purity) ölçseydik bu durumda en saf çocuk düğümleri üreten özellikleri seçmemiz gerekecektir (Tan, Steinbach, Karpatne ve Kumar, 2018).

Kullanacağımız saflığın ölçüsü bilgi (information) olarak adlandırılır ve bit olarak belirlenen birimlerle ölçülür. Ağacın bir düğümü ile ilişkili olarak yeni bir örneğin Evet (oyun oynanacak) veya Hayır (oyun oynanmayacak) şeklinde sınıflandırılmasında gereken bilgi miktarını temsil etmektedir. Elde edilen bilgi daha önce gösterilen çok aşamalı özelliğe uygun olmalıdır. Dikkat çekici bir şekilde, tüm bu özellikleri karşılayan tek bir fonksiyon olduğu görülmektedir ve bu fonksiyon bilgi miktarı veya *entropi* olarak isimlendirilir (Fürnkranz, 1999).

Tesadüfi bir X değişkenin N farklı değeri için i. değeri x_i ve olasılık değeri $p(x_i)$ olmak üzere elde edilecek bilgi kazanç miktarı aşağıdaki eşitlik yardımıyla elde edilmektedir (Kak, 2017).

$$H(X) = -\sum_{i=1}^{N} p(x_i).\log_2 p(x_i)$$

Bilgi kazanç miktarı bit cinsinden elde edildiği için logaritmanın tabanı 2 olarak tanımlanmaktadır. Bunun yanında toplam sembolü altında verilen eşitlik açıldığı taktirde aşağıdaki sonuç elde edilecektir:

entropy (p_1, p_2, …, p_n) = - p_1.logp_1 – p_2.logp_2, … , - p_n.logp_n

Buradaki (-) eksi işaretlerinin sebebi p_1,p_2,…,p_n kesirlerinden kaynaklanmatadır. Logaritma fonksiyonunun kuralı gereği A/B şeklindeki kesirli ifadelerin logaritması alınırken taban aynı olmak üzere paydaki ifadenin önüne eksi işareti konulmaktadır ($\log_a {}^{A}/_{B}$= $\log_a A$ - $\log_a B$). Her ne kadar kesirlerin önüne eksi işareti olsa da elde edilen bilgi miktarı gerçekte pozitif olacaktır. Entropide yer alan p_1,p_2,…,p_n ifadeleri toplamı 1 olan kesirler anlamına gelmektedir (Jung, Vogiatzian, Har-Shemesh, Fitzsimons ve Quax, 2014).

Örnek olması bakımından toplamda güneşli olan 2 gün, yağmurlu olan 3 gün ve bulutlu olan 4 gün için genel görünüş özelliği için elde edilecek bilgi miktarı aşağıdaki eşitlikte gösterilmiştir.

info ([2,3,4]) = entropi (2/9, 3/9, 4/9)

Yukarıdaki formülde görüldüğü gibi p_1,p_2,p_3 olarak belirtilen $\frac{2}{9}, \frac{3}{9}\ ve\ \frac{4}{9}$ kesirlerinin toplamı 1'e eşit olacaktır. Dolayısıyla çok aşamalı özelliklere ilişkin kararlar aşağıda verilen genel formül yardımıyla elde edilebilir.

$$\text{info ([2,3,4])} = -\frac{2}{9}\ \mathrm{x}\log\left(\frac{2}{9}\right) - \frac{3}{9}\ \mathrm{x}\log\left(\frac{3}{9}\right) - \frac{4}{9}\ \mathrm{x}\log\left(\frac{4}{9}\right)$$

$$= -\frac{2}{9}\ \mathrm{x}\ (\log 2 - \log 9) - \frac{3}{9}\ \mathrm{x}\ (\log 3 - \log 9) - \frac{4}{9}\ \mathrm{x}\ (\log 4 - \log 9)$$

$$= \frac{2\log 2 + 2\log 9}{9} - \frac{3\log 3 + 3\log 9}{9} - \frac{4\log 4 + 4\log 9}{9}$$

$$= \frac{-2\log 2 - 3\log 3 - 4\log 4 - 9\log 9}{9}$$

Yukarıda verilen formül bilgi miktarının pratikte nasıl hesaplandığını gösteren bir formüldür. Buna göre Şekil 4'de verilen ilk ağacın yaprak düğümü olan genel görünüş=güneşli durumu için elde edilecek bilgi miktarı aşağıda verilmiştir.

$$\text{info ([2,3])} = -\frac{2}{5}\ \mathrm{x}\log\frac{2}{5} - \frac{3}{5}\ \mathrm{x}\log\frac{3}{5}$$

$$= -\frac{2}{5}\mathrm{x}(-1{,}3219) - \frac{3}{5}\mathrm{x}(-0{,}7369)$$

=0,5287 + 4421

= 0,9708

= 0,971 bits

Benzer şekilde diğer düğümlerde yer alan yapraklar için elde edilecek bilgi miktarları da kolaylıkla hesaplanabilecektir. Ancak genel görünüş=bulutlu durumu için dört örnek durumun tamamında oyun oynanmış ve elde edilen bilgi miktarı $-1\log(1) - 0\ \mathrm{x}\log(0)$ = 0 olarak hesaplanmıştır. Buna göre genel görünüş özelliği bulutlu olduğunda tüm örnekler için oyun oynandığı için artık alt basamaklarda bölünme olmayacak ve elips şeklinde gösterilen düğüm yerine dikdörtgen şeklinde örneklerin yer aldığı sınıf değeri gösterilecektir.

Elde Edilen Sonuçların Değerlendirilmesi

Değerlendirme VM'de gerçek anlamda bir ilerleme sağlamak için anahtar rolündedir. Eldeki veri dosyasından örüntü çıkarmak ve bu örüntülerden çıkarımda bulunmanın birçok farklı yolu vardır. Fakat belirli bir problem durumunda hangi yönteminin kullanılacağını belirleme aşamasında farklı yöntemlerin nasıl çalıştığını değerlendirmek ve birbirleriyle karşılaştırmak için sistematik yollara ihtiyacımız vardır. Değerlendirme ise göründüğü kadar basit bir işlem değildir (Witten ve Frank, 2007).

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, Matwin ve Japkowicz, 2002).

Belirli bir veri seti üzerinden elde edilen sınıflandırıcının performansını tahmin etmek için kullanılacak en belirgin ölçüt tahmin edilen doğruluk, yani aslında hangi sınıfta olduğu bilinmeyen örneklerin doğru sınıflandırılma oranıdır. Her ne kadar doğruluk oranı en önemli kriter olarak görülse de, sınıflandırıcının elde ettiği sonuçları değerlendirmek için başka ölçütler de bulunmaktadır (Bramer, 2013).

Elinizde geniş bir veri kaynağı olduğunda sorun yoktur. Çünkü büyük bir eğitim setine dayanan bir model oluşturup, oluşturduğunuz bu modeli başka bir büyük veri setinde test edebilirsiniz. Fakat her ne kadar veri madenciliği "büyük veri" ile çalışılıyor olsa da pazarlama, satış ve müşteri destekli uygulamalarda nitelikli verilerin kıt olduğu bilinmektedir. Petrol kayıtları, kredi kartı uygulamaları, elektrik tüketim miktarları, başkanlık seçimleri, market satış rakamları gibi uygulamalardan geriye dönük oldukça fazla kayıt olduğundan elde edilen örnekler büyük veri tanımı içerisine girmektedir (Ramageri, 2010).

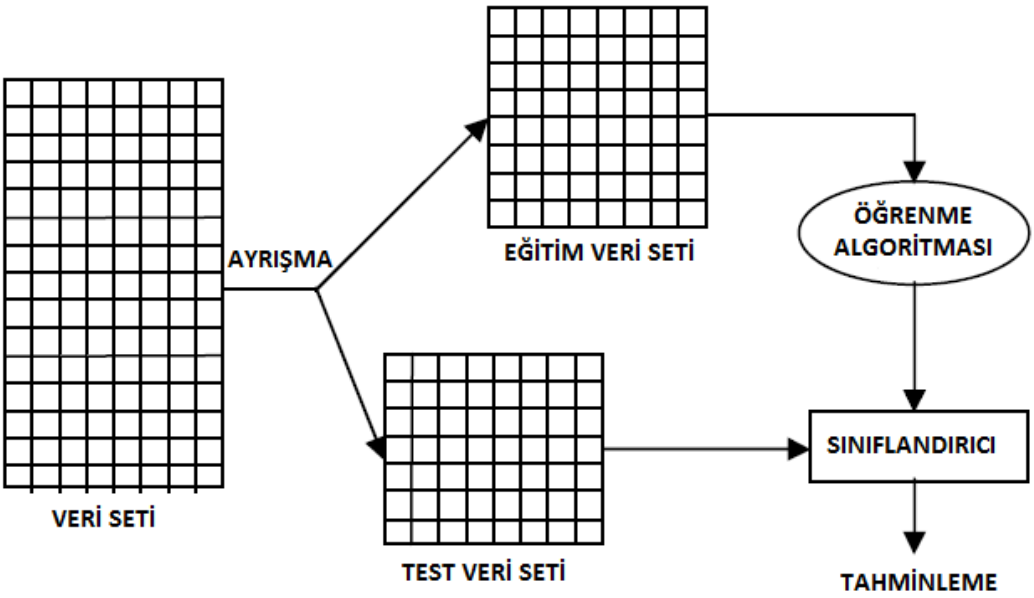
Sınırlı bir veriye dayalı olarak performansı tahmin etmek hala ilginç ve tartışmalı bir konudur. Bu aşamada elde edilen sonuçların değerlendirilmesi amacıyla birçok farklı teknik karşılaşınıza çıkacak ve bunların içinde muhtemelen sınırlı veriye dayalı uygulamalarda en çok tercih edilen değerlendirme yöntemi en doğru ve en

tutarlı sonuç veren yöntem olacaktır. Aslında belirli bir problem üzerinde farklı makine öğrenme yöntemlerinin performansını karşılaştırmak göründüğü kadar kolay değildir. Yöntemler arasında oluşacak belirgin farklılıkların tesadüfü etkilerden kaynaklanmadığından emin olmak için istatistiksel testlere ihtiyaç duyulmaktadır (Chamatkar ve Butey, 2014). Şuana kadar üstü kapalı bir şekilde tahmin edilen şeyin test örneklerini doğru bir şekilde sınıflama olduğunu varsayılmıştır. Buna rağmen, bazı durumlar sınıfların kendisinden ziyade sınıflama olasılıklarını tahmin etmeye dayanır ve diğerleri ise sınıflamadan ziyade sayısal değerleri tahmin etmeyi içermektedir. Bu sebeple her durumda farklı yöntemlere ihtiyaç duyulmaktadır. Bunların yanında elde edilen sonuçların maliyet sorunu bizim için oldukça önemlidir. Çoğu pratik veri madenciliği uygulamasında yanlış sınıflandırma hatasının maliyeti hatanın türüne bağlıdır. Örneğin pozitif bir örnek hatalı olarak negatif veya tam tersi şekilde sınıflanmış olabilir. Veri madenciliği çalışırken ve kullanılan yöntemin performansını değerlendirirken bu maliyetleri hesaba katmak oldukça önemlidir. Neyse ki çoğu öğrenme algoritmasını bu algoritmanın teknik kısımlarıyla uğraşmadan daha duyarlı hale getirmek için basit teknikler bulunmaktadır. Son olarak değerlendirmenin bütün kavramlarının felsefi bağlantılara sahip olduğu söylenebilir (Paidi, 2012). Son 2000 yıldır filozoflar bilimsel teorinin nasıl değerlendirileceği konusunda tartışmaktadırlar ve bu tartışmalar veri madenciliği ile daha da odak noktası olmuştur.

Model değerlendirmesinde kullanılabilecek yöntemler temel olarak 4 başlık altında yer almaktadır. Bunlar Bekletme (*hold-out*), K-katlı çapraz geçerleme (*k-fold cross validation*), Bir tanesinin kapsam dışı bırakma (*Leave-one-out*) ve Bootstrap yöntemidir. Değerlendirme yöntemlerinden ilki olan bekleme yaklaşımı aslında tüm verileri eğitim ve test verisi şeklinde belli oranlarda ayırıştırma işlemidir (Souza, Matwin ve Japkowicz, 2002).

Verileri eğitme ve test etme. Elde edilen sonuçların değerlendirilmesinde farklı yöntemlerin ne kadar iyi çalıştığını görebileceğiniz test veri setiniz bulunmaktadır. Ama aslında eğitim setindeki performansın nasıl olduğunu öğrenmek bağımsız bir test setindeki performansın iyi bir göstergesi olmayacaktır. Elde edilen veri setleri üzerinde yapılacak deneysel çalışmalara dayalı olarak performans sınırlarını tahmin etme yollarına ihtiyacımız olacaktır.

Bu yöntemde elinizdeki veri dosyası eğitim ve test verileri olarak ikiye ayrılmaktadır. İlk olarak sınıflandırıcı (*classifier*) olarak isimlendirilen öğrenme yöntemini oluşturmak için eğitim veri seti oluşturulur. Daha sonra test veri setinden oluşturmuş olduğunuz sınıflandırıcının örneklerin sınıflarını tahmin etmesi sağlanır. Eğer toplam N örnekten oluşan veri setinizin C tanesi doğru olarak sınıflanmış ise bu durumda tahmin edilen doğruluk oranı $p = {}^{C}/_{N}$ olacaktır. Bu formül hangi sınıfta olduğunu bilmediğiniz veriler için kullanılacak sınıflandırıcının performans ölçütlerinden biri olan tahmin edilen *doğruluk oranı* olacaktır. Görsel olarak eğitim ve test sürecinin nasıl çalıştığı Şekil 5'te gösterilmiştir (Bramer, 2013).



Şekil 5. Eğitim ve test verilerileri üzerinden öğrenme süreci

Veri madenciliğinde ilgilendiğimiz şey eski verilere dayalı geçmiş performans yerine yeni verilere dayalı olarak gelecekteki muhtemel performansın tahminidir. Zaten hali hazırda eğitim kümesindeki her bir örneğin hangi sınıfta olduğunu bilindiği için genel olarak bu sınıflandırma ile ilgilenilmemektedir. Ancak amacımız tahminden ziyade veri temizlemek ise sınıflamalara da dikkat edilmektedir. Bunun yanında eğitim setindeki hata oranı gerçekteki performansın iyi bir göstergesi olmayabilir. Çünkü sınıflandırıcı aynı eğitim verilerinden öğrendiği için bu verilere dayalı olarak yapılacak herhangi bir hesaplama işlemi de gerçek değerinden daha iyimser olacaktır.

Eğitim setindeki hata oranı yeniden yerleştirme hatası olarak isimlendirilir çünkü eğitim setindeki örnekler onlardan oluşturulan bir sınıflandırıcı tarafından yeniden sınıflandırma yapmak suretiyle hesaplanmaktadır. Her ne kadar bu değer yeni verilerin gerçek hata oranının güvenilir bir belirleyicisi olmasa da yine de bu değerin ne olduğunu bilmek faydalı olacaktır. Yeni veri üzerinden bir sınıflayıcının performansını tahmin etmek için sınıflandırıcının elde edildiği veri setinden daha farklı bir veri seti üzerinden hata oranını değerlendirilmesi gerekmektedir. Çalışmalarda hem eğitim verilerinin hem de test verilerinin araştırıldığı problemin temsili örnekleri olduğu varsayılmaktadır (Mavroforakis, 2011).

Bazı durumlarda test verileri doğadaki eğitim verilerinden farklı olabilmektedir. Witten, Frank ve Hall (2016) bu durumu şu örnekle açıklamaktadır. Elimizde kredi risk problemine ilişkin örneklerimiz bulunsun. Bu örnekler üzerinden bankanın New York ve Florida'daki şubelerinden eğitim verileri aldığını ve bu eğitim verisi üzerinden elde edilecek sınıflandırıcının Meksika'daki yeni bir şubede ne kadar iyi performans göstereceğini öğrenmek istediğini varsayalım. Muhtemelen New York sınıflandırıcısını değerlendirmek için Florida verilerini test verileri olarak kullanmak ve Florida sınıflandırıcısını değerlendirmek için New York verilerini kullanmak uygun olacaktır. Eğer veri setleri veriyi eğitme işleminden önce bir araya getirilirse test verilerindeki performans muhtemelen gelecekte tamamı ile farklı bir eyaletten elde edilecek verilerin performansının iyi bir göstergesi olmayacaktır.

Test verisinin herhangi bir şekilde sınıflandırıcıyı oluşturma aşamasında kullanılması önemlidir. Örneğin bazı öğrenme yöntemleri iki aşamalıdır ve bunlardan birinde temel yapı diğerinde bu yapıyla ilgili parametrelerin uygun hale getirilmesi yer almaktadır. Bu sebeple iki aşamada da ayrı veri setlerine ihtiyaç duyulabilmektedir. Ya da eğitim verileri üzerinden birkaç öğrenme yöntemi ortaya çıkarabilir ve sonrasından hangisinin daha iyi çalıştığını görmek için yeni bir veri kümesi üzerinden bu öğrenme yöntemlerini karşılaştırabilir. Fakat bu verilerin hiçbiri gelecekteki hata oranının bir tahminini belirlemek için kullanılamaz. Böyle durumlarda genel olarak üç farklı veri kümesinden bahsedilmektedir. Bunlar; eğitim verisi, test verisi ve geçerliliği belirlemek anlamına gelen doğrulama verisidir. Eğitim verileri sınıflandırıcıları ortaya çıkarmak amacıyla bir veya daha fazla öğrenme yöntemi ile birlikte kullanılmaktadır. Doğrulama verileri bu

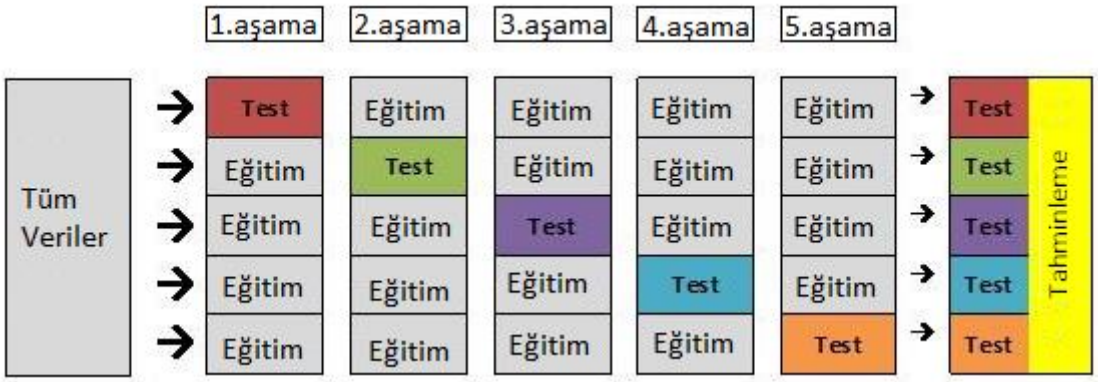
sınıflandırıcının parametrelerini optimize etmek veya belirli bir sınıflandırıcı seçmek için kullanılır. Ardından test verileri nihai olarak belirlenmiş ve optimize edilmiş yöntemin hata oranını hesaplamak için kullanılır. Belirtilen bu üç kümenin her biri de birbirinden bağımsız olarak seçilmelidir. Doğrulama kümesi optimizasyon veya seçim aşamasında iyi performans elde etmek için belirlenen eğitim setinden farklı olmalıdır ve test kümesi güvenilir bir hata oranı hesaplayabilmek için bunların her birinden farklı olmalıdır (Olson ve Delen, 2008).

Başka bir yaklaşımla hata oranı belirlendikten sonra, test verileri aktif olarak kullanılacak olan bir sınıflandırıcı oluşturmak için eğitim verilerine geri gönderilebilir. Burada yanlış olan bir durum yoktur. Bu sadece pratik olarak kullanılabilecek bir sınıflandırıcı elde edebilmek için kullanılacak veri miktarını artırmanın bir yoludur. Ayrıca önce doğrulama verisi kullanıldığında, belki de en iyi öğrenme yönteminin hangisi olduğunu belirlemek için, aynı veriler öğrenme yöntemini tekrardan eğitmek için eğitim datasına geri gönderebilir. Ancak elinizde çok fazla veriniz varsa sorun yoktur: Büyük bir örneklem alıp bunu eğitim için kullanılırsınız ve sonra başka bir bağımsız büyük örneklemi test için kullanabilir. Her iki örneklemin de birbirinden farklı olması koşuluyla, test kümesi üzerinden elde edilen hata oranı gelecekteki performansın gerçek bir göstergesi olacaktır. Genel olarak eğitim verileri ne kadar büyük olursa oluşturulacak sınıflandırıcı da o kadar iyi olur ancak belirli bir eğitim verisi için gerekli olan sınır aşıldığında geri dönüşler azalmaya başlayacaktır. Bu sebeple test verileri ne kadar çok olursa hata tahmini de o kadar doğru olacaktır. İlerleyen bölümlerde anlatılacağı gibi hata tahmininin doğruluğu istatistiksel yöntemlerle nicelleştirilebilmektedir.

Elde edilecek hata miktarı ile ilgili asıl sorun geniş bir veri kaynağı olmadığı durumlarda ortaya çıkmaktadır. Çoğu durumda eğitim verileri elle sınıflandırılmalıdır ve tabiî ki hata oranını elde etmek için test verileri de benzer şekilde sınıflandırılmaktadır. Bu durum eğitim, doğrulama ve test için kullanılabilecek veri miktarını sınırlandırır ve bu sınırlı veri ile en iyi performansın nasıl elde edileceği bir sorun haline gelir. Hatta eğitim için ayrılan verilerin bir kısmı doğrulama için de kullanılabilmektedir. Bu veri kümesinden test için belirli bir miktar (%20 - %30) tutulur ve bu işleme bekletme prosedürü denilir. Daha sonra elinizde kalan miktar eğitim için kullanılır. Bekletme yönteminde eldeki verinin bir miktarını test etmek için ayırırken geri kalan kısmı eğitim için kullanılmaktadır.

Pratik olarak eldeki verinin üçte birini (1/3) test için geri kalan üçte ikilik (2/3) kısmı ise eğitim için kullanmak oldukça yaygın bir yöntemdir. Ancak eğer gerekli ise eğitim için kullanılacak verinin bir kısmı doğrulama verisi olarak ayrılabilir. Diğer yandan bekletme yönteminde bir ikilem bulunmaktadır. Hem iyi bir sınıflandırıcı elde edilebilmek için eldeki verinin çoğunu eğitim için kullanmak isterken hem de iyi bir hata tahmini yapabilmek için test verilerinin de mümkün olduğu kadar çok olması istenmektedir (Ahmed ve Elaraby, 2014).

Çapraz Geçerlik. Veri madenciliği için kullancak verilerinizin az ve sınırlı olması durumunda elde edilen sonuçların değerlendirilmesi için bekletme yaklaşımına alternatif olarak kullanabileceğiniz değerlendirme yöntemi k-katlı çapraz geçerleme yaklaşımıdır. Toplam N örnekten oluşan veri setiniz *k* adet (5 veya 10 gibi) eşit parçaya ayrıldığını düşünün (Eğer toplam örnek sayısı k sayısına tam olarak bölünemiyorsa ayırdığınız son parçadaki örnek sayısı diğer k-1 parçadaki örnek sayılarından daha az olacağını unutmayınız). Bu durumda k adet analiz ardışık olarak gerçekleştirilmektedir. Çapraz geçerleme sürecinde sırasıyla k parçanın her biri test veri seti olarak kullanılırken diğer k-1 parçalar eğitim veri seti olarak kullanılmaktadır (Bramer, 2013). Örnek olması bakımından k=5 için elde edilen sonuçların değerlendirimesi amacıyla kullanılacak olan 5 katlı çapraz geçerleme sürecinde nasıl bir yol izlendiği Şekil 6'da görsel olarak anlatılmaktadır.



Şekil 6. Çapraz geçerlik süreci

Çapraz geçerlemede sabit bir katlanma ya da verilerin bölünme sayısına karar verilir. Örneğin şekilde görüldüğü gibi kullanılan sabit sayı 5 olsun. Daha sonra veriler yaklaşık olarak eşit beş parçaya ayrılır ve her biri sırasıyla test için kullanılırken geri kalanları eğitim için kullanılmaktadır. Yani böylece test için beşte birini kullanırken eğitim için eldeki verilerin beşte dördünü kullanmış olursunuz ve

böylece sonunda her bir örnek test için bir kez kullanılmış olacaktır. Buna beş katlı çapraz geçerlik denir ve eğer tabakalandırma işlemi yapılmışsa buna tabakalandırılmış beş katlı çapraz geçerlik denilmektedir.

Elbette eğitim (ya da test) için kullandığınız örneklem grubu evrenin iyi bir temsilcisi olmayabilir. Genelde bir örneklemin iyi bir temsilci olup olmadığını doğrudan anlaşılamamaktadır. Ancak sizin için önemli olabilecek çok basit bir kontrol vardır. Bu yaklaşımda verilerin tamamının yer aldığı çalışma evreninizde yer alan her bir sınıfın eğitim ve test verilerinde doğru bir oranda yer alması gerekmektedir. Ancak bu sayede örneklemin evreni temsil gücü artacaktır. Ancak kötü bir şansınız varsa ve aldığınız örneklemdeki bir sınıf için tüm örneklerde kayıp veri varsa ne yazık ki bu verilerden elde edilecek sınıflandırıcının test verileri üzerinde iyi bir performans göstermesini bekleyemezsiniz. Böyle bir durumda bu verilerin hiçbiri eğitim kümesinde yer almadığı için test kümesindeki verilerin aşırı temsil edilmesi sebebiyle sonuçlar daha da kötüleşecektir. Bunun yerine evrende yer alan her bir sınıfın eğitim ve test verilerinde düzgün bir şekilde temsil edildiğini garanti edebilmek amacıyla tesadüfü örnekleme yapıldığından emin olmanız gerekmektedir. Bu yönteme tabakalandırma veya tabakalı bekletme yaklaşımı adı verilmektedir. Her ne kadar bu yöntemi uygulamak genellikle yapmaya değer olsa da bu sadece eğitim ve test verilerindeki eşit olmayan veya dengesiz bir temsili ortadan kaldırmak için ilkel bir koruma olarak görülmektedir (Vanwinckelen ve Blockeel, 2012).

Belirli bir örneklem seçiminden kaynaklanan yanlılığı azaltmanın daha genel bir yolu farklı örneklem grupları üzerinden eğitim ve test verileri için tüm süreci birkaç kez tekrar etmektedir. Her bir tekrarlama işlemini verinin belli bir miktarı, örneğin eğitim için üçte ikisi, eğer mümkünse tabakalı ve rastgele seçilmiş bir şekilde eğitim için kullanılırken geri kalan miktarı test için kullanılabilir. Farklı tekrarlardan elde edilen hata oranları ortalaması alınarak daha genel bir hata oranı elde edilebilir. Buna hata oranını hesaplamak için tekrarlı bekletme yöntemi adı verilmektedir. Sadece bir bekletme yönteminin uygulandığı durumlarda, eğitim ve test verilerinin rollerini değiş-tokuş ettiğini düşünebilirsiniz. Yani test verileri üzerinden sistemi eğitirken, eğitim verileri üzerinden bu sistemi test edebilirsiniz. Böylece elde ettiğiniz iki sonucun da ortalamasını alarak eğitim ve test verilerindeki eşit olmayan temsil sorununu azaltmış olursunuz.

Elinizdeki tek ve sabit bir veri olduğunda öğrenme yönteminin hata oranını tahmin edebilmek için standart olarak 10 katlı çapraz geçerleme yaklaşımı uygulanmaktadır. Bu yaklaşımda verilerin tamamının yer aldığı çalışma evreni rastgele 10 sınıfa ayrılmaktadır ve her bir sınıfın tam veri kümesinde yaklaşık olarak aynı oranda temsil edilmesi gerekmektedir. Her bir sınıf sırasıyla test için ayrı bir kenarda bekletilirken geriye kalan dokuz sınıf üzerinden öğrenme süreci gerçekleşir ve bir kenarda bekletilen 1/10'luk veriler üzerinden hata oranı hesaplanmaktadır. Böylece öğrenme yöntemi farklı eğitim verileri üzerinden (her birinin birçok ortak noktası bulunan) toplam 10 defa tekrar edilerek gerçekleşmektedir. Son olarak 10 farklı hata teriminin ortalaması alınarak genel bir hata terimi elde edilmektedir (Witten ve Frank, 2007).

Bu aşamada 10 katlı çapraz geçerleme yönteminin kullanılma sebebi çeşitli öğrenme teknikleriyle daha büyük veri setleri üzerinden geçekleştirilen kapsamlı testlerde en iyi hata tahmini veri setini 10 parçaya ayırma işleminin en ideal olduğu ve bunun yanında bu sonucu destekleyen teorik kanıtlarında da var olmasıdır. Bu söylenenlerin hiç biri kesin bir kuram değildir ve değerlendirme konusunda en iyi yaklaşımın hangisi olduğuna ilişkin veri madenciliği ve makine öğrenme çevrelerindeki tartışmalar halen devam etse de, 10 katlı çapraz geçerleme yöntemi pratikte herkes tarafından kullanılan standart bir değerlendirme ölçütü haline gelmiştir. Deneysel çalışmalarda ayrıca tabakalı bir örnekleme yaklaşımının sonuçları biraz daha iyileştirdiği görülmektedir (Vanwinckelen ve Blockeel, 2012). Her ne kadar performans tahmininin doğru ve tutarlı bir şekilde yapılması yöntemin bir avantajı olarak görülse de birbiriyle çakışan eğitim verileri, karşılaştırma için 1.tip hatanın artması, karşılaştırma için performans varyansının tahmin edilememesi ve serbestlik derecesinin aşırı tahmin edilmesi gibi dezavantajları da bulunmaktadır (Refaeilzadeh, Tang ve Liu, 2009). Bu sebeple elinizde sınırlı bir veri olması durumunda standart bir değerlendirme yöntemi olarak tabakalandırılmış 10 katlı çapraz geçerleme yöntemi kullanılması önerilmektedir. Unutulmamalıdır ki ne tabakalandırmanın ne de 10 katlı çapraz geçerleme için oluşturulacak sınıfların tam olarak birbirine eşit olması sağlanamaz. Verinizi yaklaşık olarak birbirine eşit 10 sınıfa ayırmanız ve farklı tabakaların bu sınıflarda doğru oranda temsil edilmesi yeterli olacaktır. Dahası 5, 10 veya 20 katlı çapraz

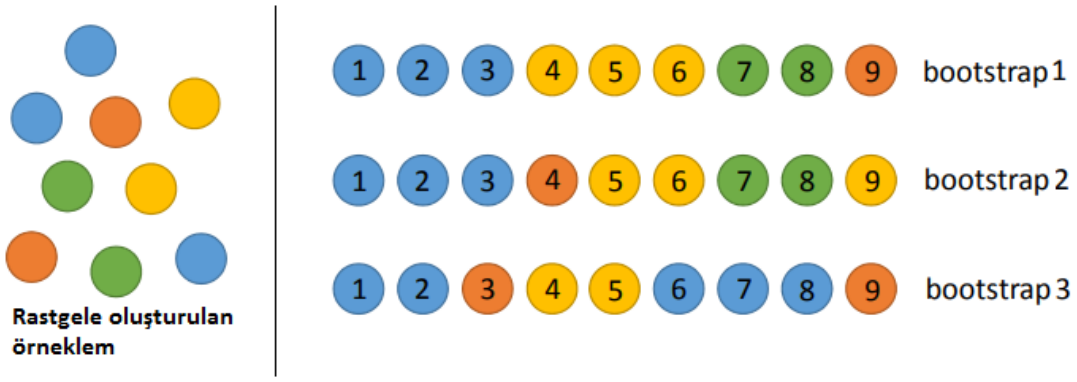
geçerlemenin her zaman iyi olacağı konusunda kesin bir kanı bulunmamaktadır (Williams, 2010).

Güvenilir bir hata tahmini için tek bir 10 katlı ve tabakalı çapraz geçerleme yeterli olmayabilir. Aynı veriler üzerinden aynı öğrenme yöntemiyle yapılan 10 katmanlı çapraz geçerleme deneyleri katmanların kendisini seçme aşamasındaki tesadüfü değişim etkisinden dolayı farklı sonuçlar üretebilmektedir. Tabakalandırma değişkenliği azaltır ancak kesinlikle tamamen ortadan kaldırdığı söylenemez. Doğru bir hata tahmini yapabilmek için çapraz geçerleme işlemini 10 kez tekrar etmek standart bir süreçtir. Bu bize 10 tabakalı çapraz geçerlemeyi ve 10 farklı sonuç üzerinden elde edilen sonuçların ortalamasını bulmamızı sağlamaktadır (Kohavi, 1995).

Bootstrap Yöntemi. Bu yöntem bir tahmincinin istatistiksel doğruluğunu değerlendirmek için Efron (1979) tarafından ortaya atılmıştır. Öte yandan Friedman (1998) bu yöntemin son 20 yılda ortaya atılmış olan en önemli teorilerden biri olduğunu ve gelecek 50 yıl içinde de en çok tartışılacak konulardan biri olduğunu belirtmiştir (Kuonen, 2018). Daha önce bahsetmiş olduğumuz kestirim yöntemlerinin bir diğeri olan Bootstrap yöntemi örnekleminin yer değiştirmesi olarak tanımlanabilecek istatistiksel yöntemlere dayanmaktadır. Bu yöntemde eğitim veya test için eldeki veri setinden bir örneklem alınır ve bu örneğin üzeri çizilir. Yani aynı örnek bir kez seçildiğinde tekrardan seçilmemektedir (Dekking, Kraaikamp, Lopuhaa ve Meester, 2005). Bu yaklaşım futbol oyunu için takımlar oluşturmaya benzer. Nasıl ki bir oyuncu farklı iki takımda yer alamıyorsa Bootstrap yönteminde de aynı örnek iki kez seçilemez. Ancak veri kümesindeki örnekler insanlar gibi değildirler. Çoğu öğrenme yöntemi aynı örneği iki defa kullanmaktadır ve eğer bir veri eğitim kümesinde iki defa yer almışsa sonuçlarda farklılık oluşacaktır. Matematiksel olarak konuşmak gerekirse aynı nesnenin birden fazla görüntüsü olması durumunda "küme" olarak tanımlanan gruplardan bahsetmek anlamlı olmayacaktır.

İstatistiksel olarak çıkarımda bulunmak için kullanılan geleneksel yaklaşımlar idealize edilmiş model ve varsayımlara dayanmaktadır. Genellikle, standart hata gibi doğruluk ölçütleri asimtotik teoriye dayanır ve küçük örneklem gruplarında bu teorinin varsayımlarının karşılanması kolay değildir. Geleneksel yaklaşımın modern bir alternatifi olan Bootstrap yöntemi, daha gerçekçi modellerin

uygulanmasına izin veren bilgisayarla yoğunluklu bir yeniden örnekleme yöntemidir (Boss, 2003). Görsel açıdan Bootstrap yönteminin daha iyi anlaşılması için Şekil 7'nin incelenmesi gerekmektedir. Şekilde görüldüğü üzere rastgele oluşturulmuş olan örneklemde yer alan birimler her bir adımda yerleri değiştirilerek yeniden örnekleme dahil edilmektedir.



Şekil 7. Bootstrap yönteminde örnekleme süreci.

Bootstrap yöntemi temel olarak 4 adımda özetlenmektedir. Birinci adımda tüm veri setindeki örneklerin yerlerini değiştirirek rastgele bir örneklem oluşturulur. İkinci adımda örneklemden bazı istatistikler elde edilir. Örnek olarak örneklemin medyan değeri hesaplanır. Üçüncü adımda ilk iki adımdaki işlemler Bootstrap dağılımını elde edebilmek için N_b defa tekrar edilir. Dördüncü ve son adımda Bootstrap dağılımı kullanılarak güven aralığı ve standart hata gibi değerler hesaplanmaktadır (Galdi ve Tagliaferri, 2017).

Bootstrap yönteminin mantığı bir eğitim kümesi oluşturabilmek için elimizdeki veri setini yer değiştirerek örnekleyebilmektir. Bu aşamada gizemli bir şekilde 0,632 Bootstrap sayısının ne olduğunun bilinmesi gerekmektedir. Bunun için n örnekten oluşan bir veri seti yer değiştirilerek n örnekten oluşan başka örneklem elde edilecek şekilde örneklenmektedir. Bu ikinci veri setinde yer alan bazı örnekler tekrar edeceği için test örnekleri olarak kullanılacak ve orijinal veri kümesinde başka yerde kullanılmış olan örneklerin olması gerekmektedir.

Bu durumda belirli bir örneğin eğitim kümesinde yer alma olasılığı $1/n$ iken bu kümede yer almama olasılığı $1 - 1/n$ olarak hesaplanmaktadır. Örnek verinin belirtilen kümede yer alma ve yer almama olasılıklarının çarpımı aşağıda verilen eşitlik 1 yardımıyla hesaplanmaktadır (Torgo, 2016).

$$\left(1 - \frac{1}{n}\right)^n \approx e^{-1} = 0{,}368 \qquad \text{Eşitlik 1}$$

Yukarıda verilen eşitlikte *e* ile gösterilen sembol doğal logaritmanın tabanıdır ve 2,7168 değerine eşittir. Bu eşitlik belirli bir örneğin herhangi bir şekilde örnekleme seçilmeme olasılığını göstermektedir. Böylece oldukça büyük bir veri kümesi için test kümesi veri setinin yaklaşık %36,80'ini kapsayacaktır ve dolayısıyla eğitim kümesi elinizdeki veri setinin %63,20'sini kapsayacaktır. Daha önce gizemli bir sayı olarak tanımlamış olduğumuz 0,632 Bootstrap sayısının nereden geldiğini öğrenmiş bulunmaktayız (Witten, Frank ve Hall, 2016).

Eğitim kümesi üzerinden bir öğrenme yöntemini eğiterek elde edilen bu sayısal değer ve test kümesi üzerinden elde edilen hata değeri gerçek hatanın kötümser bir tahmini olacaktır. Çünkü eğitim seti, boyutu n olmasına rağmen örneklerin %63'ünü kapsamaktadır ve 10 katlı çapraz geçerlemede örnekleri %90'ının kullandığı düşünüldüğünde bunun daha büyük bir oran olmadığı görülmektedir. Bunu telafi etmek için eğitim kümesi içindeki örnekler üzerinden yeniden yerleştirme işlemi sonucunda elde edilen hata ile test kümesi üzerinden elde dilen hata oranı ile birleştirilmektedir. Yeniden yerleştirme sonucunda elde edilen hata değeri gerçek hata değerinin oldukça iyimser bir tahmini olacağı için tek başına hata terimi olarak kullanılmaması gerekmektedir. Ancak Bootstrap yöntemi hata değerini eğitim ve test verileri üzerinden elde edilen hataların bir arada kullanılması yoluyla hesaplamaktadır ve aşağıda gösterilen Eşitlik 2 nihai bir hata değeri olarak ortaya çıkmaktadır.

$$\text{error} = 0{,}632 \times e_{\text{testverileri}} + 0{,}368 \times e_{\text{eğitim verileri}} \qquad \text{Eşitlik 2}$$

Ardından tüm Bootstrap yöntemi eğitim kümesi için farklı örneklemlerin yer değiştirilmesi suretiyle birkaç kez tekrarlanmaktadır ve elde edilen sonuçların ortalaması alınmaktadır.

Bootstrap yöntemi küçük veri setleri için hata oranını tahmin etmesi bakımından en iyi yöntem olarak kabul edilebilir. Buna rağmen tıpkı bir tanesini kapsam dışı bırakma yönteminde olduğu gibi tamamen yapay ve özel bir durum üzerinden sahip olduğu dezavantajın ne olduğu gösterilebilir. Bunun için daha önce ele aldığımız veri kümesi tamamiyle tesadüfi olarak iki farklı kümeye ayrılmış olsun. Bu durumda herhangi bir tahminde bulunma kuralı için gerçek hata oranı %50 olsun. Ancak eğitim kümesini hatırlayan bir taslak öğrenme yöntemi yeniden

yerleştirme sonucunda hatası (e=0) sıfır olan %100 başarılı bir sonuç elde edilecektir ve 0,632 Bootstrap yöntemi bunu 0,368 ile ağırlıklandırarak 0,632 × %50 + 0,368 × %0 olmak üzere genel hata terimi %31,60 olarak hesaplanacaktır. Ancak bu değer yanıltıcı şekilde gerçek hata değerinden daha iyimser olacaktır.

Performansın tahmin edilmesi. Veri madenciliği büyük miktarda verinin içinden bilgi çıkarmak olarak tanımlandığı için madenciliğe benzetilmektedir (Han ve Kamber, 2006). Veri madenciliği teknikleri, karar vermede yardımcı olan gizli örüntüleri ve ilişkileri keşfetmek için büyük miktarda veriyle başa çıkabilmek amacıyla yaygın olarak kullanılmaktadır. Veritabanlarından bilgi keşfi için sınıflama, kümeleme, regresyon, yapay zeka, sinir ağları, ilişki kuralları, karar ağaçları, genetik algoritma, en yakın komşuluk yöntemi gibi çeşitli algoritmalar ve teknikler kullanılmaktadır (Ahmed ve Elaraby, 2014). Veri madenciliğinde yapılacak analizler klasik istatistiksel yöntemlerde olduğu gibi betimleyici veya tahmin edici olarak ikiye ayrılmaktadır. Betimleyici veri madenciliği, büyük veri kümesinde gizlenmiş örüntüleri keşfetmek ve mantıklı kararlar alabilmek için kümeleme ve birliktelik kuralları gibi yöntemleri kullanmaktadır. Tahmine dayalı veri madenciliği, yeni bir veri kümesinin sınıfını tahmin etmek için karar ağacı, sinir ağları, kural kümeleri ve destek vektörleri gibi yöntemleri kullanarak modeller oluşturmaktadır (Mishra, Kumar ve Gupta, 2014). Her bir farklı öğrenme yönteminin performansını tahmin edebilmek için farklı bir ölçüt kullanılmaktadır.

İstatistiksel olarak başarılı veya başarısız diye sınıflanan iki sonuçla ilgilenildiğinde bu deneye ilişkin dağılıma Bernoulli* deneyi veye Bernoulli dağılımı adı verilir (Chen ve Liu, 1997). Bu dağılım için yapılan denemelere Bernoulli denemeleri denir. Bernoulli dağılımına uygun bir deneme için, deneyler aynı koşullarda tekrarlanabilir özelliklere sahip olmalı ve denemeler birbirlerinden bağımsız olmalıdır. Tekli deney olarak bilinen Bernoulli deneyinin birbirinden bağımsız olarak n defa tekrar edilmesi sonrasında çoklu deney olarak bilinen Binom dağılımı oluşmaktadır (Zhang, Hong ve Balakrishnan, 2017).

Bernoulli denemesinde başarı oranları sırasıyla p ve p.(1-q) olan iki deney için ortalama ve varyans değerleri hesaplanabilmektedir. Eğer Bernoulli

* Bernoulli dağılımı olarak da bilinen bu süreçte bir deneyde sadece başarı ve başarısızlık gibi iki sonuç bulunmaktadır. Birbirinden bağımsız denemeler için başarı olasılığı p ve başarısızlık q (1-p) olacak şekilde ihtimallerin toplamı 1 olacaktır. $f(x) = \begin{cases} 1-p, & x=0 \\ p, & x=1 \end{cases}$ 0 < p < 1

sürecinden N tane deney alınırsa ve bunlardan S tanesinde başarılı olunursa bu durumda beklenen başarı oranı f=S/N olacaktır. Daha büyük N değerleri için rastgele bir değişkenin dağılımı normal dağılıma yaklaşmaktadır (Evans, Hastings ve Peacock, 2000). Ortalaması sıfır olan rastgele bir X değişkeninin 2z güven aralığında olasılık değeri aşağıda Eşitlik 3'te gösterilmektedir.

$$P [-z \leq x \leq z] = c \qquad \text{Eşitlik 3}$$

Normal bir dağılım için olasılık değeri olan c ile buna karşılık gelen z değerleri çoğu istatistik kitabının arka kısmında yer alan tablolarda kolaylıkla görülebilmektedir. Buna rağmen tablolar geleneksel olarak biraz farklı anlama gelebilmektedirler. Örneğin x değişkeni bu aralığın dışında ve yalnızca aralığın üst kısmında değer almaz. Bu Pr [x ≥ z] durumunda geçerli olmaktadır. Tek kuyruklu olasılık olarak adlandırılan bu yaklaşımda yalnızca dağılımın üst kuyruğuna atıfta bulunulmaktadır. Şekil 8'de görüldüğü gibi normal dağılım simetrik olduğundan tek kuyruk için olasılıklar Pr [x ≤ z] veya Pr [x ≥ z] şeklinde tanımlanmaktadır.

Şekil 8. Tek kuyruk testi için olasılık dağılımları.

Bunların yanında Tablo 7'de farklı olasılık değerleri ve bu değerlere karşılık gelen standart z değerleri gösterilmektedir. Güven aralığının belirlenmesinde $\bar{x} \mp z.\ \sigma/\sqrt{n}$ genel formülü kullanılmaktadır (Jiang ve Zhao, 2014).

Tablo 7

Normal Dağılım için Güven Aralıkları

Pr [x ≥ z]	z
%0,1	3,09
%0,5	2,58
%1	2,33
%5	1,65
%10	1,28
%20	0,84
%40	0,25

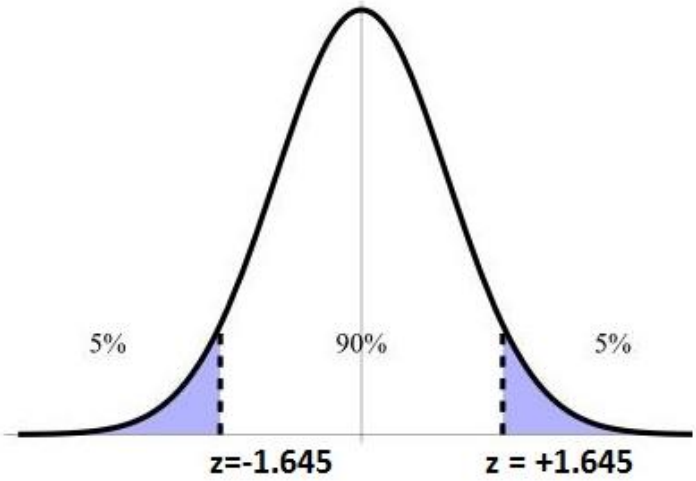
Diğer normal dağılım tablolarında olduğu gibi burada da ortalama 0 ve standart sapma 1 olarak kabul edilir. Alternatif olarak z değerlerinin ortalama ve standart sapma tarafından hesaplandığı söylenebilir. Bu sebeple P [x ≥ z] = %5 gösterimi bize x değerinin ortalamadan 1,65 standart sapma kadar büyük olma olasılığının %5 olduğunu gösterir. Dağılım simetrik olduğu için x değerinin ortalamanın 1,65 standart sapmadan fazla veya eksik olma olasılığı %10 olacaktır.

P [-1,65 ≤ x ≤ 1,65] = %90

Bu aşamada tek yapmamız gereken f olarak tanımladığımız rastgele değişkeni ortalamasını sıfır ve varyansını birim haline dönüştürebilmektedir. Bu işlemi f değerinden p ortalama değerini çıkararak ve $\sqrt{p.(1-q)\,/N}$ standart sapma değerine bölerek gerçekleştirmekteyiz. Böylece;

$$P\left[-z < \frac{f-p}{\sqrt{p.(1-p)/N}} < z \right] = c$$

Güven aralıklarını bulabilmek için gerekli yöntemin ne olduğunu belirledikten sonra sabit bir c değerinin güven aralığı için bu aralığa karşılık gelen z değerini bulmak gerekecektir. Tablo 7'yi kullanma aşamasında önce 1'den c değeri çıkarılır ve sonra sonuç ikiye bölünür. Böylece c=%90 değeri için geriye kalan %10'luk kısmı ikiye böleceğiniz için tabloda %5 değerine bakmamız yeterli olacaktır. Şekil 9 incelenerek bu durum daha iyi anlaşılacaktır.



Şekil 9. Güven aralıkları ve olasılık dağılımı.

Şekilde görüldüğü üzere çift kuyruk testi uygulandığı için c= %90 için alt ve üst kuyruklarda %5'lik ihtimallerin toplamı %10 yapacağı için P [x≥ z] = %5 değerine karşılık gelen z değeri esas alınmaktadır. Doğrusal enterpolasyon* (ara değerleme) yöntemi güven aralığının orta güven seviyeleri için kullanılabilmektedir. Bunun için daha önce verilen eşitsizlik sistemine eşittir işaretini ekleyin ve p değeri için bir eşitlik elde edebilmek amacıyla bu eşitsizlik sisteminde p'yi yalnız bırakınız.

Son adımda ikinci dereceden denklemi çözmemiz gerekmektedir. Her ne kadar bu denklemi çözmek zor olmasa da belirlenen güven aralığı içerisinden zor bir şekilde ifade edilebilen bir yapıya sahiptir.

$$p = \left(f + \frac{z^2}{2.N} \mp z.\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) / \left(1 + \frac{Z^2}{N} \right).$$

Yukarıda verilen eşitlikle $\mp$ işareti güven aralığının alt ve üst sınırlarını göstermektedir. Ne kadar da formül karışık görünse de belirli durumlarda bu eşitlikten sonuç elde etmek zor değildir.

Elde edilen bu eşitlik yardımıyla bazı sayısal örneklerin güven aralıklarını hesaplanabilir. Örneğin N=1000 için f=%75 olarak kabul edildiğinde c=%80 için Tablo 7'deki z değeri 1,28 olacak ve bu değerler için p olasılığının güven aralığı [0.732,0.767] olacaktır. Benzer şekilde N=100 için diğer c ve f değerleri aynı olduğu zaman güven aralığı [0.691,0.801] olacak şekilde biraz daha

* İki değerin ortasında yer alan ve ara değer olarak isimlendirilen sayısal değeri hesaplamak için kullanılan bir yöntemin adıdır.

genişleyecektir. Unutmayınız ki normal dağılım varsayımı sadece N>100 gibi büyük N değerleri için geçerlidir. Ancak N=10 için f=%75 ve c=%80 değerleri esas alındığında [0.549,0.881] şeklinde elde edilecek güven aralığına şüphe ile yaklaşılması gerekmektedir. Bunun sebebi N>100 varsayımının karşılanmamış olmasıdır (Witten ve Frank, 2005).

Kestirimin Gücü: Hatanın Büyüklüğünü Dikkate Alma

Bu bölüme kadar tartışılan değerlendirme işlemleri yanlış karar vermeyi ya da yanlış sınıflama yapmayı göz önüne almamaktadır (hesaba katmamaktadırlar). Hataların ne düzeyde olduğunu dikkate almadan sınıflandırma oranlarını en iyi hale getirmeye çalışmak genellikle sizi hatalı sonuçlara götürecektir. Bu aşamada iki farklı türdeki hatadan kaynaklanan sonuçlar birbirinden farklı anlamlara sahip olacaktır. Veri madenciliğinde doğru ve yanlış tahminler için maliyet tanımı göz önüne alındığında, en düşük beklenen maliyetle bir örneğin her bir sınıfta bulunma durumu için koşullu olasılık değeri kullanılarak hangi sınıfta yer alacağının tahmin edilmektedir. Maliyete duyarlı karar verme sürecinin özü, mevcut sınıftan farklı bir sınıfın daha olası olduğu durumlarda bile sanki bir sınıfın doğruymuş gibi olmasıdır. Örneğin çok mantıklı görünmesine rağmen çok büyük bir kredi transfer işlemini onaylamamak hatanın büyüklüğünü dikkate alan bir yaklaşımdır (Elkan, 2001).

Farklı hataların aynı anlama gelmediğine ilişkin başka bir örnek kredi borçlarından görülebilir. Borcunu ödemeyen birine kredi vererek yapılan hata daha önce hiç kredi çekmemiş birine kredi vermeyerek yapılan hatadan oldukça büyüktür. Başka bir örnek ise petrol sızıntısını belirlemede verilebilir. Çevreyi tehdit eden gerçek bir sızıntıyı belirleyememe durumunda yapılan hatanın maliyeti çevreyi tehdit etmeyen bir sızıntıyı tehdit olarak görmekle yapılan hatadan çok daha yüksektir. Hatanın maliyetine ilişkin verilebilecek son bir örnek ise yük tahminine ilişkindir. Aslında o trafoya çarpmayan bir fırtına için hazırlık yaparak yapılan hatanın maliyeti hazırlıksız olarak bu fırtınaya yakalanma durumunda yapılan hatadan oldukça düşüktür. Tüm bu verilen örneklerden çıkarılacak tanı ise hataları ortadan kaldıran bir makine öğrenme yöntemiyle yanlış tanımlama yapmanın maliyeti başarısızlığa uğrayan bir problemi görmezden gelen öğrenme yönteminin hata maliyetinden daha az olacaktır. Daha farklı bir örnek olarak promosyonlara ilişkin kullanıcılara atılan e-postaların maliyeti düşünülebilir. Mailine

cevap vermeyen bir aileye e-posta göndermenin maliyeti mailine cevap verebilecek bir aileye e-posta göndermeme durumundaki maliyete göre daha düşüktür (Witten ve Frank, 2005).

Evet-hayır, ödünç para verme-vermeme, kredi verme -vermeme, şüpheli bir parçayı atık veya değil, vb. şekillerde iki sınıfı olan durumlarda bu duruma ilişkin sonuçlar Tablo 8'de gösterildiği gibi olası 4 farklı şekilde ortaya çıkmaktadır. Doğru pozitif (DP) ve doğru negatif (DN) doğru sınıflama sonuçlarıdır. DP gerçek durum pozitifken test sonucunda pozitif çıkan durumlardır. Başka bir ifadeyle doğru tahmin edilen örnek sayıdır. DN gerçek durum negatifken test sonucunda negatif çıkan durumlardır. Başka bir ifadeyle doğru olarak reddedilen örneklerin sayıdır. Yanlış pozitif (YP) sonuç aslında negatif (hayır) iken yanlış bir şekilde çıktının pozitif (EVET) olarak tahmin edilmesidir. Daha yalın bir ifadeyle yanlış alarm olarak tanımlanabilir. Yanlış negatif (YN) sonuç aslında pozitif iken yanlışlıkla negatif olarak tahmin edilmesidir. Başka bir ifadeyle yanlış olarak tanımlanan örnek sayıdır.

Doğru pozitif ve yanlış pozitif kavramlarının daha iyi anlaşılması için meteoroloji tarafından yapılan hava tahminlerini ele alalım. Havanın ertesi gün için yağışlı olmamasına rağmen bir önceki gün yağışlı şeklinde tahmin edilmesi yanlış pozitiftir. Bu durumda sonuç olarak yanınıza şemsiye almanıza rağmen onu kullanmamış olursunuz ve kaybınız çok fazla değildir. Ancak yanlış negatif ertesi gün için hava yağışlı iken bir önceki gün yağış olmayacak şeklinde tahminde bulunulmasıdır. Bu durumda yanınıza şemsiye almadığınız için aşırı derecede ıslanmış olacaksınız. Bu durumda yanlış negatif sonuçları sizin için yanlış pozitif sonuçlara göre daha önemli olacaktır (Williams, 2011). Doğru pozitif oranı (DPO) gerçek durumu pozitif olan durumların tüm durumlara bölünmesiyle elde edilirken yanlış pozitif oranı gerçek durumu negatif olan durumların tüm durumlara bölünmesiyle elde edilmektedir. Tablo 8'de iki farklı sonucu olan bir tahmine ilişkin gerçek ve test sonuçlarının alabileceği değerler gösterilmektedir (Schwenke ve Schering, 2007).

Tablo 8

İki Sınıflı Bir Tahmine İlişkin Farklı Sonuçlar

		GERÇEK DURUM		
		POZİTİF	NEGATİF	TOPLAM
TEST SONUCU	POZİTİF	DOĞRU POZİTİF (DP)	YANLIŞ POZİTİF (YP)	DP+YP
	NEGATİF	YANLIŞ NEGATİF (YN)	DOĞRU NEGATİF (DN)	YN+DN
	TOPLAM	DP+YN	YP+DN	DP+YP+YN+DN

İki sınıflı bir tahmin süreci sonucunda DP, YP, YN ve DN oranları aşağıda gösterildiği şekilde elde edilmektedir (Elhamahmy, Elmahdy ve Saroit, 2010).

Yanlış pozitif oranı (YPO) : YP / (YP+DN)

Doğru pozitif oranı (DPO) : DP / (DP+YN)

Yanlış negatif oranı (YNO): YN / (YN+DP)

Doğru negatif oranı (DNO) : DN / (DN+YP)

Sonuçlara ilişkin oranların ardından genel başarı oranı doğru sınıflamaların sayısının toplam sınıflama sayısına bölümüdür.

Genel Başarı = $\frac{DP+DN}{DP+DN+YP+YN}$

Bu sınıflama için hata oranı ise 1'den genel başarı oranının çıkarılması sonucunda elde edilmektedir. Duyarlılık ve seçicilik kavramlarının da bilinmesinde yarar vardır.

Duyarlılık: Testin gerçek pozitif durumlar içinden pozitif olan durumları ayırma yeteneğidir.

Duyarlılık (Sensitivity) = $\frac{DP}{DP+YN}$

Seçicilik: Testin gerçek negatif durumlar içinden negatif olan durumları ayırma yeteneğidir.

Seçicilik (Specifity) = $\frac{DN}{DN+YP}$

Yukarıda anlatılan duyarlılık ve seçicilik değerleri, testin araştırılan durumla ilgili olan ve ilgili olmayanları birbirinden ne kadar iyi ayırt edip etmediğini tanımlamaktadır. Çok sınıflı bir tahminde, testten elde edilen sonuçlar karışıklık (*confusion*) matrisi adı verilen iki boyutlu her bir sınıf için satır ve sütunları olan bir tablo üzerinde gösterilmektedir. Matrisin her bir elemanı satırı gerçek değeri, sütunu ise tahmin edilen (test sonucu) örnek sayısını gösterecek şekilde tanımlanmaktadır (bazı durumlarda satırlar test sonuçlarını, sütunlar ise gerçek durumu göstermektedir). İyi sonuçlar esas köşegen üzerindeki (sol üst köşeden sağ alt köşeye çizilen doğru parçası) sayıların oldukça büyük ve köşegen dışındaki elemanların küçük hatta ideal olanı sıfır olması durumunda elde edilir.

Örnek olması bakımından Saa (2016) tarafından yapılan deneysel çalışmadan elde edilen sonuçlar Tablo 9'da gösterilmiştir. Dönem not ortalamalarına göre öğrenciler a=mükemmel (3,60-4,00), b=oldukça iyi (3,00-3,59), c=iyi (2,50-2,99) ve d=geçti (2,00-2,50) olacak şekilde 4 farklı sınıfa ayrılmıştır. Toplamda 270 öğrenci ile gerçekleştirilen çalışmada cinsiyet, aile büyüklüğü, gelir durumu, anne-baba eğitim düzeyi ve arkadaşlarla geçirilen süre gibi 20 farklı değişken yardımıyla öğrencilerin hangi sınıfta yer alacakları tahmin edilmeye çalışılmıştır.

Tablo 9

Dört Sınıflı Bir Tahmine İlişkin Farklı Sonuçlar

		Gerçek Sınıf				
		a	b	c	d	hassaslık
Tahmin Edilen sınıf	a	23	12	8	6	46,94
	b	20	40	30	29	33,61
	c	11	10	18	12	35,29
	d	6	19	12	14	27,45

(a) C4.5

		Gerçek Sınıf				
		a	b	c	d	hassaslık
Tahmin Edilen sınıf	a	20	12	7	6	44,44
	b	25	39	35	34	29,32
	c	9	11	18	8	39,13
	d	6	19	8	13	28,26

(b) ID3

Tablo 9 dört sınıflı bir örneğe ilişkin sınıflama sonuçlarını göstermektedir. Birinci örnek uygulamada C4.5 algoritması altında toplam 270 (onaltı hücrede yer alan sayıların toplamı) örneğin 23+40+18+14 tanesi (esas köşegen üzerindeki sayılar toplamı) doğru olarak tahmin edilmiş ve başarı oranı %35,19 (95/270)

olarak hesaplanmıştır. İkinci örnek uygulamada ID3 algoritması altında toplam 270 örneğin 90 tanesi doğru tahmin edilmiş ve başarını oranı %33,33 (90/270) olarak belirlenmiştir. Buna göre genel olarak ID3 algoritmasının doğru sınıflama oranı daha yüksek olarak belirlenmiştir. Bunun yanında her iki öğrenme yönteminin farklı sınıf değerleri için hassaslık ölçütlerinin farklılık gösterdiği belirlenmiştir.

Kappa İstatistiği. Kappa istatistiği ya da Kappa değeri beklenen ve gözlenen değerlerin karşılaştırıldığı sayısal bir değerdir. Bunun yanında tesadüfi şans faktörünü de hesaba kattığından daha az yanıltıcı bir değerdir. Beklenen ve gözlenen doğruluk hesabı Kappa istatistiğinde aynı anda kullanılmakta ve karışıklık matrisi üzerinden kolaylıkla belirlenebilmektedir. Konunun anlaşılması bakımından aşağıdaki örneğin incelenmesi faydalı olacaktır. Bu örnekte makine öğrenme yönteminin hayvanların etiketleri üzerindeki verilere göre onları kedi ve köpek olarak sınıflandırdığını düşünün. Eğitim alanında sütundaki değerler puanlayıcılardan biri iken satırdaki değerler puanlayıcılardan diğeri olarak kabul edilse de makine öğrenmede puanlayıcılardan biri gerçek değere, diğeri ise tahmin edilen (veya test sonucu) değere karşılık gelecektir. Nihai olarak satır ve sütunlardan hangisinde gerçek hangisinde tahmin edilen değerin yer alacağı fark etmese de açıklık olması bakımından sütunlar gerçek değerleri satırlar ise makine öğrenme yöntemi tarafından tahmin edilen ya da başka bir ifadeyle beklenen değerleri göstersin. Hayvanların gerçekte ne olduğu sütunlarda, makine öğrenme yöntemi tarafından tahmin edilen sınıflar ise satırlarda olmak üzere sonuçlar Tablo 10'da gösterilmiştir.

Tablo 10

Hayvanların Türlerine İlişkin Gerçek Durum ve Tahmin Sonuçları

		Gerçek Durum	
		KEDİ	KÖPEK
Tahmin Edilen Sonuç	KEDİ	10	7
	KÖPEK	5	8

Tablo olarak verilen matrise göre toplamda 30 hayvandan oluşan bir örnek veri setinin olduğu ve bunlardan ilk sütunda yer alan 15 tanesinin gerçekte kedi iken geri kalan 15 tanesinin köpek olduğu görülmektedir. Makine öğrenme yöntemiyle toplam 30 hayvandan 17 (10+7) tanesi kedi olarak tahmin edilirken 13 (5+8) tanesi köpek olarak tahmin edilmiştir.

Gözlenen doğruluk oranı basitçe tüm matris üzerinden doğru olarak sınıflanmış örneklerin sayısının toplam örnek sayısına bölümü sonucunda elde edilmektedir. Esas köşegen üzerinde yer alan ve aslında kedi iken makine öğrenme yöntemi tarafından kedi olarak tahmin edilen 10 hayvan ile aslında köpek iken makine öğrenme yöntemi tarafından köpek olarak tahmin edilen 8 hayvanın toplam hayvan sayısına bölümü ile (10+8/30=0,6) sınıflama için gözlenen doğruluk oranı %60 olacaktır. Kappa istatistiğini elde edebilmek için bir tane daha değere ihtiyaç vardır. Bu ise beklenen doğruluk değeridir.

Beklenen doğruluk verilen matrise dayalı olarak herhangi bir tesadüfi sınıflandırıcının başarılı olması anlamına gelmektedir. Beklenen doğruluk değerini elde edebilmek için önce kediler için hesaplanan frekans değerini beklenen frekans değeri ile çarpıp daha sonra toplam gözlem sayısına bölmemiz gerekmektedir. Örnek sınıflamada gerçekte 15 kedi sınıflandırıcı tarafından 17 olarak hesaplanmış ve toplam kedi-köpek sayısı 30 olduğu için bu değer 15*17/30=8,5 olarak hesaplanır. Aynı işlem köpekler için yapılırsa aslında 15 köpek olmasına rağmen sınıflandırıcı tarafından 13 köpek olduğu tahmin edildiği için bu değer 15*13/30=6,5 olarak hesaplanır. Son aşamada bu iki değer toplanarak tekrardan toplam gözlem sayısına bölünerek beklenen doğruluk oranı elde edilir. Verilen örnekte (8.5+6.5)/30=0.5 olduğu için beklenen doğruluk oranı %50 olarak hesaplanmıştır. Son olarak beklenen (0.50) ve gözlenen (0.60) doğruluk oranlarının belirlenmesinin ardından Kappa istatistiği aşağıda verilen eşitlik yardımıyla hesaplanmaktadır.

$$Kappa\ (\kappa) = \frac{Gözlenen\ doğruluk - Beklenen\ doğruluk}{1 - Beklenen\ doğruluk}$$

Verilen formüle göre örnek uygulama için hesaplanan Kappa istatistiği (0.60-0.50) / (1-0.50) = 0.20 değerine eşit olacaktır. Böylece özünde Kappa istatistiği makine öğrenme yöntemi tarafından tesadüfi bir sınıflandırıcının beklenen doğruluk değerini kontrol ederek elinizdeki örneklerin gerçeğe ne kadar yakın bir şekilde sınıflandırıldığının ölçüsünü vermektedir.

Kappa istatistiğinin yorumlanması konusunda kesin standart bir yorum olmasa da genel olarak 0.00-0.20 arası düşük, 0.21-0.40 arası kayda değer, 0.41-0.60 arası orta düzeyde; 0.61-0.80 arası önemli ve 0.81-1.00 arası mükemmel olarak tanımlanmaktadır (Landis ve koch, 1977). Ancak başka bir sınıflamada ise

0.75'ten büyük Kappa değerleri mükemmel, 0.40-0.75 arası makul düzeyde iyi ve 0.40'dan aşağı olanlar zayıf olarak tanımlanmaktadır (Fleisis,1971). Özet olarak, Kappa istatistiği veri setindeki beklenen ve gözlenen sınıflara ayrılma işleminde anlaşmayı ölçmek için kullanılan bir değerlendirmedir ve şansla doğru sınıflama oranını düzeltmeye çalışır. Ancak daha önce bahsedilen doğrudan sınıflama oranlarında şansla doğru sınıflama veya puanlayıcılar arası şans eseri anlaşma durumlarını dikkate almamaktadır.

Sayısal Tahminin Değerlendirilmesi. Bu bölüme kadar anlatılan tüm değerlendirme ölçütleri sayısal tahminden ziyade sınıflama sonuçlarıyla ilgilidir. Performansın değerlendirilmesi için eğitim seti kullanmak yerine bekletme ve çapraz geçerlik gibi birbirinden bağımsız testler kullanırken uygulanan temel ilkeler sayısal tahminlerde de eşit derecede uygundur. Ancak hata oranları dikkate alınarak elde edilen temel değerlendirme ölçüleri bu bölümde uygun olmamaktadır. Çünkü hatalar sadece var ya da yok şeklinde değil de farklı büyüklüklerde ortaya çıkmaktadır (Witten ve Frank, 2005).

Bazı alternatif ölçü birimleri sayısal tahminin başarısını değerlendirebilmek amacıyla kullanılmaktadır. Test örnekleri için tahmin edilen değerler P_1, P_2, P_3, … , P_n ve bunların gerçek değerleri a_1, a_2, a_3, … , a_n olsun. Bu noktada P_i ortalaması daha önce belirli bir tahmini *i.* test örneği için tahminin sayısal değerine atıfta bulunmaktadır.

Ortalama karesel hata temel ve yaygın olarak kullanılan bir ölçüttür ve bazen bu değerin karekökü tahmin edilen değerin kendisi ile aynı boyutta olması amacıyla kullanabilmektedir. Doğrusal regresyon gibi birçok matematiksel yöntem ortalama karesel hatayı kullanmaktadır çünkü matematiksel olarak müdahale edilebilecek en kolay ölçüt durumundadır ve matematikçilerin ifadesiyle "iyi huylu" bir ölçüttür.

Ortalama mutlak hata alternatif bir ölçüttür. Burada hataların işareti dikkate alınmadan sadece sayısal değerlerinin ortalaması alınmaktadır. Ortalama karesel hata uç değerlerin etkisini daha da abartma eğilimindeyken mutlak hata da böyle bir durum söz konusu değildir. Mutlak hata hesaplanırken farklı büyüklükteki hatalar büyüklüklerine göre eşit şekilde değerlendirilirler.

Bazen de mutlak hata yerine bağıl hata değeri daha önemli olabilmektedir. Örneğin %10'luk hata değeri, 500 tahminden 50 tanesinin hatalı ise aynı öneme sahiptir ve 2 tahminde %2'lik bir hata varsa mutlak hataların ortalaması anlamlı olmayacak ve bağıl hatalar daha uygun olacaktır. Bu etki hataların kareleri veya ortalama mutlak hata hesaplamasında bağıl hataları kullanarak göz önüne alınmaktadır.

Tablo 11'de gösterilen bağıl hatalar biraz farklı bir anlama sahiptir. Hata değeri, basit bir tahminleyicinin kullanılması halinde göreceli olarak ne elde edileceğine bağlı olarak hesaplanmaktadır. Burada bahsi geçen tahminleyicinin eğitim veri setinden elde edilen gerçek değerlerin ortalamasıdır. Böylece bağıl kareli hata toplam kareli hata olur ve varsayılan tahminleyici tarafından üretilen toplam kareli hataya bölünerek bu değer normalleştirmektedir.

Bir sonraki hata ölçütü bağıl mutlak hata (göreceli mutlak hata) adıyla bilinir ve aslında benzer normalleştirme işlemiyle elde edilmiş mutlak hataların toplamıdır. Bahsedilen bu üç bağıl hata ölçütlerinde hatalar, ortalama değerleri tahmin eden basit tahminleyici tarafından normalleştirilmektedir (Witten, Frank ve Hall, 2016).

Tablo 11'de gösterilen son hata ölçütü, a'lar ile p'ler arasındaki ilişkiyi istatistiksel olarak belirleyen korelasyon katsayısıdır. Korelasyon katsayısı mükemmel ilişki anlamına gelen 1 ile ilişki yok anlamına gelen 0 arasında değer almaktadır. Eğer iki değişken arasında negatif yönde ve mükemmel düzeyde bir ilişki varsa korelasyon katsayısı -1 olmaktadır (Tabachnick ve Fidel, 2007). Elbette mantığa uygun tahmin yöntemleri için negatif korelasyon değerleri elde edilmeyecektir. Bu aşamada korelasyon diğer ölçütlerden biraz farklıdır çünkü korelasyon katsayısı ölçekten bağımsızdır ve belirli bir tahminlerden elde edilen veri kümesi üzerinde çalışıyorsanız ve tüm tahminler sabit bir faktör ile çarpılıyor ve gerçek değerler değişmiyorsa, bu durumda elde edeceğiniz hatalar da değişmeyecektir. Bu sabit faktör, hem pay kısmında bulunan her S_{PA} hem de payda da bulunan her S_P ifadesinde bulunacağı için, gerekli sadeleştirme işlemi sonrası ortadan kaybolacaktır. Ancak bu uygulama normalleştirme işlemine rağmen bağıl hata değerleri için doğru değildir. Bunun sebebi tüm tahmin edilen değerleri büyük bir sabit sayı ile çarparsanız, bu durumda gerçek değerler ile

tahmini değerler arasındaki fark hızla değişecek ve benzer şekilde hata yüzdeleri de değişecektir (Hossin ve Sulaiman, 2015).

Tablo 11

Sayısal Tahminin Değerlendirmesinde Kullanılan Performans Ölçütleri

SAYISAL TAHMİN İÇİN PERFORMANS ÖLÇÜTLERİ*	
PERFORMANS ÖLÇÜTÜ	FORMÜL
Ortalama karesel hata	$\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}$
Ortalama hataların kare kökü	$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$
Ortalama mutlak hata	$\frac{\lvert(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2\rvert}{n}$
Kareli bağıl hata	$\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(p_1 - \bar{a})^2 + \cdots + (p_n - \bar{a})^2}$ iken $\bar{a} = \frac{1}{n}\sum_i a_i$
Bağıl hataların karekökü	$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(p_1 - \bar{a})^2 + \cdots + (p_n - \bar{a})^2}}$
Göreceli mutlak hata	$\frac{\lvert(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2\rvert}{\lvert(p_1 - \bar{a})^2 + \cdots + (p_n - \bar{a})^2\rvert}$
Korelasyon katsayısı	$\frac{S_{PA}}{\sqrt{S_P S_A}}$ iken $S_{PA} = \frac{\sum_i (p_i - \bar{p}) * (a_i - \bar{a})}{n-1}$ $S_P = \frac{\sum_i (p_i - \bar{p})^2}{n-1}$ ve $S_A = \frac{\sum_i (a_i - \bar{a})^2}{n-1}$

* p değerleri özelliğin tahmin edilen değerlerini ve a değerleri ise gerçek değerleri göstermektedir.

Tablo 11'de gösterilen ölçütlerden hangisinin daha uygun olduğu üzerinde çalışılan duruma bağlı olarak belirlenebilmektedir. Bu değerlendirme ölçütlerinden farklı hata değerleri yaptğınız çalışmaya ve kullandığınız değişkenlerin durumuna göre farklılık gösterecektir. Kareli hata ölçütleri ile hataların karekökü ölçütleri büyük uyuşmazlıkları küçüklere kıyasla daha ağırlıklandırılabilinirken mutlak hata ölçütleri bunu yapamazlar. Hataların karekökünü almak tahmin edilen değerlerin aynı boyuta indirgenmesinin sağlayacaktır. Bağıl hata rakamları çıktı değişkenin temel tahmin edilebilirliğini veya tahmin edilemezliğini dengelemeye çalışmaktadır;

eğer tahminler ortalama değere oldukça yakın bir yayılım gösterirse, tahminimizin iyi olmasını beklersiniz ve bağıl hata bunun telafisini sağlamaktadır. Aksi taktirde, bir durum için elde ettiğiniz hata değeri diğer durum için elde ettiğiniz hata değerinden daha farklı olacaktır. Oprea (2014) tarafından yapılan deneysel çalışmada insan kaynakları, pazarlama ve eğitim alanı olmak üzere 3 farklı veri seti üzerinden ID3, Naive Bayes, Çok katmanlı sınıflama ve k-enyakın komşuluk yöntemlerinin farklı değerlendirme ölçütlerine göre nasıl bir performans gösterdiği araştırılmıştır. Çalışma sonucunda elde edilen bulgular Tablo 12'de gösterilmiştir.

Tablo 12

Dört farklı sayısal tahmin modelleri için performans ölçütleri

	Veri seti	ID3	Naive Bayes	Multilayer perceptron	K–Nearest Neighbor
Doğruluk oranı	İnsan kaynakları	93,39	78,74	91,09	93,39
	Pazarlama	99,29	82,27	96,45	99,29
	Eğitim	90,67	65,30	90,38	90,09
Hata	İnsan kaynakları	6,61	21,26	8,91	6,61
	Pazarlama	0,71	17,73	3,55	0,71
	Eğitim	9,33	34,69	9,62	9,91
F-ölçütü	İnsan kaynakları	0,93	0,77	0,90	0,93
	Pazarlama	0,99	0,82	0,96	0,99
	Eğitim	0,89	0,65	0,90	0,89
Roc eğrisi alanı	İnsan kaynakları	0,99	0,91	0,97	0,95
	Pazarlama	1,00	0,92	0,97	0,99
	Eğitim	0,99	0,82	0,99	0,91

Tablo 12'de elde edilen sonuçlara göre doğruluk oranları dikkate alındığında 3 farklı veri setinde de en iyi sonuçları ID3 öğrenme yöntemi tarafından elde edildiği görülmektedir. Sınıflama doğruluğu bakımından en az başarılı olan yöntem ise Naive Bayes olarak belirlenmiştir. Hata ölçütleri bakımından en düşük hata değerlerinin ID3 algoritması tarafından elde edilirken en yüksek hata değerleri Naive Bayes yöntemi tarafından hesaplanmaktadır. F-ölçütlerine göre karşılaştırma yapıldığında ID3 ve k-enyakın komşuluk yöntemleri aynı düzeyde başarı gösterirken en düşük geçerlik değeri Naive Bayes tarafından elde edilmiştir. Son olarak ROC eğrisi altında kalan alanlar karşılaştırıldığında en başarılı yöntemlerin sırasıyla ID3, Çok katmanlı sınıflama, k-enyakın komuşuk ve Naive Bayes olduğu görülmektedir. Buna göre farklı veri setleri kullanılsa da

karşılaştırılan 4 yöntemden en başarılı öğrenme yönteminin ID3 algortimasına dayalı sınıflama yöntemleri olduğu ve ele alınan veri setlerinde en başarısız olan yöntemin ise Naive Bayes olduğu belirlenmiştir.

Sayısal tahminle ilgili iki farklı öğrenme yöntemini karşılaştırırken, daha önce veri madenciliği yöntemlerinin karşılaştırılması başlığı altında verilen ve t-testi gibi istatistiksel yöntemlere dayanan kıyaslama işlemleri burada da uygulanabilmektedir. Buradaki tek farklılık artık başarı oranı yerine ortalama hataların karekökü gibi uygun bir performans ölçütü kullanılarak anlamlılık testleri yardımıyla iki yöntem arasındaki farklılık test edilmektedir.

Veri Madenciliği ile İlgili Araştırmalar

Liu ve Schumann (2005) tarafından yapılan çalışmada veri madenciliğinde değişken sayısını azaltmak için kullanılan dört farklı yöntemin sınıflama ve tahminleme süreçlerindeki performansları karşılaştırılmıştır. M5, yapay sinir ağları, lojistik regresyon ve en yakın komşuluk algoritmalarının karşılaştırıldığı çalışmada 72 farklı özellik kullanılmış ve en yakın komşuluk yönteminin dışındaki diğer yöntemlerde özellik sayısı arttıkça hata oranlarının azaldığı belirlenmiştir. Bunun yanında CON (Consistency-based Feature Selection Algorithm) ve WRP (The Wrapper Feature Selection Algorithm) algoritmalarının tüm yöntemler için değişken sayısını azaltmada etkili olduğu tespit edilmiştir. Gerçek veriler üzerinden kredi puanlarının hesaplandığı çalışmada en yakın komşuluk yönteminin değişken sayısından en fazla etkilenen yöntem olduğu ortaya çıkarılmıştır.

Tadesse, Wilhite, Hayes, Harms ve Goddard (2005) tarafından yapılan çalışmada VM yöntemleri kullanılarak yaklaşık 50 yıllık veriler yardımıyla okyanus parametreleri ve kuraklık endeksleri arasındaki ilişkiler belirlenmeye çalışılmıştır. MOWCATL (Minimal Occurrences With Constraints and Time Lags) algoritmasının kullanıldığı çalışmada birliktelik kuralları yardımıyla ardışık tekrar eden olaylar arasındaki ilişkiler belirlenmiştir. Böylece uzun süreli kuraklıkta okyanussal ve atmosferik olayların belirleyici olduğu ortaya çıkarılmıştır. Çalışmada ayrıca kısa ve uzun süreli tahminlerde VM yöntemlerinin zaman serilerinden daha esnek bir yapıya sahip olduğu belirlenmiştir.

C. Romero, Ventura, Hervas, ve Gonzales (2006) tarafından yapılan çalışmada öğrencilerin ders başarısını tahmin etmek amacıyla bir grup sınıflama algoritmasının performansı karşılaştırılmıştır. Araştırmacılar öğrencilerin ders başarısı; geçti, kaldı, iyi, mükemmel şeklinde dörtlü kategorik bir yapıya dönüştürülerek kullanılmıştır. Algoritma performansları 10 katlı çapraz geçerlilik yöntemi ile genelleştirilerek değerlendirilmiştir. Araştırma bulguları incelendiğinde sınıflama algoritmalarının doğru sınıflama oranlarının %60-%70 arasında değiştiği belirlenmiştir. Araştırmacılar doğru sınıflama oranının %70'in üzerine çıkmamasının öğrencilerin tamamının katılmadığı değişkenlerden kaynaklanabileceğini belirtmişlerdir (ör: öğrencilerin %30'u forumlara veya quizlere katılmamış). Araştırma sonuçları sınıflama sonucu elde edilen kuralların ve bilgilerin öğretmenlerin yeni öğrencileri sınıflamasında ya da karar verme

süreçlerinde kullanılabileceğini belirtmişlerdir. Bunun yanında araştırmacılar öğretmenlerin bu bilgileri sorun yaşayan öğrencilere zamanında müdahale etmek amacıyla kullanabileceklerini belirtmişlerdir.

Aydın (2007) tarafından yapılan çalışmada Anadolu Üniversitesi Uzaktan Eğitim Sisteminde eğitim gören öğrencilere ilişkin farklı kaynaklardaki veriler bir araya getirilerek veri madenciliği uygulaması gerçekleştirilmiştir. Çalışmada öğrenci başarısını tahmin etmeye yönelik C5.0 karar ağacı algoritmasının kullanıldığı bir tahmin modeli önerilmiştir. Önerilen modelin karar kuralları sisteme entegre edilerek öğrenci başarı tahmini amacıyla kullanılabileceği öngörülmektedir. Mezun olan öğrencilere yönelik çalışmada "K-means" algoritması kullanılarak beş küme elde edilmiştir. Kümeleme analizi ile elde edilen bilgilerin bilgisayar kullanımı ve öğrenci başarısı arasındaki ilişkiyi doğrular nitelikte olduğu görülmüştür.

Arabacı (2007) tarafından çalışmada örnek veri tabanında bilgi keşfi sürecinin yapılması ve modelin değerlendirilmesinde kullanılan iki ayrı programın karşılaştırılması amaçlanmıştır. Çalışmada hastanedeki servislerden gönderilen idrar, kan vb. örneklerin incelenmesi sonucu elde edilen gram negatif basillere ait verileri, WEKA ve YALE programlarını kullanarak karşılaştırmıştır. Çalışmada Yale programının Weka'ya kıyasla büyük veri setlerinde depolama sorunu yaşadığı ve diğer koşullar aynı tutulduğunda her iki programın da benzer sonuçları ürettiği belirlenmiştir. Bunun yanında WEKA programının YALE programından yaklaşık 4 kat daha yavaş olduğu ortaya çıkarılmıştır. Çalışma yöntem karşılaştırmadan ziyade iki programın performansını karşılştırmak üzereni yapılandırılmıştır.

Gschwind (2007) tarafından yapılan çalışmada veri madenciliği yöntemleri yardımıyla kiracıların geç ödeme davranışları tahmin edilmeye çalışılmıştır. Son 10 yılda yaklaşık 9,5 milyon kişiden elde edilen veriler SAS Enterprise Miner programında lojistik model, karar ağacı ve yapay sinir ağı algoritmaları yardımıyla analiz edilmiştir. Analiz sonucunda modelleri karşılaştırmak amacıyla kullanılan logaritmik artış grafiğine (lift chart) göre verilerin %10'luk kısmına karar ağacının ve yapay sinir ağlarının aynı düzeyde tahmin yaparken diğer kısımlarda yapay sinir ağlarının daha doğru tahmin yaptığı belirlenmiştir.

Hayashi, Hsieh ve Setiono (2009) tarafından yapılan çalışmada veri madenciliği yöntemleri ile fast food franchise için tüketici tercihini analiz etmek amacıyla karar ağacı ve yapay sinir ağı yöntemleri karşılaştırılmıştır. Bu kapsamda marka tercihlerini belirleyen faktörleri anlamak için Tayvan'daki 800 katılımcıdan toplanan veriler karar ağacı ve sinir ağı modelleri yardımıyla test edilmiştir. Markayı tercih eden tüketiciler ile markayı almayan tüketiciler arasında ayrım yapmak için sınıflandırma kuralları oluşturulmuş ve hem karar ağacı hem de sinir ağı modelleri eğitim verilerinin %80'den fazla olması durumunda 0,70 ve daha yüksek düzeyde doğru sınıflama yaptıkları belirlenmiştir. Ancak yapay sinir ağı modellerinin kural karmaşıklığı ve girdi değişkeninin çok olması durumunda daha iyi yordama yaptığı belirlenmiştir.

Li ve Sarkar (2009) tarafından yapılan çalışmada veri madenciliği yöntemlerinden karar ağacı ve sınıflama algoritmalarının kategorik ve sürekli değişkenler yardımıyla farklı koşullar altında nasıl bir güce sahip olduğu belirlenmeye çalışılmıştır. Sınıflama özelliğinin değişmesi durumunda sınıflarda yer alan birey sayılarının nasıl bir değişim gösterdiğinin araştırıldığı çalışmada C4.5 ve Lojistik regresyon yöntemleri karşılaştırılmış ve lojistik regresyon yönteminin altı farklı veri setinin tamamında daha iyi sonuçlar elde ettiği belirlenmiştir. Çalışmada ayrıca veri setleri arasında değişiklik yapılması durumunda sınıflama hatalarının da arttığı belirlenmiştir.

Aydoğdu (2011) tarafından yapılan çalışmada elektronik öğrenme ortamlarının veri madenciliği teknikleri ile değerlendirilmesi amaçlanmıştır. Elektronik öğrenme modelleri geliştirmek ve elektronik öğrenme sistemlerinin verimliliğine katkı sağlayan temel başarı faktörlerinin belirlenmeye çalışıldığı çalışmada elde edilen veriler veri madenciliği methotları ile SPSS programı kullanılarak analiz edilmiştir. Çalışma sonucunda elde edilen bulgulara göre bağımsız faktörlerin birçoğunun başarıdaki varyansı farklı seviyelerde açıklayabildiği sonucu çıkarılmıştır. Çalışmada ayrıca elektronik dersin içeriği, sertifikalı olup olmama gibi özelliklerin elektronik öğrenme başarısına daha güçlü etkisi olduğu görülmüştür.

Ehrenberg (2011) tarafından yapılan çalışmada hastalıklar arasındaki ilişkileri belirlemek amacıyla veri madenciliği yöntemlerinden metin madenciliği tekniği ile 10 yıllık ve toplam 4700 hastaya ait veriler yardımıyla aslında aralarında

ilişki olmadığı düşünülen migren ve saç dökülmesi gibi hastalıklar arasında ilişki olduğu belirlenmiştir. Bunun yanında hastalıkları harita üzerinde sınıflayarak hastaların hangi hastalıklar üzerinde kümelendiği ve bu hastaların hangi özelliklere sahip olduğu incelenerek aslında ilişkili olduğu düşünülen bazı hastalıkların arasında bir ilişki olmadığı veya aslında ilişkisiz olduğu düşünülen bazı hastalıklar arasında önemli ilişkiler olduğu tespit edilmiştir.

Abdous, He ve Yen (2012) tarafından yapılan çalışmada VM yöntemleri ile interaktif eğitim ve final başarıları arasındaki ilişkiler belirlenmeye çalışılmıştır. Çalışmada öğrencilerin çevrimiçi öğrenme davranışlarını ve derslerindeki performanslarını analiz etmek için regresyon ve eğitsel veri madenciliği yöntemlerini kullanan karma bir yaklaşım izlenmiştir. Öğrencilerin tartışmaya katılımı ve oturum açma sıklıklarının yanı sıra öğretmenlerine gönderdikleri sohbet mesajları ve soru sayısı, öğrencilerin final notlarıyla birlikte analiz edilmiştir. SPSS Clementine programının kullanıldığı çalışmada öğretmene gönderilen mesaj ve öğrencilerin kendi aralarındaki konuşma sıklığının final başarıları üzerinde etkili olduğu belirlenmiştir.

Kurt ve Erdem (2012) tarafından yapılan çalışmada başarılı ve başarısız öğrencilerin profilleri belirlenerek uygun önlem ve çözümler önerilmesi amaçlanmıştır. Öğrencilerin akademik başarılarına etki eden faktörlerin belirlenmesinde öğrencilerin kişisel, sosyal, ekonomik ve barınmayla ilgili demografik özelliklerini içeren toplam 38 soruluk Likert tipi anket hazırlanmıştır. Öğrencilerin başarılarına etki eden faktörleri bulabilmek amacıyla CRT, Chaid, Neural Network, Apriori, K-Means modelleri denenerek amaçlanan sonuçlara ulaşılmaya çalışılmıştır. Çalışma sonucunda veri Madenciliği yöntemlerinin oldukça etkili sonuçlar verdiği görülmüştür.

Lopez ve diğerleri (2012) tarafından yapılan çalışmada öğrencilerin Moodle öğrenme yönetim sistemindeki forum kullanım verilerinin ders başarısının önemli bir göstergesi olup olmadığının belirlenmesi amaçlanmıştır. Bir diğer araştırma problemi ise sınıf değişkeninin (ders başarısı) bilinmediği durumlarda kümeleme algoritmaları ile aynı sonuca ulaşılıp ulaşılmayacağının test edilmesi yönündedir. Araştırmacılar, öğrencilerin forum katılımları ile ilgili olarak 8 değişken belirlemişlerdir ve bu değişkenleri kullanarak öğrencilerin ders başarılarını (geçti – kaldı) tahmin etmeye çalışmışlardır. Bu amaçla birçok sınıflama algoritmasını

karşılaştırarak doğru sınıflama oranına göre performanslarını belirlemişlerdir. Araştırma bulgularında 8 değişken kullanarak yapılan sınıflamada en iyi sınıflama oranı %87,7 ile BayesNet algoritması ile elde edilirken yapılan nitelik (özellik) seçme işlemi sonucu belirlenen 6 değişken ile yapılan sınıflamada Naive Bayes Simple algoritması %89,4 doğru sınıflama oranına ulaşılmıştır. Kümeleme algoritmalarında ise Maksimum Beklenti (Expectation Maximization=EM) algoritması ile %89,4 oranında doğru sınıflama sağlanmıştır.

Yung, Hsu ve Rice (2012) tarafından yapılan çalışmada öğrencilerin öğrenme kayıtları, demografik özellikleri ve dönem sonunda yapılan anketler yardımıyla K-12 programının değerlendirilmesi amaçlanmıştır. Toplam 7.539 öğrenciden elde edilen veriler veri madenciliği sınıflama ve tahminleme algoritmaları yardımıyla test edilmiştir. Çalışmada öğrencilerin ortak özelliklerini ortaya çıkarmak için kümeleme analizi uygulanmış ve öğrenci performansını ile ders ve öğretim elemanına yönelik memnuniyet düzeylerini tahmin etmek için karar ağacı analizi uygulanmıştır. Araştırma sonucunda karar vermede ayrıntılı bilgi üretmek için veri madenciliğinin program değerlendirmesine nasıl dahil edilebileceğini gösterilmiştir. Bunun yanında öğrenci başarılarını yordamak amacıyla elde edilen karar ağacına göre sırasıyla okulun türü, cinsiyet ve yaş değişkenlerinin en iyi yordayıcılar olduğu belirlenmiştir.

Angeli ve Valanides (2013) tarafından yapılan çalışmada eğitsel veri madenciliği yöntemleriyle alana bağlı ve alandan bağımsız öğrenenlerin karmaşık problemleri çözebilme becerileri incelenmiştir. Çalışmada öğrencilerin problem çözme performansları ile sıralı problem için bilgisayar modelleme aracı arasındaki etkileşimleri incelenmiştir. Model-It yazılımı üzerinden öğrencilerin değişkenler arasındaki ilişkileri kendilerinin modellemesine izin verilen çalışmada alana bağımlı, yarı bağımlı ve alandan bağımsız olarak sınıflanan üç grup öğrencinin problem çözme performanslarında anlamlı farklılık olduğu belirlenmiştir. Çalışmada alandan bağımsız olarak sınıflanan öğrencilerin problem çözme performanslarının diğerlerinden daha iyi olduğu ortaya çıkarılmıştır.

Tollenaar ve Heijden (2013) tarafından yapılan çalışmada son dönemde suç işlemiş bireylerin mahkûmiyet hikâyeleri kullanılarak genel suç tekrarı, şiddeti tekrar etme ve cinsel suçu tekrar etme şeklinde üç tür sınıf oluşturulmuştur. Araştırmada modern istatistiksel tekniklerden veri madenciliği ve makine öğrenme

yönteminden suçluların sınıflarına ilişkin tahminlerin, klasik istatistiksel yöntemlerden lojistik regresyon ve diskriminant analizine göre nasıl bir performans gösterdiğinin belirlenmesi amaçlanmıştır. Bu modeller, geniş performans ölçütleri kapsamında karşılaştırılmış ve sonuçlar klasik yöntemlerin modern çağdaş yöntemlere eşit veya daha iyi olduğunu göstermektedir. Bunun yanında ilk veri setinde lojistik regresyonun genel olarak en iyi sonuçlara sahip olurken ikinci veri setinde lojistik regreyon ve diskriminant analizi benzer sonuçlar ürettiği belirlenmiştir.

Bilen, Hotoman, Aşkın ve Büyüklü (2014) tarafından yapılan çalışmada İstanbul ilinde 2011 yılında LYS sınavına giren 42 farklı lise türü, başarı durumlarına göre kümelenmiş ve kümelere ayrışmada hangi test türlerinin etkili olduğunun belirlenmesi amacıyla eğitsel veri madenciliği yöntemlerinden karar ağacı ve kümeleme teknikleri kullanılmıştır. Çalışma grubunda yer alan okulların farklı puan türlerinin her biri için farklı başarı seviyelerini gösteren toplam 5 kümeye ayrıldığı belirlenmiştir. Fen lisesi, Özel fen liseleri, Anadolu liseleri ve Anadolu öğretmen liseleri test türlerinin tamamı için en yüksek başarı gösteren kümede yer aldıkları belirlenmiştir. Bunun yanında CHAID algoritması yardımıyla oluşturulan karar ağacı modellerinde okulların kümelere ayrılmasında hangi testlerin daha fazla etkiye sahip olduğu belirlenmiştir.

Aksoy (2014) tarafından yapılan çalışmada matematik alanında üstün yetenekli öğrencilerin belirlenmesi için veri madenciliği yöntemlerinden karar ağaçları ve kümeleme analizi yöntemleri kullanılmıştır. Çalışmanın örneklemini 5., 6., 7. ve 8. sınıf ortaokul öğrencisi olmak üzere toplam 735 öğrenci oluşturmakta ve bunların 234 tanesi matematik alanında üstün yetenekli olarak belirlenmiştir. Çalışmada oluşturulan karar ağacının doğruluk oranı % 70,73 olarak belirlenmiş ve sınıflamada en önemli değişkenin çoklu zeka alanı olduğu ortaya çıkmıştır.

Akçapınar (2014) tarafından yapılan çalışmada çevrimiçi öğrenme ortamındaki etkileşim verileri kullanılarak öğrencilerin akademik performanslarının veri madenciliği yöntemleri ile modellenmesi amaçlanmıştır. Dersi bırakma eğilimi olan öğrencilerin ve dönem sonundaki olası başarısızlıkların erkenden tahmini açısından önemli görülen çalışma 76 öğrenci ile 14 haftalık bir yarıyılda gerçekleştirilmiştir. Araştırma sonucunda öğrencilerin çevrimiçi öğrenme ortamındaki etkileşim verileri kullanılarak dönem sonundaki akademik

performanslarının başarılı bir şekilde tahmin edilebileceği belirlenmiştir. CN2 kuralları ve kNN algoritmaları ile dersten kalan ve geçen öğrencilerin %86'sı doğru olarak sınıflamıştır. Öğrencilerin dönem sonu akademik performanslarının daha önceki haftalardan tahmin edilip edilemeyeceği ile ilgili analizler incelendiğinde ise üçüncü hafta gibi kısa bir sürede bunun %74 oranında doğru olarak tahmin edilebileceği görülmüştür.

Jiang, Javaad ve Golab (2015) tarafından yapılan çalışmada mühendislik fakültesi için hazırlık kurslarının 10 yıl boyunca eğitim verdiği 250.000 den fazla öğrenciden elde edilen verilerle kurs kalitesini ve eğitici değerlendirmesini etkileyen faktörler belirlenmeye çalışılmıştır. WEKA programının kullanıldığı çalışmada VM yöntemleri yardımıyla kurs kalitesini belirlemek amacıyla beş farklı bileşen için lojistik denklemler elde edilmiştir. Çalışmada bileşenlerden biri için sorulara verilen yanıt en iyi yordayıcı iken diğer bileşen için kurs düzeyinin en iyi yordayıcı olduğu görülmüştür. Elde edilen bu sonuca göre kurlardaki eğitim kalitesini artırmak için eğitmenlerin tutumlarını ve görsel sunumlarını artırmaları, soruları iyi ve açık bir şekilde cevaplamaları ve değerlendirme yöntemlerini geliştirmeleri gerektiği belirlenmiştir.

Sinharay (2016) tarafından yapılan çalışmada TIMMS sınavına ait veriler üzerinden R programındaki algoritmalar farklı eğitim ve test setlerinde duyarlılık, kesinlik ve kappa istatistikleri bakımından karşılaştırılmıştır. Veri setlerinin büyük veya küçük olması Random forest, Boosting ve Classification tree algoritmalarında çok fazla bir farklılık yaratmazken MLR yönteminde güvenirlik ve hata indeksleri bakımından önemli farklılık oluştuğu gözlemlenmiştir. Çalışmada ayrıca tüm veri setlerinde Random forest yönteminin en iyi performans gösteren öğrenme yöntemi olduğu ve Boosting yönteminin de geleneksel yöntemlerden daha iyi sonuçlar elde ettiği belirlenmiştir.

Abad ve Lopez (2017) tarafından yapılan çalışmada akademik başarıyla ilişkili faktörlerin belirlenmesi amacıyla 11-15 yaş aralığındaki 18.935 öğrenciden elde edilen veriler veri madenciliği yöntemleri ile analiz edilmiştir. Sınıflama tekniklerinden karar ağacı öğrenme algoritmasının kullanıldığı çalışmada başarı üzerinde öğrenciyle ilişkili faktörlerin okul ve sosyal faktörlerden daha etkili olduğu belirlenmiştir. Amerika'nın Meksika eyaletinde gerçekleştirilen çalışmada okullar arasında başarıyı yordayan en önemli değişkenin uyuşturucu kullanımı olduğu ve

sonrasında okuma alışkanlığı, öz saygı ve evdeki kaynakların diğer önemli yordayıcılar olduğu belirlenmiştir.

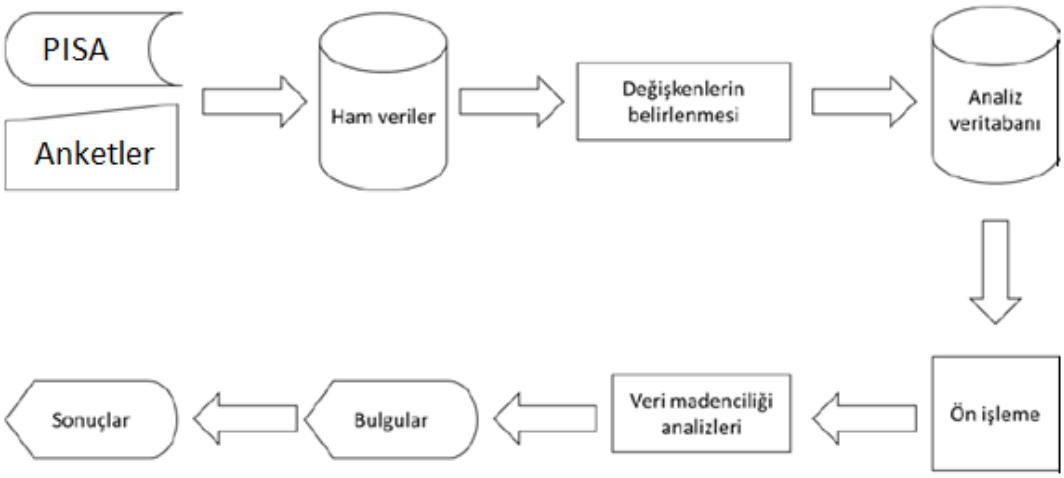
Güldal ve Çakıcı (2017) tarafından yapılan çalışmada sadece 70 öğrenci ile veri madenciliği sınıflandırma algoritmalarından Naive Bayes, Karar Ağacı (C4.5) ve k-enyakın komşuluk yöntemi k=1, 3 ve 5 değeri için farklı değerlendirme ölçütlerine göre karşılaştırılmıştır. Çalışmada en yüksek doğruluk oranları sırasıyla k=3 için en yakın komşuluk, J48, k01 ve k=5 için en yakın komşuluk ve Naive Bayes yöntemleri tarafından elde edilmiştir. F değeri bakımından en iyi sonuçların sıralaması doğruluk oranları ile aynı olup öğrencileri başarılı-başarısız olarak sınıflamada en fazla bilgi veren yordayıcı değişkenlerin sırasıyla ödev gönderme sayısı, portala giriş sayısı, katılma sıklığı, ders notu tıklama sayısı ve cinsiyet olduğu belirlenmiştir. Çalışmada farklı algritmalar yardımıyla öğrencileri başarılı ve başarısız olarak sınıflamada ele alınan sınıflama algoritmalarının doğruluk değerlerinin %55.7 ile %64.3 arasında değişkenlik gösterdiği belirlenmiştir.

İlgili araştırmalar bir bütün olarak değerlendirildiğinde yapay sinir ağlarının (YSA) daha yüksek bir sınıflama oranına sahip olduğu yönündeki bulguların fazla olduğu görülmektedir. Ancak her ne kadar yapay sinir ağlarının Gschwind (2007) tarafından yapılan çalışmada görüldüğü gibi lojistik model ve karar ağaçlarından daha yüksek bir sınıflama yaptığı belirlense de YSA tarafından yapılan sınıflama işlemlerinde değişkenlerin ağdaki etki düzeyleri ve ne kadar katkı sağladıkları her bir değişken için ayrı ayrı belirlenememektedir. YSA'da sadece sınıflama sonuçlarına bakarak ele alınan değişkenlere ilişkin etkili veya az etkili şeklinde genel yorumlar yapılabilmektedir. Ancak eğitim alanında girdi değişkenlerinin yapısı ve çıktı değişkeni üzerindeki özgül etkileri karar alma sürecinde önemli bir yere sahip olduğu için bu konuda oldukça detaylı bilgi veren veri madenciliği yöntemlerinin yapay sinir ağlarından daha değerli olarak düşünülmektedir. Nitekim VM'de her bir değişkenin farklı koşullarda nasıl bir performans gösterdiği rapor edilirken YSA'da sonuca ilişkin kararların verildiği ağ program tarafından oluşturulmakta ve bu süreçteki işlemler hakkında bilgi alınamamaktadır. Bunun yanında Hayashi, Hsieh ve Setiono (2009) karar ağacı ve yapay sinir ağlarının benzer düzeyde sınıflama yaptıklarını belirlendiğinden değişkenler hakkında daha fazla bilgi veren veri madenciliğine dayalı karar ağaçlarının yapısı ve performanslarının incelenmesi önemli görülmektedir. Bu sebeple Weka gibi veri

madenciliği programlarının içerisinde yer alan tüm öğrenme yöntemlerinin farklı koşullar altında performanslarının incelenmesi önemli görülmektedir. Bu sayede ileride yapılacak çalışmalarda araştırmacılara ellerindeki verinin yapısı ve miktarı göz önüne alınarak en uygun öğrenme yönteminin hangisi olduğu gösterilecektir.

Bölüm 3
Yöntem

Bu bölümde araştırmanın türü, çalışma grubu, araştırma koşulları ve verilerin analizine ilişkin bilgilere yer verilmektedir. Bu aşamada araştırmada izlenen süreç ve sürece ilişkin her bir bileşen görsel olarak daha iyi anlaşılması amacıyla Şekil 10'da gösterilmiştir.



Şekil 10. Araştırma süreci.

Ülkemizde 15 yaş grubundaki öğrencilerinin PISA öğrenci anketinde yer alan alt ölçeklere verdikleri yanıtlar yardımıyla Fen okuryazarlığı başarılarını tahmin etmek ve bu süreçte kullanılan veri madenciliği tahminleme yöntemlerinin güvenirlik ve geçerlik değerlerinin ne düzeyde olduğunun belirlenmek amacıyla Şekil 10'da verilen adımlar takip edilmiştir. Buna göre öncelikle PISA öğrenci anketinden ham veriler elde edilmiş ve ardından değişkenler belirlenmiştir. Sonrasında analiz edilecek veriler ön işleme tabi tutulmuş ve veri madenciliği yöntemleri ile analiz edilmiştir. Analiz sonucunda elde edilen bulgular ve sonuçlara dayalı olarak önerilerde bulunulmuştur.

Araştırmanın Türü

Bu araştırmada öğrencilerin PISA 2015 öğrenci anketinde yer alan sorulara verdikleri yanıtlar yardımıyla Fen okuryazarlığı bakımından başarılarını tahmin etmek ve bu süreçte ele alınan güvenirlik ve geçerlik değerlerinin incelemesi amaçlanmaştır. Çalışmanın ilk aşamasında öğrencilerin PISA başarılarını tahmin etmede kullanılan sınıflama yöntemlerinin doğruluk derecesi belirlenmiş ardından

ikinci aşamada farklı algoritmalar yardımıyla elde edilen sonuçların geçerliği incelenmiştir. Araştırma, PISA fen puanları açısından başarılı ve başarısız olarak sınıflanan öğrencilerin duyuşsal özellikleri ölçen alt testler ve sosyodemografik indeksler yardımıyla başarı durumlarına ilişkin tahminleme yapılması bakımından betimsel araştırma modelindedir. Çalışmada veri madenciliği yöntemleriyle elde edilen sınıfların ve PISA fen başarını tahmin etmede kullanılan farklı algoritmaların güvenirlik ve geçerlik değerlerinin ne düzeyde olduğunun belirlenmesi amaçlanmış ve bu kapsamda bir kuramın söz konusu modellerini test etme amacını taşımasından dolayı çalışma betimsel araştırma niteliğindedir (Fraenkel, Wallen ve Hyun, 2012).

Araştırmanın Evreni ve Örneklemi

Araştırmanın amaçları doğrultusunda çalışma grubunu OECD tarafından düzenlenen PISA 2015 sınavına katılan ve örgün eğitime kayıtlı olan 15 yaş grubu öğrencileri oluşturmaktadır. Sınava 72 ülkeden toplam 540.000'e yakın öğrenci katılmış ve bunların 5895 tanesi Türkiye'den katılan öğrencilerdir. PISA 2015 Türkiye uygulamasında 15 yaş grubu öğrenci evreni 1.324.089 öğrenci, uygulamaya katılabilecek ulaşılabilir Türkiye evreni ise 925.366 öğrenci olarak belirlenmiştir. PISA araştırmasında okul örneklemi, tabakalı seçkisiz örnekleme yöntemiyle belirlenmektedir (MEB, 2016).

Veri Toplama Süreci

Araştırmada kullanılan veriler 2017 yılında paylaşıma açılan ve OECD'nin resmi internet sayfası olan http://www.oecd.org/pisa/data/2015database/ adresinden elde edilmiştir. SPSS veri dosyası formatında yer alan öğrenci anketinden ülke kodu TR olan 5895 öğrenciye ilişkin veriler analiz kapsamında veri kaynağı olarak kullanılmıştır.

Veri Toplama Araçları

Araştırmada veri toplama aracı olarak PISA öğrenci anketinde yer alan ve aşağıda isimleri ve kodları yazılı olan değişkenler Tablo 13'de gösterilmiştir.

Tablo 13

Çalışma Kapsamında Kullanılan Değişkenlere İlişkin Tanımlayıcı Bilgiler

Değişkenin Adı	Kodu	Değişkenin Adı	Kodu
Disiplin ortamı	DISCLISCI	Fen öğrenme süresi (hafta)	SMINS
Öğretmen desteği	TEACHSUP	Toplam öğrenme süresi (hafta)	TMINS
Sorgulamaya dayalı fen eğitimi	IBTEACH	Okula ait hissetme	BELONG
Öğretmen merkezli eğitim	TDTEACH	Test kaygısı	ANXTEST
Çevreye duyarlık	ENVAWARE	Tutum, tercih ve öz inançlar	MOTIVAT
Fenden keyif alma	JOYSCIE	Birlikte çalışama isteği	COOPERATE
Fen konularına ilgi	INTBRSCI	Ailenin duygusal desteği	EMOSUPS
Araçsal motivasyon	INSTSCIE	Algılanan geri bildirim	PERFEED
Fen özyeterliği	SCIEEFF	Öğretmenin adil olması	unfairteacher
Bilimsel inançlar	EPIST	Evdeki eğitim kaynakları	HEDRES
Fen yaşantısı	SCIEACT	Evdeki eğitimsel eşyalar	HOMEPOS
Öğrenciden beklenen mesleki statü	BSMJ	BİT kaynakları	ICTRES
Anne eğitim düzeyi	MISCED	Ailenin mal varlığı	WEALTH
Baba eğitim düzeyi	FISCED	Sosyo ekonomik durum indeksi	ESCS
Okul dışı ders çalışma süresi (hafta)	OUTHOURS	Fen Okuryazarlığı	PV1SCIE

Tablo 13 incelendiğinde VM yöntemleri ile fen okuryazarlığını yordamak amacıyla kullanılan bağımsız değişkenler ile fen okuryazarlığı bağımlı değişkenlerinin veri dosyasındaki kodları görülmektedir. Her ne kadar bilgi teknolojileri kullanımı, öğrencinin geçmiş fen deneyimi, ailenin öğrenme için sağladığı destek, öğrencinin değiştirdiği okul sayısı ve ailenin algıladığı okul kalitesi değişkenleri analize dahil edilmek istense de Türkiye örnekleminde bu soruların boş bırakılması ve cevaplanmaması sebebiyle analize dahil edilecek değişken sayısı 29 olarak kararlaştırılmıştır.

Verilerin Analizi

Araştırmanın ilk aşamasında amaç öğrencilerin PISA verilerini kullanarak fen okuryazarlığı performanslarını tahmin edecek bir model oluşturmaktır. Öğrencilerin PISA sınavında gösterdikleri performans “Başarılı” ve “Başarısız” şeklinde kodlandığı için bu bir sınıflama problemidir ve veri madenciliği sınıflama yöntemleri kullanılarak analiz edilmiştir (Romero, OlmoveVentura, 2013). Bu sebeple literatürde sınıflama ve tahmin amacıyla sıklıkla kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge lojistik Regresyon yöntemleri kullanılarak analizler gerçekleştirilmiştir.

Araştırmanın ikinci aşamasında; araştırma kapsamında belirlenen sınıflama algoritmalarıyla başarılı ve başarısız olarak sınıflanan öğrencilerin elde edilen ölçme sonuçlarının güvenirlik ve geçerlik değerleri incelenmiştir.

Çalışmada Weka programından elde edilen sonuçların dış geçerliğini belirlemek amacıyla aynı eğitim ve test veri seti üzerinden her iki programın doğru sınıflama oranları karşılaştırılmıştır. PISA 2015 sınavına katılan 5865 öğrenciden elde edilen verilerin %66,6'sı eğitim (N=3870) ve %33,3'ü test (N=1995) veri seti olarak ayrılmıştır. Weka programında gerçekte başarısız olan 1392 öğrenci öğrenme yöntemi tarafından başarısız ve gerçekte başarılı olan 1357 kişi öğrenme yöntemi tarafından başarılı olarak sınıflandırılarak genel başarı oranı %71,03 (2749/3870) olarak belirlenmiştir. Aynı verilerin Matlab2017b programına eğitim ve test veri seti olarak ayrı ayrı tanıtılmasının ardından eğitim veri setinde gerçekte başarısız olan 1136 öğrenci öğrenme yöntemi tarafından başarısız ve gerçekte başarılı olan 1039 kişi öğrenme yöntemi tarafından başarılı olarak sınıflandırılarak genel başarı oranı %70,30 (2175/3095) olarak belirlenmiştir. Aynı şekilde test veri seti için WEKA programında gerçekte başarısız olan 978 öğrenci öğrenme yöntemi tarafından başarısız ve gerçekte başarılı olan 491 kişi öğrenme yöntemi tarafından başarılı olarak sınıflandırılarak genel başarı oranı %73,63 (1469/1995) olarak belirlenmiştir. Matlab programında gerçekte başarısız olan 203 öğrenci öğrenme yöntemi tarafından başarısız ve gerçekte başarılı olan 192 kişi öğrenme yöntemi tarafından başarılı olarak sınıflandırılarak genel başarı oranı %68,00 (395/581) olarak belirlenmiştir. Genel olarak değerlendirildiğinde Weka programının doğru sınıflama oranı %71,03 iken Matlab programında %69,90 olarak belirlenmiştir. Matlab programında eğitim, test ve doğrulama veri setleri için elde edilen karışıklık matrisi Ek-F'de gösterilmiştir.

Verilerin Analizinde Kullanılan Yazılım ve Algoritmalar

Verilerin analizinde Java tabanlı WEKA 3.8 (Waikato Environment for Knowledge Analysis) paket programlarından yararlanılmıştır. WEKA programı tarımsal verinin işlenmesi amacıyla Yeni Zelanda'daki Waikato Üniversitesi tarafından geliştirilmiştir (Kuyucu, 2012). Verilerin analizinde WEKA programının seçilmesinin sebeblerinden bir tanesi programın farklı alanlarda yaygın olarak kullanılmaya başlaması ve bir diğer sebebi ise sistemin açık kaynak kodlu olmasıdır. Bu sayede araştırmacıların kendilerinin de sisteme kod yazarak farklı

güvenirlik ve geçerlik katsayıları geliştirmelerine olanak sağlanmaktadır. Tablo 14'de Weka programı tarafından kullanılan algoritmalar ve bunların işlevleri gösterilmektedir. Tablo 14 incelendiğinde; Bayesci sınıflandırıcı, ağaçlar, kurallar, tembel sınıflandırıcılar* ve çeşitli kategorileri başlıkları görülmektedir (Witten, Frank ve Hall,2016).

Tablo 14

Weka Programındaki Algoritmalar ve İşlevleri

YÖNTEM	ADI	İŞLEVİ
BAYES	AODE	Ortalaması alınmış, tek bağımlı tahmin
	Bayesnet	BAYES ağlarını öğrenme
	Complement naive bayes	Tamamlayıcı NAIVE BAYES sınıflandırıcısı oluşturma
	Naive bayes	Olasılığa dayalı standart NAIVE BAYES sınıfları
	Naive bayes multinomial	NAIVE BAYES in çok terimli versiyonu
	Naive bayes simple	NAIVE BAYES in basit uygulaması
	Naive bayes updateable	Bir seferde tek bir örneği öğrenen artımlı NAIVE BAYES sınıflandırıcı
AĞAÇLAR	Adtree	Alternatif karar ağaçları oluşturma
	Decision stump	Tek düzeyli karar ağacı oluşturma
	ID3	Temel böl ve yönet karar ağacı alternatif
	J48	C4.5 karar ağacı öğrenmesi
	LMT	Lojistik modele dayalı ağaç oluşturma
	M5p	M5 modeli ağaç öğrenmesi
	Nbtree	Yapraklarda NAIVE BAYES sınıflama yöntemi ile karar ağacı oluşturur
	Randomforest	Rastgele ormanlar oluşturur
	Randomtree	Her düğümde rastgele özelliklerin belirli bir sayısına göre karar ağacı oluşturur
	Reptree	Azaltılmış hatayı kullanan hızlı ağaç öğrenmesi
	Userclassıfıer	Kullanıcılara kendi karar ağaçlarını oluşturma
KURALLAR	Conjunctiverule	Temel birleştirici öğrenme kuralı
	Decisiontable	Çoğunluk sınıfına göre basit karar tablosu oluşturma

* LAZY CLASSIFIER: Bu öğrenme algoritmasında eğitim örnekleri gerçek sınıflama işlemine kadar kullanılmaktadır. Başka bir ifadeyle bir test örneği gelmediği müddetçe sınıflandırıcı eldeki veri seti tarafından eğitilmemektedir.

	Jrip	Hızlı ve etkili kural için RIPPER algoritması
	M5rules	M5 yöntemi kullanılarak ağaçlardan kurallar oluşturma
	Nnge	İç içe geçmemiş genel örnekler kullanarak kural oluşturmada en yakın komşuluk
	Oner	12 sınıflandırıcı
	Part	J4.8 kullanarak oluşturulan kısmi karar ağaçlarından kurallar elde etmek
	Prism	Kurallar için basit kapsama algoritmaları
	Ridor	Dalgalanmalı kural öğrenme
	Zeror	Kategorik ise çoğunluk sınıfını sayısal ise ortama değeri tahmin etme
	Leanstmedsq	Ortalama yerine medyan kullanarak dayanıklı regresyon
	Linearregression	Standart doğrusal regresyon
	Logistic	Doğrusal lojistik regresyon modelleri oluşturma
	Multilayerperceptron	Geri yayılmalı sinir ağı
FONKSİYONLAR	Pace regression	En küçük kareler yöntemine dayalı doğrusal regresyon modeli oluşturma
	Rbfnetwork	Radyo temeli ağ oluşturur
	Simple linear regression	Tek bir özelliğe dayalı olarak doğrusal regresyon modeli öğrenir
	Simple logistic	Seçili özellikle doğrusal lojistik modelleri oluşturur
	SMO	Destek vektör sınıflandırması için ardışık minimal optimizasyon algoritması

En popüler karar ağaçlarından biri olan ID3 Austurya Sidney üniversitesinden J.Ross Quinlen tarafından temeli atılan ve bilgi kazanım ölçütüne (*information gain criterion*) dayalı kararlı bir yöntemdir. Daha sonrasında bağımlı değişkenin sayısal olması, kayıp veriler, gürültülü veriler (noisy data) ve ağaçlardan kural oluşturma gibi birçok yenilik ve iyileştirmeye sahip C4.5 ağaç yöntemi ID3 esas alınarak geliştirilmiştir.

Karar kütüğü (*Decision Stump*) yöntemi Boosting (*artırma*) yöntemleri kullanılarak tasarlanmıştır. Bu yöntemde sonuç değişkeni kategorik veya sayısal olabilmekte ve elinizdeki veri setleri için tek düzeyli ve en az iki dallanması olan ağaç üretirken kayıp veri sorununu bunları üçüncü bir ağaç dalı olarak modele eklemek suretiyle çözebilmektedir.

Rastgele ağaç (*RondomTree*) 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.

Rastgele orman (RandomForest) yönteminde ise rastgele ağaçların bir araya getirilmesiyle orman yapıları elde edilmektedir.

Regresyon ağacı (*REPTree*) bilgi kazanım ve varyans azaltma yöntemiyle bir karar veya regresyon ağacı oluşturmaktadır. Varyans azaltması indirgenmiş hata budaması yöntemiyle gerçekleşmektedir. Bu yöntem C4.5 gibi eksik verileri parçalara ayırmak suretiyle üstesinden gelebilmektedir. Her bir yaprakta yer alacak en düşük örnek sayısı kullanıcılar tarafından belirlenebilir. Maksimum ağaç derinliği dallanma için eğitim setinin minimum oranı ve dallanma için düzey sayısı da seç düğmesinin yanındaki komut satırı tıklanarak açılan pencerede değişiklik yaparak istenildiği şekilde düzenlenebilmektedir.

Naive Bayes Ağaç (*NBTrees*) yöntemi Naive Bayes ve karar ağaçlarının bir araya gelmesiyle oluşturulmuş karma bir yöntemdir. Bu yöntemde yapraklarda Naive Bayes sınıflandırıcısı tarafından sınıflandırılmış örnek sayıları olan ağaçlar oluşturulmaktadır. Ağaç oluşurken, bir düğümden sonra tekrardan bir bölünme olup olmayacağı veya Naive Bayes yönteminin kullanılıp kullanılmayacağı çapraz geçerleme yöntemiyle belirlenmektedir (Kohavi,1996).

Lojistik Model Ağaçları (*LMT*) hedef veya sonuç değişkeni olarak tanımlanan özelliğin iki kategorili veya çok kategorili, sayısal veya sınıflama düzeyinde özellikler ve kayıp veriler ile çalışabilmektedir. Bir düğümde lojistik regresyon fonksiyonunu uyguladığınızda, analiz için kaç iterasyon yapılacağını belirlemek için çapraz geçerleme yöntemi kullanılır ve böylece her düğümde farklı bir iterasyon yerine ağacın tamamı boyunca aynı sayı kullanılmaktadır. Bu işlem çalışma süresini önemli ölçüde artırırken elde edilen sonuçların doğruluğu üzerinde çok az bir etkiye sahiptir. Alternatif olarak ağaç boyunca kullanacak iterasyon sayısını kendiniz ayarlayabilirsiniz. Normalde çapraz geçerlik sayısının azaltılması yanlış sınıflandırma hatasıdır. Fakat bunun yerine olasılıkların ortalama hatalarının karekökü de seçilebilir. Ağaçtaki bölünme ölçütü C4.5'in bilgi miktarına dayalı olarak ya da lojit artık değerlerine göre belirlenmektedir. Burada artık değerlerin saflığını artırmak için uğraşılmaktadır.

Alternatif ağaç (*ADTrees*) yönteminde boosting yöntemi kullanılarak alternatif karar ağacı üretilmektedir ve iki sınıflı problemler için optimizasyon yapılmaktadır. Bu yöntemde ağacın kökünde daima bir tahmin düğümü yer

almaktadır. Ağaçta tekrarlanan iterasyon sayısı veri seti ile istenilen doğruluk-karmaşıklık arasındaki dengeyi sağlayabilecek şekilde ayarlanan bir parametredir. Her iterasyonda ağaca bir tane bölünmüş ve iki tane tahmin olmak üzere toplam üç düğüm eklenir ve bu işlem düğümler birleşmediği sürece devam eder. Varsayılan arama yöntemi tüm sayıların kullanıldığı ayrıntılı arama (*exhaustive search*) yöntemi iken diğerleri buluşsal veya sezgisel olarak tanımlanmaktadırlar ve daha hızlı işlem yapma kapasitesine sahiptirler (Witten ve Frank, 2005).

Bölüm 4

Bulgular ve Yorumlar

Çalışma kapsamında ilgili literatür taraması sonucunda öğrencilerin fen okuryazarlığı üzerinde etkili olduğu belirlenen değişkenlerin isimleri, kodları, en düşük ve en yüksek değerleri ile aritmetik ortalamaları Tablo 15'te gösterilmiştir.

Tablo 15

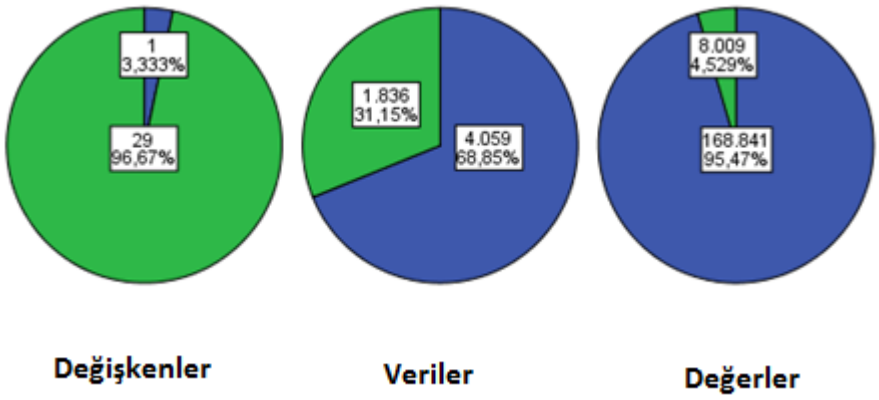
Çalışma Kapsamında Kullanılacak Değişkenlere İlişkin İstatistikler

Değişkenin Adı	Kodu	min	max	ort	kayıp oranı
Disiplin ortamı	DISCLISCI	-2,416	1,883	-0,134	%10,7
Öğretmen desteği	TEACHSUP	-2,719	1,447	0,196	%10,9
Sorgulamaya dayalı fen eğitimi	IBTEACH	-3,340	3,182	0,320	%10,9
Öğretmen merkezli eğitim	TDTEACH	-2,447	2,078	-0,059	%11,1
Çevreye duyarlık	ENVAWARE	-3,376	3,293	0,552	%4,2
Fenden keyif alma	JOYSCIE	-2,115	2,163	0,124	%4,4
Fen konularına ilgi	INTBRSCI	-2,581	2,730	-0,065	%8,5
Araçsal motivasyon	INSTSCIE	-1,930	1,735	0,375	%5,1
Fen özyeterliği	SCIEEFF	-3,756	3,277	0,336	%5,1
Bilimsel inançlar	EPIST	-2,790	2,155	-0,192	%4,8
Fen yaşantısı	SCIEACT	-1,757	3,361	0,687	%5,3
Öğrenciden beklenen mesleki statü	BSMJ	10	89	63,35	%7,7
Anne eğitim düzeyi	MISCED	0	6	2,20	%1,6
Baba eğitim düzeyi	FISCED	0	6	2,67	%1,6
Okul dışı ders çalışma süresi (hafta)	OUTHOURS	0	70	25,54	%8,8
Fen öğrenme süresi (hafta)	SMINS	0	800	198,82	%4,7
Toplam öğrenme süresi (hafta)	TMINS	100	3000	1558,75	%6,5
Okula ait hissetme	BELONG	-3,129	2,612	-0,436	%1,5
Test kaygısı	ANXTEST	-2,505	2,549	0,318	%1,3
Tutum, tercih ve öz inançlar	MOTIVAT	-3,087	1,854	0,613	%1,5
Birlikte çalışama isteği	COOPERATE	-3,332	2,287	0,000	%1,6
Ailenin duygusal desteği	EMOSUPS	-3,078	1,099	-0,267	%1,1
Algılanan geri bildirim	PERFEED	-1,525	2,499	0,351	%11,3
Öğretmenin adil olması	unfairteacher	1	24	10,25	%1,8
Evdeki eğitim kaynakları	HEDRES	-4,370	1,180	-0,583	%1,5
Evdeki eğitimsel eşyalar	HOMEPOS	-6,710	5,150	-1,432	%0,6
BİT kaynakları	ICTRES	-3,270	3,500	-1,190	%1,2
Ailenin mal varlığı	WEALTH	-6,960	4,090	-1,487	%0,9
Sosyo ekonomik durum indeksi	ESCS	-5,130	3,120	-1,447	%0,6
Fen Okuryazarlığı	PV1SCIE	197,72	707,89	422,45	%0,0

Tablo 15 incelendiğinde veri madenciliği yöntemleri ile fen okuryazarlığını yordamak amacıyla kullanılacak değişken sayısının 30 olduğu belirlenmiştir. Her ne kadar bilgi teknolojileri kullanımı, öğrencinin geçmiş fen deneyimi, ailenin öğrenme için sağladığı destek, öğrencinin değiştirdiği okul sayısı ve ailenin algıladığı okul kalitesi değişkenleri analize dahil edilmek istense de Türkiye

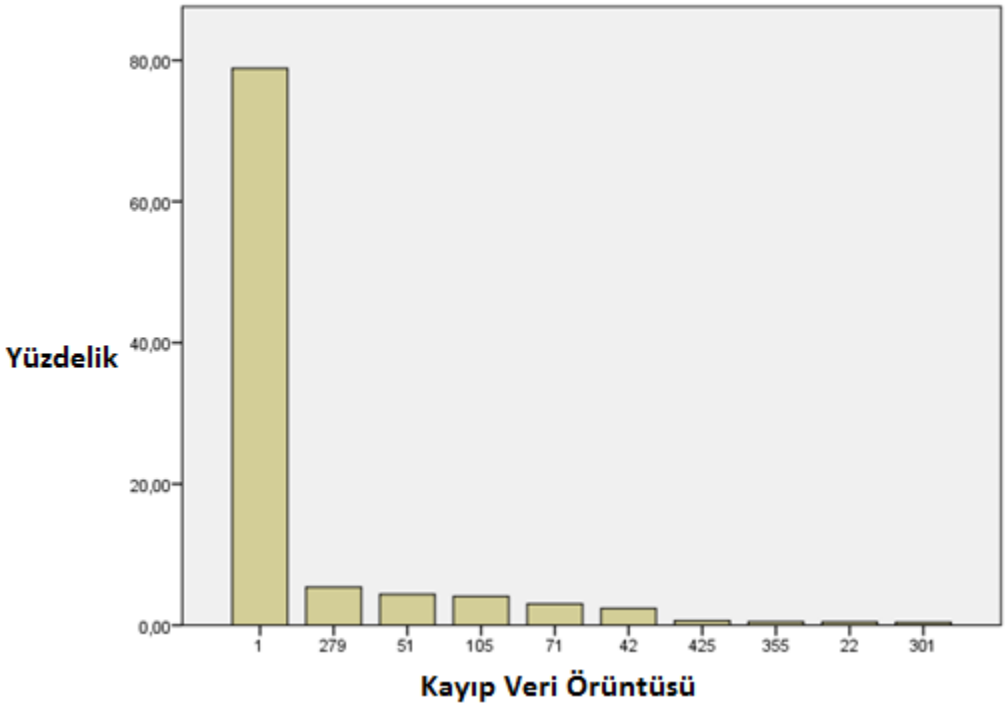
örnekleminde bu soruların boş bırakılması ve cevaplanmaması sebebiyle analize dahil edilecek bağımsız değişken sayısı 29 olarak kararlaştırılmıştır.

Bu işlemin ardından veri setindeki kayıp verilerin dağılımı incelenmiş ve sonuçlar Şekil 11'de gösterilmiştir.



Şekil 11. Kayıp verilere ilişkin özetleyici bilgiler.

Veri dosyasındaki kayıp verilerin oranlarına ilişkin grafikler incelendiğinde birinci daire grafiği veri setindeki her 29 değişkenden 1 tanesinde kayıp veri olduğunu, ikinci grafik 5895 öğrenciden elde edilen verilerden 1836 tanesinde en az 1 tane kayıp veri olduğunu ve üçüncü grafik ise PISA verilerinin tamamı için %4,53 oranında kayıp veri sorunu olduğunu gösterilmektedir. Bunun yanında veri setindeki kayıp verilerin nasıl bir örüntüye sahip olduğunu göstermek için elde edilen sütun grafiği Şekil 12'de gösterilmiştir.



Şekil 12. Kayıp veri örüntüsü grafiği.

Elde edilen bu sonuca göre veri setindeki kayıp verilerin yaklaşık %80'i tek bir örüntüye sahip olduğu ve bu nedenle veri setindeki kayıp veriler arasında bir ilişki olmadığı belirlenmiştir. Kayıp veri dağılımının istatistiksel olarak tamamiyle tesadüfi olup olmadığını belirlemek amacıyla yapılan analiz sonucunda ki-kare 11534,05 serbestlik derecesi 8236 ve belirlenen serbestlik derecesinde ki-kare değeri anlamlı bulunmuştur ($p<.01$). Elde edilen bu sonuca göre kayıp verilerin tamamiyle tesadüfi olarak dağıldığı hipotezi reddedilmiştir. Başka bir ifadeyle tamamiyle tesadüfi olarak dağılmayan kayıp verilerin silinmesi veri kaybına ve dolayısıyla sonuçların yanlı olmasına sebep olacağı için kayıp veri atama yöntemlerinin uygulanmasına karar verilmiştir. Alan yazında gelişmiş kayıp veri atama teknikleri (Soley-Bori, 2013) arasında yer alan Çoklu Atama (*Multiple Imputation=MI*) yöntemi kullanılarak kayıp verilerin yerine SPSS 21 programı tarafından otomatik değer atama yöntemi kullanılarak farklı modeller yardımıyla yeni değerler atanmıştır. Her bir değişken için hangi atama yönteminin kullanıldığı Tablo 16'da gösterilmiştir.

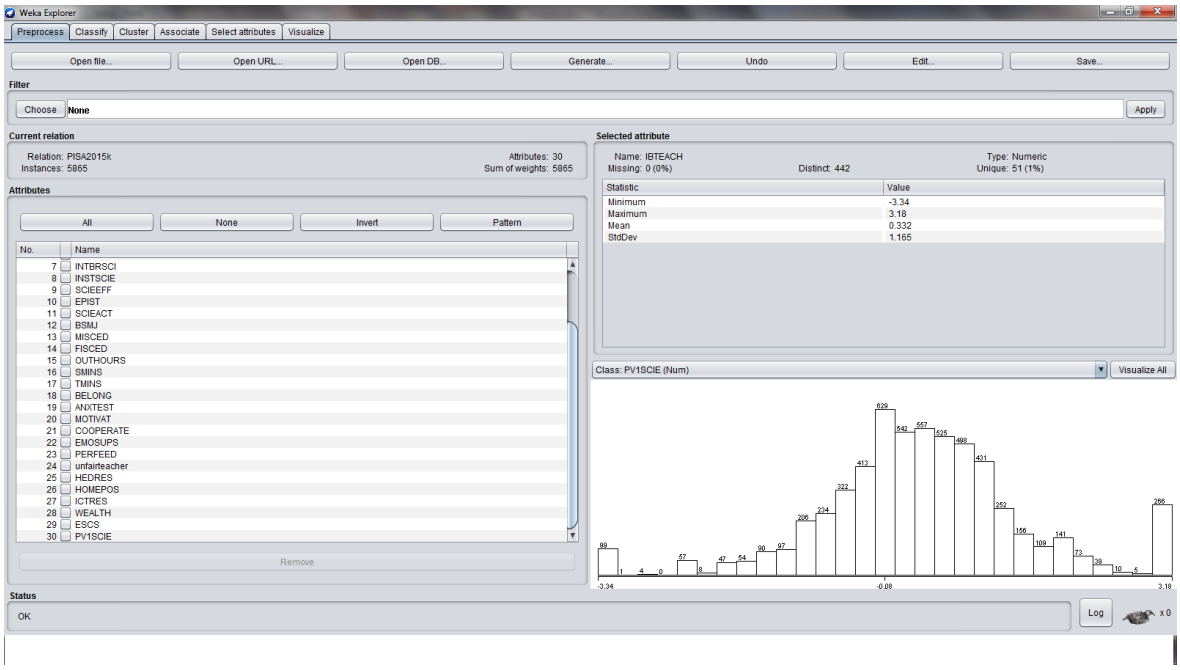
Tablo 16

Kayıp veri atama yöntemine ilişkin sonuçlar

İşlemin Adı	Tanımlama
Atama yöntemi	Tam Koşullu Tanımlama
Atama sayısı	5
Sürekli değişkenler için kullanılan model	Doğrusal Regresyon
Modele dahil edilen etkileşimler	Yok
Kayıp verilerin en yüksek oranı	100,0%
Atama modelinde yer alacak en yüksek parametre sayısı	100

Tablo 16 incelendiğinde kayıp verilerin yerine doğrusal regresyon yöntemiyle değer atandığı görülmektedir. Bu işlemlerin ardından kayıp veri içermeyen tam verilerin Weka programına aktarılması amacıyla gerekli dönüşümler yapılmış ve analiz işlemlerine geçilmiştir.

Öğrencilerin PISA okuryazarlık puanları bakımından Türkiye ortalaması olan 425,00 puanının altında olanlar başarısız (0), bu puanın üzerinde olanlar başarılı (1) olarak kodlanmış ve elde edilen iki kategorili sonuç değişkeni için veri dosyasındaki 5865 öğrencinin 2769'u başarılı (%47,20) olurken 3096'sı başarısız (%52,8) olarak sınıflanmıştır. Öğrencilerin PISA 2015 öğrenci anketinden elde ettikleri sonuçların Weka programına aktarılması sonucunda elde edilen ekran arayüzü Şekil 13'te gösterilmiştir.



Şekil 13. PISA 2015 verilerinin weka programına aktarılması.

Şekil 13'de görüldüğü üzere PISA Fen okuryazarlığını yordamak amacıyla kullanılacak değişken sayısının 29 olması sebebiyle ilk olarak farklı algoritmalar yardımıyla en iyi özellikler belirlenmeye çalışılmıştır. Bu aşamada eğitim veri seti üzerinden kullanılan algoritmalar BestFirst ve Greedy Stepwise olup her bir yöntemle elde edilen sıralama sonuçları Tablo 17'de gösterilmiştir.

Tablo 17

PISA Okuryazarlığını En İyi Yordayan Değişkenler

Sıra No	Best First -forward	Best First -backward	Greedy Stepwise
1	Sorgulamaya dayalı fen eğitimi	Sorgulamaya dayalı fen	Sorgulamaya dayalı fen
2	Çevreye duyarlık	Çevreye duyarlık	Çevreye duyarlık
3	Bilimsel inançlar	Bilimsel inançlar	Bilimsel inançlar
4	Öğrenciden beklenen mesleki statü	Öğrenciden beklenen	Öğrenciden beklenen
5	Okul dışı ders çalışma süresi	Okul dışı ders çalışma	Okul dışı ders çalışma
6	Fen öğrenme süresi	Fen öğrenme süresi	Fen öğrenme süresi
7	Toplam öğrenme süresi	Toplam öğrenme süresi	Toplam öğrenme süresi
8	Öğretmenin adil olması	Öğretmenin adil olması	Öğretmenin adil olması
9	Evdeki eğitim kaynakları	Evdeki eğitim kaynakları	Evdeki eğitim kaynakları
10	BİT Kaynakları	BİT Kaynakları	BİT Kaynakları
11	Sosyo ekonomik durum indeksi	Sosyo ekonomik durum	Sosyo ekonomik durum

Tablo 17 incelendiğinde fen okuryazarlığını yordayan en iyi ilk 11 özelliğin farklı yöntemler kullanılsa da değişmediği belirlenmiştir. Elde edilen bu sonucun geçerliğini belirlemek amacıyla 10 katlı çapraz geçerleme yöntemi uygulanmış ve belirlenen özeliklerin her bir katta nasıl bir performans gösterdikleri Tablo 18'de gösterilmiştir.

Tablo 18

PISA Okuryazarlığını En İyi Yordayan Değişkenler için Çapraz Geçerlik Sonuçları

Sıra No	Özellik	Başarı olduğu düzey sayısı	Başarı oranı (%)
1	**Disiplin ortamı**	**1**	**10**
2	Öğretmen desteği	0	0
3	**Sorgulamaya dayalı fen eğitimi**	**10**	**100**
4	Öğretmen merkezli eğitim	0	0
5	**Çevreye duyarlık**	**10**	**100**
6	Fenden keyif alma	0	0
7	Fen konularına ilgi	0	0
8	Araçsal motivasyon	0	0
9	Fen özyeterliği	0	0

10	**Bilimsel inançlar**	**10**	**100**
11	Fen yaşantısı	0	0
12	**Öğrenciden beklenen mesleki statü**	**10**	**100**
13	Anne eğitim düzeyi	0	0
14	Baba eğitim düzeyi	0	0
15	**Okul dışı ders çalışma süresi (hafta)**	**10**	**100**
16	**Fen öğrenme süresi (hafta)**	**10**	**100**
17	**Toplam öğrenme süresi (hafta)**	**10**	**100**
18	**Okula ait hissetme**	**1**	**10**
19	**Test kaygısı**	**6**	**60**
20	Tutum, tercih ve öz inançlar	0	0
21	**Birlikte çalışama isteği**	**1**	**10**
22	**Ailenin duygusal desteği**	**1**	**10**
23	Algılanan geri bildirim	0	0
24	**Öğretmenin adil olması**	**4**	**40**
25	**Evdeki eğitim kaynakları**	**8**	**80**
26	**Evdeki eğitimsel eşyalar**	**2**	**20**
27	**BİT kaynakları**	**10**	**100**
28	Ailenin mal varlığı	0	0
29	**Sosyo ekonomik durum indeksi**	**10**	**100**

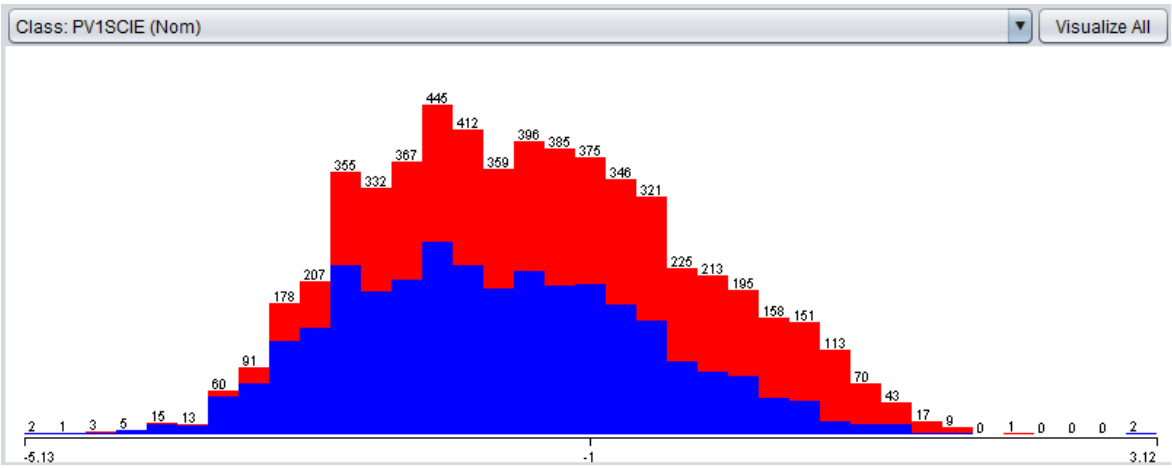
Tablo 18 incelendiğinde sorgulamaya dayalı fen eğitimi, çevreye duyarlık, bilimsel inançlar, öğrenciden beklenen mesleki statü, okul dışı ders çalışma süresi, fen öğrenme süresi, toplam öğrenme süresi, Bilgi ve İletişim Teknolojisi kaynakları ile sosyo ekonomik durum indeksi 10 katlı çapraz geçerlemenin her aşamasında PISA okuryazarlığı üzerinde anlamlı bir etkide bulunurken, evdeki eğitim kaynakları 8, test kaygısı 6, öğretmenin adil olması özelliği 4, evdeki eğitimsel eşyaların 2 ve disiplin ortamı, okula ait hissetme, birlikte çalışma isteği ile ailenin duygusal desteği özelliklerinin 1 katmanda anlamlı etkide bulunduğu belirlenmiştir. Elde edilen bu sonuca göre PISA okuryazarlığını yordayacak değişken sayısı için ölçütün en az 4 katmanda anlamlı bir etkiye sahip olan değişkenler olması sebebiyle modele dahil edilecek değişkenler ile değişken kodları ve başarı olduğu katman sayısı Tablo 19'da gösterilmiştir.

Tablo 19

Veri Madenciliğinde Kullanılacak Değişkenlere İlişkin Özet Bilgiler

Sıra No	Değişkenin Adı	Kodu	min	max	Başarı oranı
1	Sorgulamaya dayalı fen eğitimi	IBTEACH	-3,340	3,182	%100
2	Çevreye duyarlık	ENVAWARE	-3,376	3,293	%100
3	Bilimsel inançlar	EPIST	-2,790	2,155	%100
4	Öğrenciden beklenen mesleki statü	BSMJ	10	89	%100
5	Okul dışı ders çalışma süresi (hafta)	OUTHOURS	0	70	%100
6	Fen öğrenme süresi (hafta)	SMINS	0	800	%100
7	Toplam öğrenme süresi (hafta)	TMINS	100	3000	%100
8	Test kaygısı	ANXTEST	-2,505	2,549	%60
9	Öğretmenin adil olması	unfairteacher	1	24	%40
10	Evdeki eğitim kaynakları	HEDRES	-4,370	1,180	%80
11	BİT kaynakları	ICTRES	-3,270	3,500	%100
12	Sosyo ekonomik durum indeksi	ESCS	-5,130	3,120	%100

Bu işlemlerin sonucunda 12 bağımsız değişken tarafından PISA fen okuryazarlığını yordamak amacıyla kullanılan veri madenciliği yöntemlerinin güvenirlik ve geçerlik değerlerinin karşılaştırılması aşamasına geçilmiştir. Bu aşamada Weka programında PISA fen okuryazarlığı kategorik değişkene dönüştürülmüş ve her bir özellik için mavi ve kırmızı renklerle ilgili özellik için sonuç değişkeninin dağılımı gösterilmektedir. Örnek olması bakımından -5,13 ile 3,12 aralığında değişkenlik gösteren sosyo ekonomik durum indeksi özelliği için başarılı ve başarısız olarak tahmin edilen öğrenci dağılımları Şekil 14'te gösterilmiştir.



Şekil 14. Sosyo ekonomik durum indeksi özelliği için dağılım grafiği.

PISA fen okuryazarlığı bakımından başarılı olarak tanımlananlar kırmızı, başarısız olarak tanımlananlar mavi renkle gösterilmektedir. PISA fen okuryazarlığı bakımından öğrencileri başarılı ve başarısız olarak tahmin etmek amacıyla kullanılacak değişkenlerin belirlenmesinin ardından çalışmanın alt problemlerine ilişkin analizlere geçilmiştir.

Birinci Alt Probleme İlişkin Bulgular

PISA öğrenci anketiyle ölçülen değişkenlerden yararlanarak başarıyı tahmin etmede kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge lojistik Regresyon yöntemleri yardımıyla edilen ölçme sonuçlarının 10 katlı çapraz geçerleme yöntemi yardımıyla elde edilen güvenirlik ve geçerlik değerleri ne düzeydedir?

Çalışmanın birinci alt probleminde veri setini eğitim ve test verisi olarak ayırmadan doğrudan 10 katlı çapraz geçerleme yöntemiyle analizler gerçekleştirilmiştir. Bu aşamada farklı yöntemlerle elde edilen sınıflama sonuçlarının güvenirlik ve geçerlik değerleri rapor edilmiştir.

Decision Stump Yöntemine İlişkin Sonuçlar

Decision Stump yöntemiyle elde edilen sonuçların güvenirlik değerleri hesaplanmış ve sonuçlar Tablo 20'de gösterilmiştir.

Tablo 20

Decision Stump Yöntemiyle Elde Edilen Güvenirlik Değerleri

Yöntem	Sonuç
Doğru olarak sınıflanan örnek sayısı	3744
Doğru sınıflama oranı	63,83
Kappa istatistiği	0,276
Mutlak hatanın ortalaması	0,460
Hataların ortalama karekökü	0,479
Göreceli mutlak hata	%92,32
Göreceli hataların karekökü	%96,10
Toplam Örnek Sayısı	5865

Tablo 20 incelendiğinde PISA Türkiye örnekleminde 5865 öğrenciden 3744 tanesinin doğru olarak sınıflandırıldığı ve doğru sınıflama oranının %63,83 olduğu görülmektedir. Bunun yanında kappa istatistiğinin 0,28 ve mutlak hatanın ortalamasının 0,46 olduğu belirlenmiştir. Bunların haricinde veri madenciliğinde kullanılan geçerlik ölçütlerinden Doğru pozitif, Yanlış pozitif, Duyarlılık, Geri çağırma, F-Ölçütü, Matthews korelasyon katsayısı (MCC), ROC eğrisi altında

kalan alan (ROC Area) ve Duyarlık-Hassalık Alanı (PRC Area) ölçütlerine ilişkin sonuçlar Tablo 21'de gösterilmiştir.

Tablo 21

Decision Stump Yöntemiyle Elde Edilen Geçerlik Ölçütleri

Sınıf	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
Başarısız	0,632	0,355	0,666	0,632	0,649	0,277	0,629	0,621
Başarılı	0,645	0,368	0,611	0,645	0,627	0,277	0,629	0,568
Genel	0,638	0,361	0,640	0,638	0,639	0,277	0,629	0,596

Tablo 21'de görüldüğü üzere başarılı ve başarısız olarak sınıflandırma sonuçları genel olarak incelendiğinde doğru pozitif oranı 0,638, yanlış pozitif oranı 0,361, duyarlık 0,640, geri çağırma 0,638, f-ölçütü 0,639 Matthew korelasyon katsayısı (MKK) 0,277, Roc eğrisi altında kalan alan 0,629 ve duyarlık geri çağırma eğrisi altında kalan alanın 0,596 olduğu belirlenmiştir. Buna göre en yüksek değerlendirme ölçütünün duyarlık ve en düşük değerlendirme ölçütünün MKK olduğu belirlenmiştir.

Hoeffding Tree Yöntemine İlişkin Sonuçlar

Hoeffding Tree yöntemiyle elde edilen güvenirlik değerleri Tablo 22'de gösterilmiştir.

Tablo 22

Hoeffding Tree Yöntemiyle Elde Edilen Güvenirlik Değerleri

Yöntem	Sonuç
Doğru olarak sınıflanan örnek sayısı	4085
Kappa istatistiği	0,390
Mutlak hatanın ortalaması	0,357
Hataların ortalama karekökü	0,455
Göreceli mutlak hata	%71,76
Göreceli hataların karekökü	%91,18
Toplam Örnek Sayısı	5865

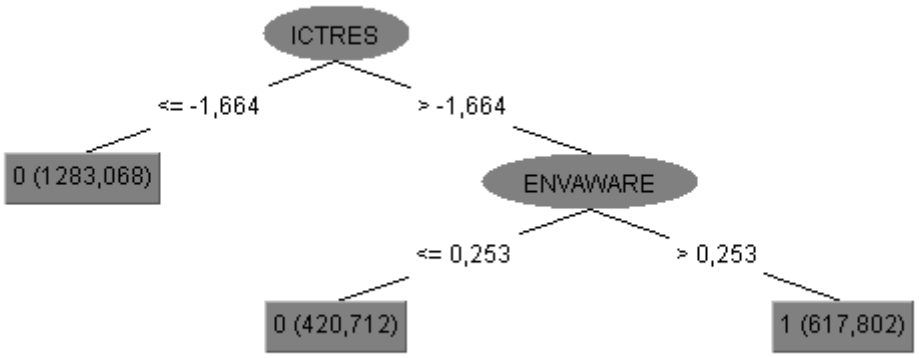
Tablo 22 incelendiğinde PISA Türkiye örnekleminde 5865 öğrenciden 4085 tanesinin doğru olarak sınıflandırıldığı ve doğru sınıflama oranının %69,65 olduğu görülmektedir. Bunun yanında kappa istatistiğinin 0,39 ve mutlak hatanın ortalamasının 0,36 olduğu belirlenmiştir. Bunların yanında sınıflara ilişkin detaylı geçerlik ölçütleri Tablo 23'de gösterilmiştir.

Tablo 23

Hoeffding Tree Yöntemiyle Elde Edilen Geçerlik Ölçütleri

Sınıf	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
Başarısız	0,724	0,335	0,708	0,724	0,716	0,390	0,758	0,750
Başarılı	0,665	0,276	0,683	0,665	0,674	0,390	0,758	0,726
Genel	0,697	0,307	0,696	0,697	0,696	0,390	0,758	0,739

Tablo 23'de görüldüğü üzere başarılı ve başarısız olarak sınıflandırma sonuçları genel olarak incelendiğinde doğru pozitif oranı 0,697, yanlış pozitif oranı 0,307, duyarlık 0,696, geri çağırma 0,697, f-ölçütü 0,696, Matthew korelasyon katsayısı 0,390, Roc eğrisi altında kalan alan 0,758 ve duyarlık geri çağırma eğrisi altında kalan alanın 0,739 olduğu belirlenmiştir. Buna göre en yüksek değerlendirme ölçütünün DP oranı ile geri çağırma ve en düşük değerlendirme ölçütünün YP oranı olduğu belirlenmiştir. Öte yandan Hoefding tree yöntemi için program tarafından oluşturulan ağaç yapısı Şekil 15'te gösterilmiştir.



Şekil 15. Hoeffding tree yöntemiyle elde edilen ağaç yapısı.

PISA fen okuryazarlığı bakımından öğrencileri sınıflamada ilk olarak Bilgi ve İletişim Teknolojileri kaynakları ve sonrasında çevreye duyarlık değişkenlerinin etkili olduğu belirlenmiştir. BİT kaynakları bakımından kesme puanının -1,664 olduğu ve bu değerin altında kalanların başarısız olarak sınıflanırken bu değerin üzerinden olanların çevreye duyarlık puanlarına bakıldığı ve çevreye duyarlık için kesme puanının 0,253 olduğu belirlenmiştir.

J.48 Yöntemine İlişkin Sonuçlar

J.48 yöntemiyle elde edilen güvenirlik değerleri Tablo 24'de gösterilmiştir.

Tablo 24

J.48 Yöntemiyle Elde Edilen Güvenirlik Değerleri

Yöntem	Sonuç
Doğru olarak sınıflanan örnek sayısı	3976
Kappa istatistiği	0,351
Mutlak hatanın ortalaması	0,382
Hataların ortalama karekökü	0,492
Göreceli mutlak hata	%76,79
Göreceli hataların karekökü	%98,67
Toplam Örnek Sayısı	5865

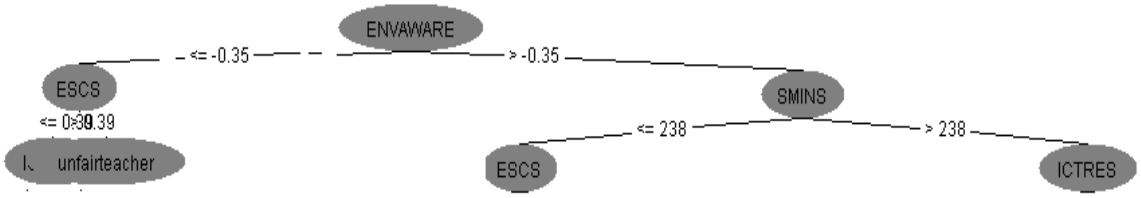
Tablo 24 incelendiğinde PISA Türkiye örnekleminde 5865 öğrenciden 3976 tanesinin doğru olarak sınıflandırıldığı ve doğru sınıflama oranının %67,79 olduğu görülmektedir. Bunun yanında kappa istatistiğinin 0,35 ve mutlak hatanın ortalamasının 0,38 olduğu belirlenmiştir. Bunların yanında sınıflara ilişkin detaylı geçerlik ölçütleri Tablo 25'te gösterilmiştir.

Tablo 25

J.48 Yöntemiyle Elde Edilen Geçerlik Ölçütleri

Sınıf	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
Başarısız	0,732	0,383	0,681	0,732	0,706	0,352	0,694	0,666
Başarılı	0,617	0,268	0,673	0,617	0,644	0,352	0,694	0,630
Genel	0,678	0,328	0,678	0,678	0,677	0,352	0,694	0,649

Tablo 25'te görüldüğü üzere başarılı ve başarısız olarak sınıflandırma sonuçları genel olarak incelendiğinde doğru pozitif oranı 0,678, yanlış pozitif oranı 0,328, duyarlık 0,678, geri çağırma 0,678, F-ölçütü 0,677, Matthew korelasyon katsayısı 0,352, Roc eğrisi altında kalan alan 0,694 ve duyarlık geri çağırma eğrisi altında kalan alanın 0,649 olduğu belirlenmiştir. Buna göre en yüksek değerlendirme ölçütünün ROC alanı ve en düşük değerlendirme ölçütünün YP oranı olduğu belirlenmiştir. Öte yandan J.48 yöntemi için program tarafından oluşturulan ağaç yapısının bir kısmı Şekil 16'da gösterilmiştir.



Şekil 16. J.48 yöntemiyle elde edilen ağaç yapısı.

PISA fen okuryazarlığı bakımından öğrencileri sınıflamada ilk olarak çevreye duyarlık ve sonrasında sosyo ekonomik durum indeksi ile haftalık fen öğrenme süresi değişkenlerinin etkili olduğu belirlenmiştir. Bir alt düzeydeki sınıflamada ise yine sosyo ekonomik durum indeksi ve BİT kaynakları değişkenlerinin etkili olduğu görülmektedir. Çevreye duyarlık bakımından kesme puanının program tarafından ham puanlar üzerinden hesaplanmı ve -0,35 olduğu belirlenmiştir. Bu değerin altında kalanların başarısız olarak sınıflanırken bu değerin üzerinden olanların haftalık fen öğrenme sürelerine bakıldığı ve fen öğrenme süresi için kesme puanının 238 olduğu belirlenmiştir.

Lojistik Model Yöntemine İlişkin Sonuçlar

Lojistik Model (LM) yöntemiyle elde edilen güvenirlik değerleri Tablo 26'da gösterilmiştir.

Tablo 26

Lojistik Model Yöntemiyle Elde Edilen Güvenirlik Değerleri

Yöntem	Sonuç
Doğru olarak sınıflanan örnek sayısı	4125
Doğru sınıflama oranı	70,33
Kappa istatistiği	0,403
Mutlak hatanın ortalaması	0,381
Hataların ortalama karekökü	0,437
Göreceli mutlak hata	%76,58
Göreceli hataların karekökü	%87,70
Toplam Örnek Sayısı	5865

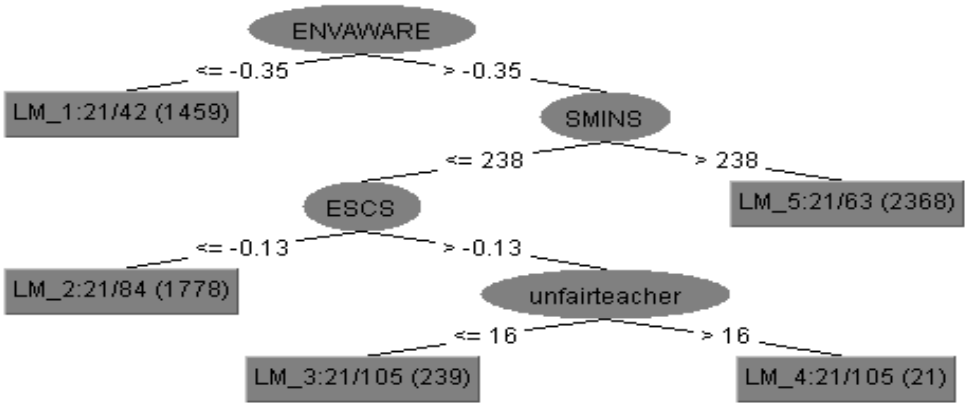
Tablo 26 incelendiğinde PISA Türkiye örnekleminde 5865 öğrenciden 4125 tanesinin doğru olarak sınıflandırıldığı ve doğru sınıflama oranının %70,33 olduğu görülmektedir. Bunun yanında kappa istatistiğinin 0,40 ve mutlak hatanın ortalamasının 0,38 olduğu belirlenmiştir. Bunların yanında sınıflara ilişkin detaylı geçerlik ölçütleri Tablo 27'de gösterilmiştir.

Tablo 27

Lojistik Model Yöntemiyle Elde Edilen Geçerlik Ölçütleri

Sınıf	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
Başarısız	0,739	0,337	0,711	0,739	0,725	0,404	0,778	0,787
Başarılı	0,663	0,261	0,695	0,663	0,679	0,404	0,778	0,743
Genel	0,703	0,301	0,703	0,703	0,703	0,404	0,778	0,767

Tablo 27'de görüldüğü üzere başarılı ve başarısız olarak sınıflandırma sonuçları genel olarak incelendiğinde doğru pozitif oranı 0,703, yanlış pozitif oranı 0,301, duyarlık 0,703, geri çağırma 0,703, F-ölçütü 0,703, Matthew korelasyon katsayısı 0,404, Roc eğrisi altında kalan alan 0,778 ve duyarlık geri çağırma eğrisi altında kalan alanın 0,767 olduğu belirlenmiştir. Buna göre en yüksek değerlendirme ölçütünün ROC alanı ve en düşük değerlendirme ölçütünün YP oranı olduğu belirlenmiştir. Öte yandan Lojistik Model yöntemi için program tarafından oluşturulan ağaç yapısının bir kısmı Şekil 17'de gösterilmiştir.



Şekil 17. Lojistik model yöntemiyle elde edilen ağaç yapısı.

PISA fen okuryazarlığı bakımından öğrencileri sınıflamada ilk olarak çevreye duyarlık ve sonrasında haftalık fen öğrenme süresi özelliğinin etkili olduğu belirlenmiştir. Bir alt düzeydeki sınıflamada ise sosyo ekonomik durum indeksi ve son alt basamakta öğretmenin adil olması değişkeninin etkili olduğu görülmektedir. Çevreye duyarlık bakımından kesme puanının -0,35 olduğu ve bu değerin altında kalanların 1 numaralı lojistik modele göre sınıflanırken bu değerin üzerinden olanların haftalık fen öğrenme sürelerine bakıldığı ve fen öğrenme süresi için

kesme puanının 238 olduğu belirlenmiştir. Elde edilen bu sonuç J.48 yöntemi ile benzerlik göstermektedir. Şekil 15 incelendiğinde karar ağacının yapraklarında LM_1, LM_2, LM_3, LM_4 ve LM_5 şeklinde beş farklı lojistik modele dayalı denklem oluşturulduğu görülmektedir. Geleneksel istatistiksel analizlerde bağımlı değişkenin bağımısız değişkenler tarafından yordanmasına ilişkin tek bir regresyon denklemi oluşturulurken Lojistik modelde beş farklı regresyon denklemi elde edilmektedir. Bu nedenle lojistik denklemin başarı oranını diğer yöntemlerden daha yüksek olmaktadır. Bunun yanında her bir lojistik denklemde sabit sayıların ve değişkenlerin önünde yer alan katsayıların farklılık gösterdiği belirlenmiştir.

REPTree Yöntemine İlişkin Sonuçlar

REPTree yöntemiyle elde edilen güvenirlik değerleri Tablo 28'de gösterilmiştir.

Tablo 28

REPTree Yöntemiyle Elde Edilen Güvenirlik Değerleri

Yöntem	Sonuç
Doğru olarak sınıflanan örnek sayısı	3946
Doğru sınıflama oranı	67,28
Kappa istatistiği	0,341
Mutlak hatanın ortalaması	0,391
Hataların ortalama karekökü	0,472
Göreceli mutlak hata	%78,56
Göreceli hataların karekökü	%94,65
Toplam Örnek Sayısı	5865

Tablo 28 incelendiğinde PISA Türkiye örnekleminde 5865 öğrenciden 3946 tanesinin doğru olarak sınıflandırıldığı ve doğru sınıflama oranının %67,28 olduğu görülmektedir. Bunun yanında kappa istatistiğinin 0,34 ve mutlak hatanın ortalamasının 0,39 olduğu belirlenmiştir. Bunların yanında sınıflara ilişkin detaylı geçerlik ölçütleri Tablo 29'da gösterilmiştir.

Tablo 29

REPTree Yöntemiyle Elde Edilen Geçerlik Ölçütleri

Sınıf	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
Başarısız	0,716	0,376	0,681	0,716	0,698	0,342	0,718	0,715
Başarılı	0,624	0,284	0,663	0,624	0,643	0,342	0,718	0,657
Genel	0,673	0,332	0,672	0,673	0,672	0,342	0,718	0,688

Tablo 29'da görüldüğü üzere başarılı ve başarısız olarak sınıflandırma sonuçları genel olarak incelendiğinde doğru pozitif oranı 0,673, yanlış pozitif oranı 0,332, duyarlık 0,672, geri çağırma 0,673, F-ölçütü 0,672, Matthew korelasyon katsayısı 0,342, Roc eğrisi altında kalan alan 0,718 ve duyarlık geri çağırma eğrisi altında kalan alanın 0,688 olduğu belirlenmiştir. Buna göre en yüksek değerlendirme ölçütünün ROC alanı ve en düşük değerlendirme ölçütünün YP oranı olduğu belirlenmiştir. Öte yandan REPTree yöntemi için program tarafından oluşturulan ağaç yapısının bir kısmı Şekil 18'de gösterilmiştir.

Şekil 18. REPTree yöntemiyle elde edilen ağaç yapısı.

PISA fen okuryazarlığı bakımından öğrencileri sınıflamada ilk olarak haftalık fen öğrenme süresi ve sonrasında çevreye duyarlık ile BİT kaynakları özelliklerinin etkili olduğu belirlenmiştir. Bir alt düzeydeki sınıflamada ise çevreye duyarlık değişkeninin etkili olduğu görülmektedir. Ağacın ilk sınıflamasında haftalık fen öğrenme süresi bakımından kesme puanının 232,5 olduğu ve bu değerin altında kalanların çevreye duyarlık puanlarına bakılırken bu değerin üzerinde olanların BİT kaynakları puanına bakılmaktadır.

Random Forest Yöntemine İlişkin Sonuçlar

Random Forest yöntemiyle elde edilen güvenirlik değerleri Tablo 30'da gösterilmiştir.

Tablo 30

Random Forest Yöntemiyle Elde Edilen Güvenirlik Değerleri

Yöntem	Sonuç
Doğru olarak sınıflanan örnek sayısı	4177
Kappa istatistiği	0,420
Mutlak hatanın ortalaması	0,378
Hataların ortalama karekökü	0,434
Göreceli mutlak hata	%76,01
Göreceli hataların karekökü	%86,92
Toplam Örnek Sayısı	5865

Tablo 30 incelendiğinde PISA Türkiye örnekleminde 5865 öğrenciden 4177 tanesinin doğru olarak sınıflandırıldığı ve doğru sınıflama oranının %71,22 olduğu görülmektedir. Bunun yanında kappa istatistiğinin 0,42 ve mutlak hatanın ortalamasının 0,38 olduğu belirlenmiştir. Bunların yanında sınıflara ilişkin detaylı geçerlik ölçütleri Tablo 31'de gösterilmiştir.

Tablo 31

Random Forest Yöntemiyle Elde Edilen Geçerlik Ölçütleri

Sınıf	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
Başarısız	0,761	0,342	0,713	0,761	0,736	0,421	0,785	0,791
Başarılı	0,658	0,239	0,711	0,658	0,683	0,421	0,785	0,761
Genel	0,712	0,294	0,712	0,712	0,711	0,421	0,785	0,777

Tablo 31'de görüldüğü üzere başarılı ve başarısız olarak sınıflandırma sonuçları genel olarak incelendiğinde doğru pozitif oranı 0,712, yanlış pozitif oranı 0,294, duyarlık 0,712, geri çağırma 0,712, F-ölçütü 0,711, Matthew korelasyon katsayısı 0,421, Roc eğrisi altında kalan alan 0,785 ve duyarlık geri çağırma eğrisi altında kalan alanın 0,777 olduğu belirlenmiştir. Buna göre en yüksek değerlendirme ölçütünün ROC alanı ve en düşük değerlendirme ölçütünün YP oranı olduğu belirlenmiştir.

Random Tree Yöntemine İlişkin Sonuçlar

Random Tree yöntemiyle elde edilen güvenirlik değerleri Tablo 32'de gösterilmiştir.

Tablo 32

Random Tree Yöntemiyle Elde Edilen Güvenirlik Değerleri

Yöntem	Sonuç
Doğru olarak sınıflanan örnek sayısı	3690
Kappa istatistiği	0,256
Mutlak hatanın ortalaması	0,370
Hataların ortalama karekökü	0,609
Göreceli mutlak hata	%74,40
Göreceli hataların karekökü	%121,98
Toplam Örnek Sayısı	5865

Tablo 32 incelendiğinde PISA Türkiye örnekleminde 5865 öğrenciden 3690 tanesinin doğru olarak sınıflandırıldığı ve doğru sınıflama oranının %62,92 olduğu

görülmektedir. Bunun yanında kappa istatistiğinin 0,26 ve mutlak hatanın ortalamasının 0,37 olduğu belirlenmiştir. Bunların yanında sınıflara ilişkin detaylı geçerlik ölçütleri Tablo 33'de gösterilmiştir.

Tablo 33

Random Tree Yöntemiyle Elde Edilen Geçerlik Ölçütleri

Sınıf	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
Başarısız	0,644	0,388	0,650	0,644	0,647	0,256	0,628	0,607
Başarılı	0,612	0,356	0,606	0,612	0,609	0,256	0,628	0,554
Genel	0,629	0,373	0,629	0,629	0,629	0,256	0,628	0,582

Tablo 33'de görüldüğü üzere başarılı ve başarısız olarak sınıflandırma sonuçları genel olarak incelendiğinde doğru pozitif oranı 0,629, yanlış pozitif oranı 0,373, duyarlık 0,629, geri çağırma 0,629, F-ölçütü 0,629, MKK 0,256, ROC eğrisi altında kalan alan 0,628 ve duyarlık geri çağırma eğrisi altında kalan alanın 0,582 olduğu belirlenmiştir. Buna göre en yüksek değerlendirme ölçütlerinin birbirine eşitken en düşük değerlendirme ölçütünün MKK olduğu belirlenmiştir. Öte yandan Random Tree yöntemi için program tarafından oluşturulan ağaç yapısının bir kısmı Şekil 19'da gösterilmiştir.

Şekil 19. Random tree yöntemiyle elde edilen ağaç yapısı.

PISA fen okuryazarlığı bakımından öğrencileri sınıflamada ilk olarak haftalık fen öğrenme süresi ve sonrasında çevreye duyarlık ile Evdeki eğitim kaynakları özelliklerinin etkili olduğu belirlenmiştir. Bir alt düzeydeki sınıflamada ise çevreye duyarlık değişkenini ile toplam öğrenme süresi ve bilimsel inançlar özelliklerinin etkili olduğu görülmektedir. Ağacın ilk sınıflamasında haftalık fen öğrenme süresi bakımından kesme puanının 232,5 olduğu ve bu değerin altında kalanların

çevreye duyarlık puanlarına bakılırken bu değerin üzerinde olanların Evdeki eğitim kaynakları puanına bakılmaktadır.

Ridge Regresyon Yöntemine İlişkin Sonuçlar

Bağımsız değişkenler arasında çoklu doğrusal bağlantı olduğunda bu sorunu Ridge Regresyon yöntemiyle ortadan kaldıran ve son yıllarda oldukça popüler olan Ridge temelli Lojistik Regresyon yöntemiyle elde edilen güvenirlik değerleri Tablo 34'te gösterilmiştir (Cessie ve Houwelingen, 1992).

Tablo 34

Ridge Regresyon Yöntemiyle Elde Edilen Güvenirlik Değerleri

Yöntem	Sonuç
Doğru olarak sınıflanan örnek sayısı	4141
Doğru sınıflama oranı	70,61
Kappa istatistiği	0,409
Mutlak hatanın ortalaması	0,384
Hataların ortalama karekökü	0,438
Göreceli mutlak hata	%77,11
Göreceli hataların karekökü	%87,82
Toplam Örnek Sayısı	5865

Tablo 34 incelendiğinde PISA Türkiye örnekleminde 5865 öğrenciden 4141 tanesinin doğru olarak sınıflandırıldığı ve doğru sınıflama oranının %70,61 olduğu görülmektedir. Bunun yanında kappa istatistiğinin 0,41 ve mutlak hatanın ortalamasının 0,38 olduğu belirlenmiştir. Bunların yanında sınıflara ilişkin detaylı geçerlik ölçütleri Tablo 35'te gösterilmiştir.

Tablo 35

Ridge Regresyon Yöntemiyle Elde Edilen Geçerlik Ölçütleri

Sınıf	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
Başarısız	0,739	0,330	0,714	0,739	0,726	0,409	0,777	0,786
Başarılı	0,670	0,261	0,679	0,670	0,683	0,409	0,777	0,742
Genel	0,706	0,298	0,706	0,706	0,706	0,409	0,777	0,765

Tablo 35'te görüldüğü üzere başarılı ve başarısız olarak sınıflandırma sonuçları genel olarak incelendiğinde doğru pozitif oranı 0,706, yanlış pozitif oranı 0,298, duyarlık 0,706, geri çağırma 0,706, F-ölçütü 0,706, Matthew korelasyon katsayısı 0,409, Roc eğrisi altında kalan alan 0,777 ve duyarlık geri çağırma eğrisi altında kalan alanın 0,765 olduğu belirlenmiştir. Buna göre en yüksek

değerlendirme ölçütünün ROC alanı ve en düşük değerlendirme ölçütünün YP oranı olduğu belirlenmiştir.

Farklı Yöntemlerin Karşılaştırılması

Farklı yöntemler yardımıyla elde edilen güvenirlik değerlerinin karşılaştırılması amacıyla Ağırlıklandırılmış genel ortalama değerleri karşılaştırılmıştır. Her bir yöntem için doğru sınıflanan örnek sayıları, doğru sınıflama oranları, Kappa istatistiği, mutlak hatanın ortalaması, hataların ortalama karekökü ve göreceli mutlak hata değerleri Tablo 36'da gösterilmiştir.

Tablo 36

Farklı Yöntemler Kullanılarak Elde Edilen Güvenirlik Ölçütleri

Algoritma	Doğru Sınıflama sayısı	Doğru Sınıflama oranı	Kappa istatistiği	Ortalama mutlak hata	Karekök hata	Göreli mutlak hata	Göreli karekök hata
1.Decision Stump	3744	63,83	0,276	0,460	0,479	92,32	96,10
2.Hoefding Tree	4085	69,65	0,390	**0,357**	0,455	**71,76**	91,18
3. J.48	3976	67,79	0,351	0,382	0,492	76,79	98,67
4.Lojistik Model	4125	70,33	0,403	0,381	0,437	76,58	87,70
5. RepTree	3946	67,28	0,341	0,391	0,472	78,56	94,65
6. Random Forest	**4177**	**71,22**	**0,420**	0,378	**0,434**	76,01	**86,92**
7. Random Tree	3690	62,92	0,256	0,370	0,609	74,40	121,98
8. Ridge Lojistik Reg.	4141	70,61	0,409	0,384	0,438	77,11	87,82

Tablo 36 incelendiğinde doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği, karekök hata ve göreceli karekök hata değerleri bakımından en iyi sonuçların Random Forest yöntemiyle elde edilirken mutlak hatanın ortalaması ve göreceli mutlak hata bakımından en iyi sonuçların Hoefding Tree yöntemiyle elde edildiği belirlenmiştir. Bunun yanında doğru sınıflanan örnek sayısı ve oranı, Kappa istatistiği, hataların ortalama karekökü ve göreceli hataların karekökü bakımından en kötü sonuçların Random tree tarafından elde edilirken mutlak hatanın ortalaması ve göreceli mutlak hata bakımından en kötü sonuçların Decision Stump yöntemiyle elde edildiği belirlenmiştir.

Çalışmada Decision stump, Hoefding tree, J.48, Lojistik model, RepTree, Random Forest, Random Tree ve Ridge Lojistik Regresyon yöntemleri kullanılarak

başarılı ve başarısız olarak sınıflama sonuçları için elde edilen ağırlıklandırılmış ortalamalara dayalı geçerlik ölçütleri Tablo 37'de gösterilmiştir.

Tablo 37

Farklı Algoritmalar Kullanılarak Elde Edilen Geçerlik Ölçütleri

Algoritma	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
1.Decision Stump	0,638	0,361	0,640	0,638	0,639	0,277	0,629	0,596
2.Hoefding Tree	0,697	0,307	0,696	0,697	0,696	0,390	0,758	0,739
3. J.48	0,678	0,328	0,678	0,678	0,677	0,352	0,694	0,649
4.Lojistik Model	0,703	0,301	0,703	0,703	0,703	0,404	0,778	0,767
5. RepTree	0,673	0,332	0,672	0,673	0,672	0,342	0,718	0,688
6. Random Forest	**0,712**	**0,294**	**0,712**	**0,712**	**0,711**	**0,421**	**0,785**	**0,777**
7. Random Tree	0,629	0,373	0,629	0,629	0,629	0,256	0,628	0,582
8. Ridge Lojistik Reg.	0,706	0,298	0,706	0,706	0,706	0,409	0,777	0,765

Tablo 37 incelendiğinde Random Forest yöntemiyle elde edilen tüm geçerlik ölçütlerinin diğer yöntemlerden daha yüksek olduğu belirlenmiştir. Güvenirlik ve geçerliğe ilişkin elde edilen sonuçlar bir bütün olarak değerlendirildiğinde eğer kategorik bir bağımlı değişken için bireyleri sınıflama yapmak istenirse en iyi sonuçların sırasıyla Random forest, Ridge lojistik regresyon, Lojistik model, Hoefding tree, J.48, RepTree, Decision Stump ve Random Tree algoritmaları şeklinde olacağı görülmektedir. Eğer Kappa istatistiği dikkate alınarak bir sıralama yapmak istenirse en iyi sonuçların sırasıyla Random forest, Ridge lojistik regresyon, Lojistik model, Hoefding tree, J.48, RepTree, Decision Stump ve Random Tree algoritmaları şeklinde olacağı görülmektedir. Buna göre doğru sınıflama veya kappa istatistiğini esas almanın sıralamada bir değişikliğe neden olmayacağı görülmektedir.

İkinci Alt Probleme İlişkin Bulgular

PISA öğrenci anketiyle ölçülen değişkenlerden yararlanarak başarıyı tahmin etmede kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree,

Random Forest, Random Tree ve Ridge lojistik Regresyon yöntemleri yardımıyla edilen ölçme sonuçlarının test edilen verilerin %5, %10, %15, %20, %25, %30 ve %35 olması durumunda güvenirlik ve geçerlik sonuçları ne düzeydedir?

Çalışmanın ikinci alt probleminde çapraz geçerleme yöntemi kullanılmadan doğrudan veri setinin eğitim ve test verisi olarak belli oranlarda ayrılması yöntemiyle analizler gerçekleştirilmiştir.

Decision Stump Yöntemine İlişkin Sonuçlar

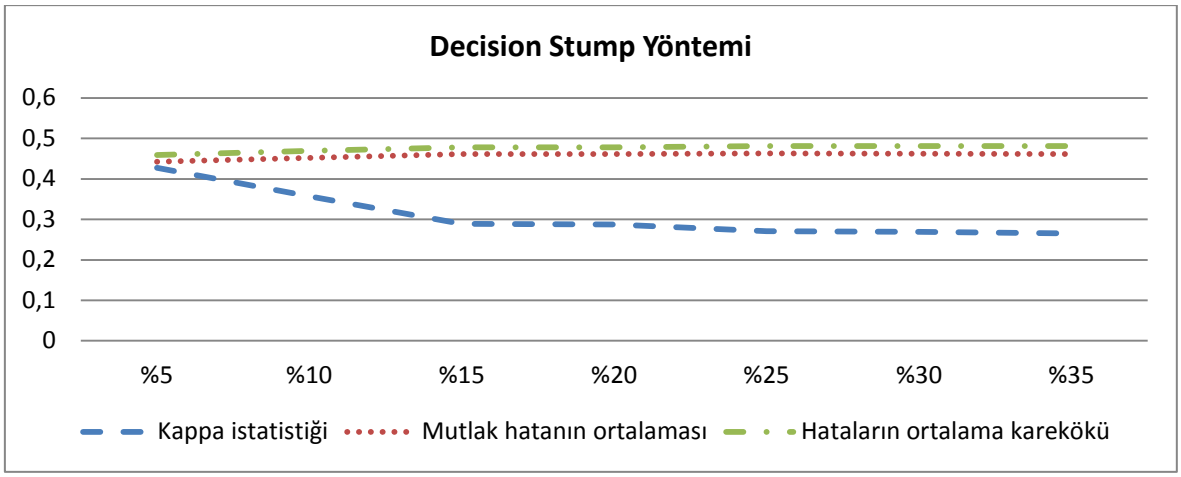
Decision Stump yöntemiyle test edilen verilerin farklı yüzdeliklerde elde edilen sonuçlarına ilişkin güvenirlik değerleri belirlenmiş ve sonuçlar Tablo 38'de gösterilmiştir.

Tablo 38

Decision Stump Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri

Yöntem	%5	%10	%15	%20	%25	%30	%35
Toplam örnek sayısı	293	586	880	1173	1466	1759	2053
Doğru sınıflanan örnek sayısı	209	398	567	754	931	1116	1298
Doğru sınıflama oranı (%)	71,31	67,91	64,43	64,27	63,50	63,44	63,22
Kappa istatistiği	0,427	0,357	0,289	0,287	0,271	0,269	0,265
Mutlak hatanın ortalaması	0,442	0,452	0,461	0,461	0,463	0,462	0,461
Hataların ortalama karekökü	0,459	0,469	0,478	0,478	0,481	0,481	0,481
Göreceli mutlak hata	88,51	90,41	92,25	92,36	92,68	92,59	92,62
Göreceli hataların karekökü	91,69	93,68	95,55	95,62	96,05	96,23	96,38

Tablo 38 incelendiğinde test edilen veri setinin oranı arttıkça doğru sınıflama oranının azaldığı, diğer istatistiklerin ise monoton bir şekilde arttığı belirlenmiştir. Buna göre test edilen veri oranı attıkça kappa istatistiği haricindeki diğer istatistiklerin arttığı görülmektedir. Bu sonuca göre Decision Stump yönteminin eğitilen veya test edilen örneklem oranından etkilendiği ve test edilen veri setinin büyüklüğü arttıkça hata değerlerinin de arttığı belirlenmiştir. Görsel açıdan daha iyi anlaşılması amacıyla kappa istatistiği, mutlak hatanın ortalaması ve hataların ortalama karekökünün farklı yüzdeliklerdeki test verilerinde nasıl bir değişkenlik gösterdiği Şekil 20'de verilmiştir.



Şekil 20. Decision stump yöntemiyle elde edilen hata ölçütlerinin değişimi.

Bu işlemin ardından Decision Stump yöntemiyle elde edilen Doğru pozitif, Yanlış pozitif, Duyarlılık, Geri çağırma, F-Ölçütü, Matthews korelasyon katsayısı (MCC), ROC eğrisi altında kalan alan (ROC Area) ve Duyarlık-Hassalık Alanı (PRC Area) ölçütlerine ilişkin sonuçlar Tablo 39'da gösterilmiştir.

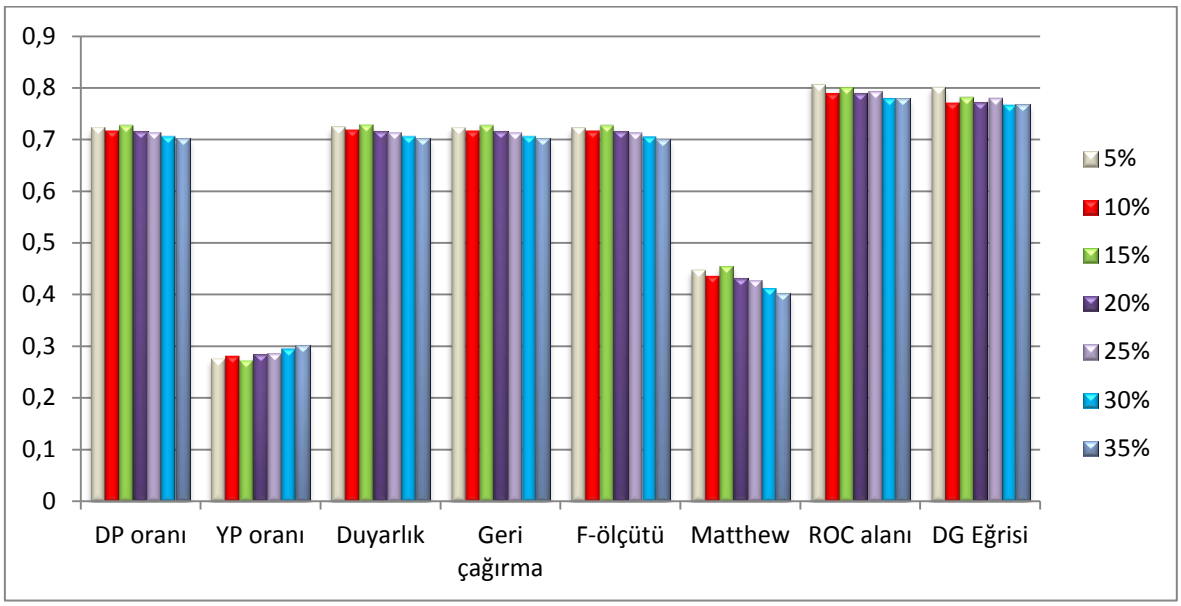
Tablo 39

Decision Stump Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri

Oran	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
%5	0,713	0,286	0,715	0,713	0,713	0,428	0,714	0,653
%10	0,679	0,322	0,679	0,679	0,679	0,358	0,679	0,621
%15	0,644	0,354	0,649	0,644	0,642	0,294	0,645	0,594
%20	0,643	0,355	0,648	0,643	0,640	0,292	0,644	0,593
%25	0,635	0,363	0,640	0,635	0,633	0,276	0,636	0,587
%30	0,634	0,364	0,636	0,634	0,634	0,271	0,635	0,586
%35	0,632	0,366	0,634	0,632	0,632	0,266	0,633	0,585

Tablo 39'da görüldüğü üzere test edilen veri setinin oranı arttıkça doğru pozitif oranının monoton olarak azalırken yanlış pozitif oranının monoton bir şekilde arttığı görülmektedir. Bunun yanında test edilen veri setinin oranı arttıkça duyarlığın ve geri çağırmanın azaldığı görülmektedir. Matthew, ROC alanı ve DG eğrisinin test edilen veri seti arttıkça monoton bir şekilde azaldığı ve benzer

şekilde F-ölçütünün de azalma eğiliminde olduğu belirlenmiştir. Elde edilen sonuçlar Şekil 21'de görsel olarak sunulmaktadır.



Şekil 21. Decision stump yöntemiyle farklı yüzdeliklerdeki geçerlik ölçütleri.

Hoeffding Tree Yöntemine İlişkin Sonuçlar

Hoeffding tree yöntemiyle test edilen verilerin farklı yüzdeliklerde elde edilen sonuçların güvenirlik değerleri elde edilmiş ve sonuçlar Tablo 40'da gösterilmiştir.

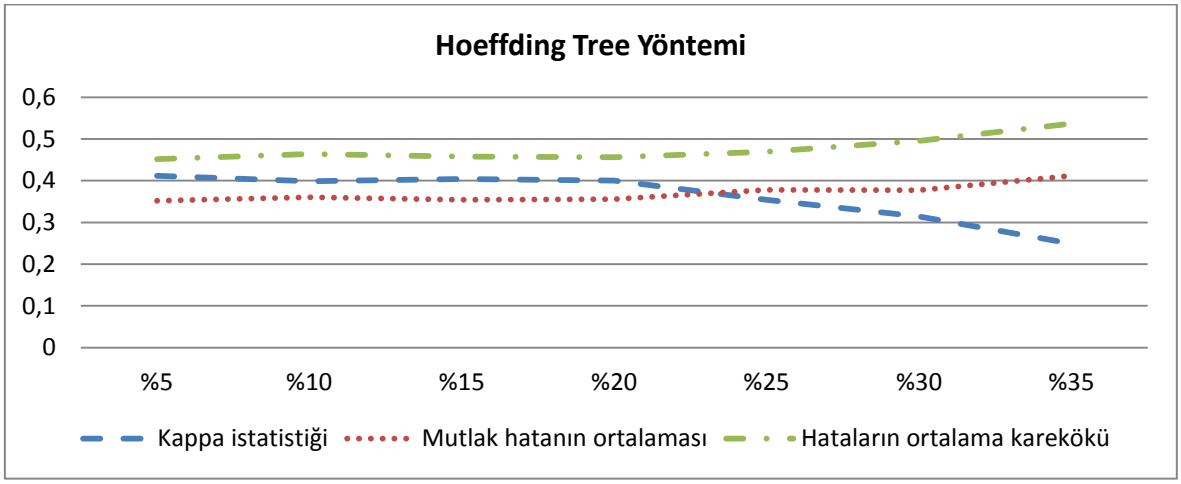
Tablo 40

Hoeffding Tree Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri

Yöntem	%5	%10	%15	%20	%25	%30	%35
Toplam örnek sayısı	293	586	880	1173	1466	1759	2053
Doğru sınıflanan örnek sayısı	207	410	618	821	994	1161	1297
Doğru sınıflama oranı (%)	70,64	69,96	70,22	69,99	67,80	66,00	63,17
Kappa istatistiği	0,412	0,399	0,404	0,400	0,354	0,314	0,249
Mutlak hatanın ortalaması	0,352	0,360	0,354	0,356	0,378	0,377	0,412
Hataların ortalama karekökü	0,452	0,464	0,458	0,456	0,469	0,495	0,537
Göreceli mutlak hata	70,48	72,10	70,86	71,36	75,74	75,57	82,60
Göreceli hataların karekökü	90,44	92,54	91,48	91,25	93,76	99,09	107,63

Tablo 40 incelendiğinde test edilen veri setinin oranı arttıkça doğru sınıflama oranının test edilen verinin %15 olması durumunda arttığı ancak daha sonra azaldığı, kappa istatistiğinin ise test edilen veri setinin oranı %15 olduğunda bir miktar arttığı ancak daha sonra tekrardan monoton bir şekilde azaldığı belirlenmiştir. Buna göre test edilen veri oranı attıkça kappa istatistiği haricindeki

diğer istatistiklerin arttığı görülmektedir. Bu sonuca göre Hoeffding tree yönteminin eğitilen veya test edilen örneklem oranından etkilendiği ve test edilen veri setinin büyüklüğü arttıkça hata değerlerinin de arttığı belirlenmiştir. Görsel açıdan daha iyi anlaşılması amacıyla kappa istatistiği, mutlak hatanın ortalaması ve hataların ortalama karekökünün farklı yüzdeliklerdeki test verilerinde nasıl bir değişkenlik gösterdiği Şekil 22'de verilmiştir.



Şekil 22. Hoeffding tree yöntemiyle elde edilen hata ölçütlerinin değişimi.

Hoeffding tree yöntemiyle elde edilen Doğru pozitif, Yanlış pozitif, Duyarlılık, Geri çağırma, F-Ölçütü, Matthews korelasyon katsayısı (MCC), ROC eğrisi altında kalan alan (ROC Area) ve Duyarlık-Hassalık Alanı (PRC Area) ölçütlerine ilişkin sonuçlar Tablo 41'de gösterilmiştir.

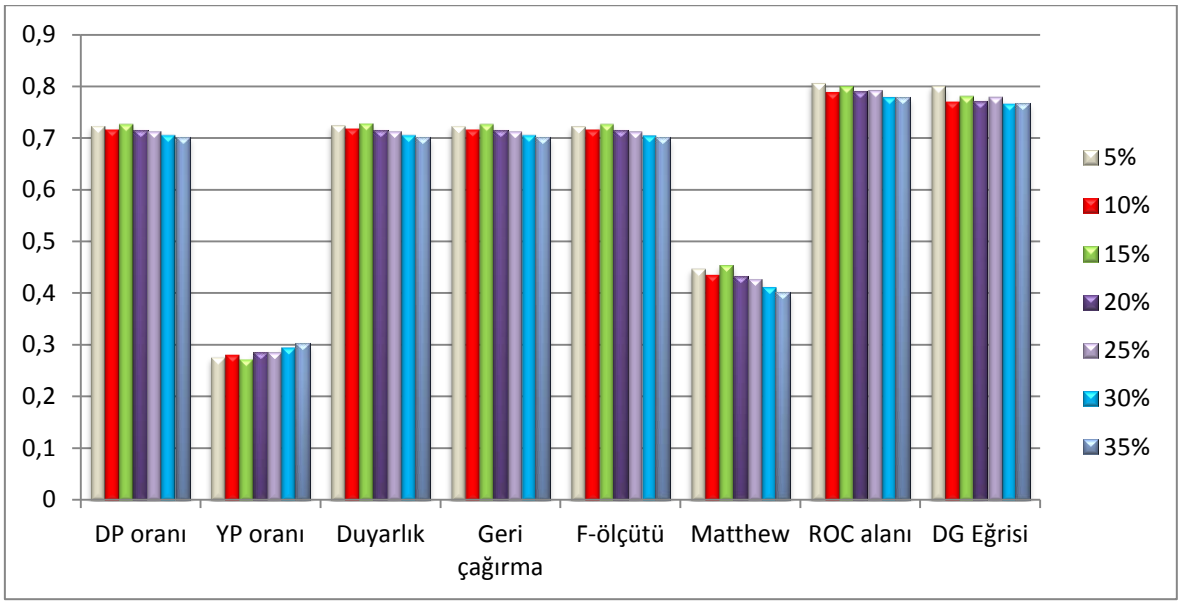
Tablo 41

Hoeffding Tree Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri

Oran	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
%5	0,706	0,294	0,706	0,706	0,706	0,413	0,767	0,758
%10	0,700	0,300	0,700	0,700	0,700	0,399	0,748	0,725
%15	0,702	0,297	0,703	0,702	0,702	0,406	0,758	0,727
%20	0,700	0,299	0,701	0,700	0,700	0,401	0,760	0,739
%25	0,678	0,325	0,688	0,678	0,673	0,365	0,719	0,695
%30	0,660	0,348	0,670	0,660	0,653	0,327	0,694	0,666
%35	0,632	0,387	0,657	0,632	0,609	0,279	0,623	0,620

Tablo 41'de görüldüğü üzere test edilen veri setinin oranı arttıkça doğru pozitif oranının monoton olarak azalırken yanlış pozitif oranının monoton bir

şekilde arttığı görülmektedir. Bunun yanında test edilen veri setinin oranı arttıkça duyarlığın ve geri çağırmanın azaldığı görülmektedir. Matthew, ROC alanı ve DG eğrisinin test edilen veri seti arttıkça monoton bir şekilde azaldığı ve benzer şekilde F-ölçütünün de azalma eğiliminde olduğu belirlenmiştir. Elde edilen sonuçlar Şekil 23'te gösterilmiştir.



Şekil 23. Hoeffding tree yöntemiyle farklı yüzdeliklerdeki geçerlik ölçütlerinin değişimi.

J.48 Yöntemine İlişkin Sonuçlar

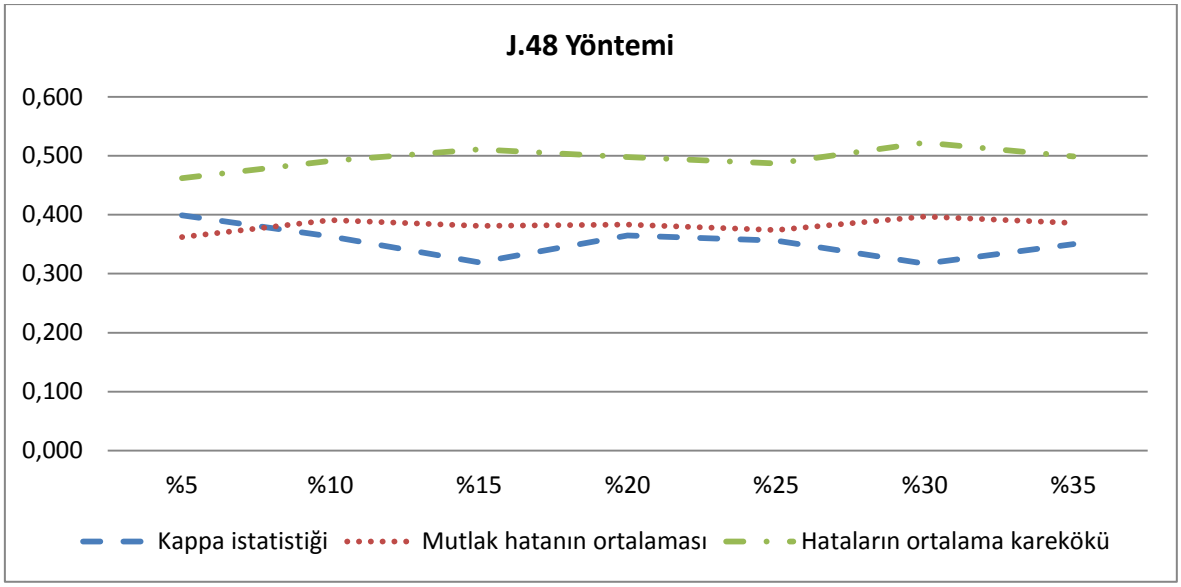
J.48 yöntemiyle test edilen verilerin farklı yüzdeliklerde elde edilen sonuçların güvenirlik değerleri elde edilmiş ve sonuçlar Tablo 28'de gösterilmiştir.

Tablo 42

J.48 Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri

Yöntem	%5	%10	%15	%20	%25	%30	%35
Toplam örnek sayısı	293	586	880	1173	1466	1759	2053
Doğru sınıflanan örnek sayısı	205	399	581	801	994	1159	1389
Doğru sınıflama oranı (%)	69,96	68,08	66,02	68,28	67,80	65,88	67,65
Kappa istatistiği	0,399	0,363	0,319	0,365	0,356	0,317	0,350
Mutlak hatanın ortalaması	0,362	0,391	0,381	0,383	0,374	0,397	0,386
Hataların ortalama karekökü	0,462	0,491	0,511	0,498	0,487	0,522	0,499
Göreceli mutlak hata	72,60	78,22	76,25	76,76	74,95	79,56	77,48
Göreceli hataların karekökü	92,40	97,91	102,17	99,56	97,29	104,39	99,84

Tablo 42 incelendiğinde test edilen veri setinin oranı arttıkça doğru sınıflama oranının önce azaldığı sonra bir miktar artıp %20'de en yüksek değeri ulaştığı ve sonra tekrar azaldığı, kappa istatistiğinin ise önce azaldığı sonra bir miktar arttığı ancak daha sonra tekrar azaldığı belirlenmiştir. Başka bir ifadeyle kappa istatistiğinin tutarlı sonuçlar üretmediği görülmektedir. Benzer şekilde test edilen veri oranı attıkça diğer istatistiklerinde hem arttığı hem de azaldığı görülmektedir. Görsel açıdan daha iyi anlaşılması amacıyla kappa istatistiği, mutlak hatanın ortalaması ve hataların ortalama karekökünün farklı yüzdeliklerdeki test verilerinde nasıl bir değişkenlik gösterdiği Şekil 24'te verilmiştir.



Şekil 24. J.48 yöntemiyle elde edilen hata ölçütlerinin değişimi.

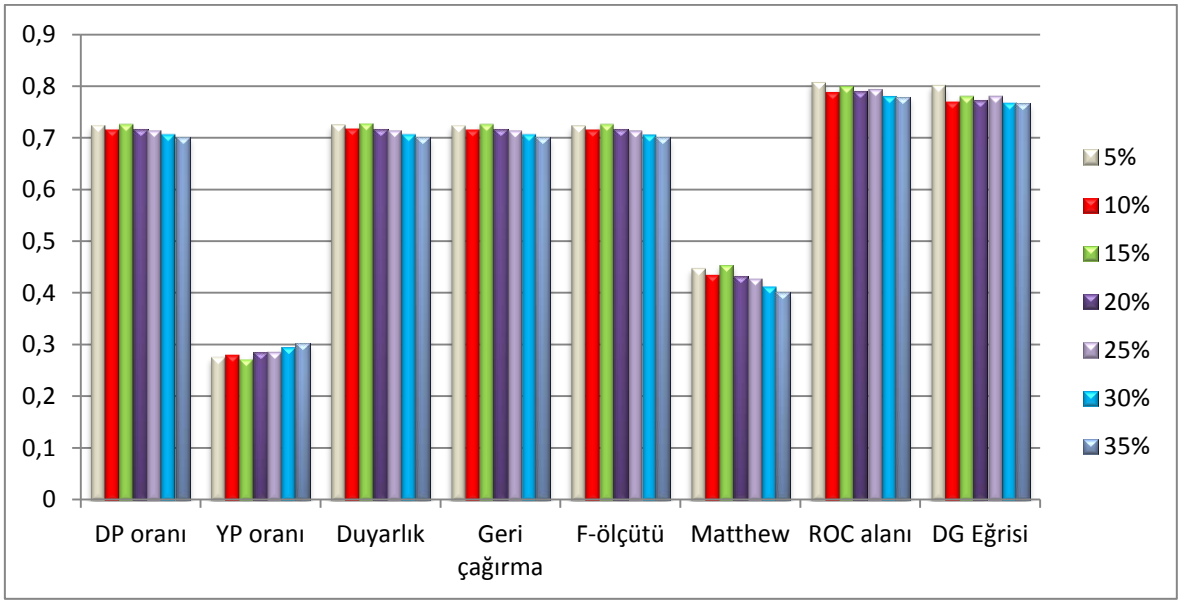
J.48 yöntemiyle elde edilen Doğru pozitif, Yanlış pozitif, Duyarlılık, Geri çağırma, F-Ölçütü, Matthews korelasyon katsayısı (MCC), ROC eğrisi altında kalan alan (ROC Area) ve Duyarlık-Hassalık Alanı (PRC Area) ölçütlerine ilişkin sonuçlar Tablo 43'de gösterilmiştir.

Tablo 43

J.48 Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri

Oran	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
%5	0,700	0,300	0,700	0,700	0,700	0,400	0,749	0,713
%10	0,681	0,317	0,685	0,681	0,680	0,367	0,693	0,651
%15	0,660	0,341	0,661	0,660	0,659	0,321	0,680	0,636
%20	0,683	0,318	0,683	0,683	0,683	0,366	0,688	0,642
%25	0,678	0,322	0,678	0,678	0,678	0,356	0,708	0,664
%30	0,659	0,342	0,659	0,659	0,659	0,317	0,654	0,609
%35	0,677	0,327	0,677	0,677	0,676	0,351	0,682	0,641

Tablo 43'de görüldüğü üzere test edilen veri setinin oranı arttıkça doğru pozitif ve yanlış pozitif oranlarının hem arttığı hem de azaldığı görülmektedir. Bunun yanında test edilen veri setinin oranı arttıkça duyarlığın ve geri çağırmanın da hem azalıp hem de arttığı belirlenmiştir. Genel olarak bakıldığında test edilen veri setinin %20 olması durumunda değerlendirme ölçütlerinin değişkenlik yönünde farklılaşma olmakta ve ölçütlerin tutarlı sonuçlar üretmediği görülmektedir. Elde edilen sonuçlar Şekil 25'te görselleştirilmiştir.



Şekil 25. J.48 yöntemiyle farklı yüzdeliklerdeki geçerlik ölçütlerinin değişimi.

Lojistik Model Yöntemine İlişkin Sonuçlar

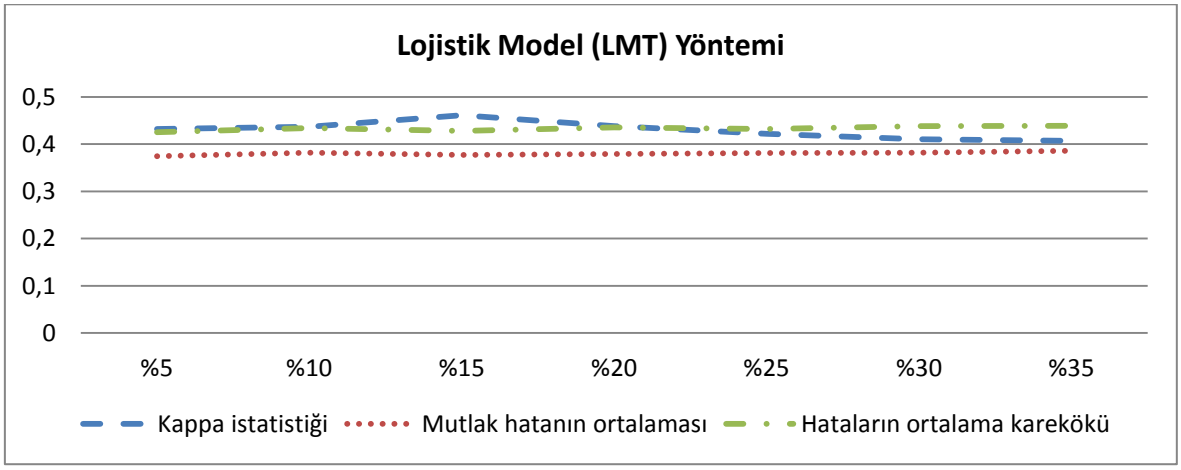
Lojistik model yöntemiyle test edilen verilerin farklı yüzdeliklerde elde edilen sonuçların güvenirlik değerleri elde edilmiş ve sonuçlar Tablo 44'de gösterilmiştir.

Tablo 44

Lojistik Model Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri

Yöntem	%5	%10	%15	%20	%25	%30	%35
Toplam örnek sayısı	293	586	880	1173	1466	1759	2053
Doğru sınıflanan örnek sayısı	210	421	643	844	1043	1242	1448
Doğru sınıflama oranı (%)	71,67	71,84	73,06	71,95	71,14	70,60	70,53
Kappa istatistiği	0,432	0,437	0,461	0,438	0,422	0,410	0,407
Mutlak hatanın ortalaması	0,374	0,382	0,377	0,379	0,381	0,382	0,386
Hataların ortalama karekökü	0,425	0,434	0,428	0,435	0,432	0,438	0,439
Göreceli mutlak hata	74,98	76,42	75,49	75,96	76,34	76,53	77,43
Göreceli hataların karekökü	84,95	86,63	85,63	86,99	86,39	87,59	87,91

Tablo 44 incelendiğinde test edilen veri setinin oranı arttıkça doğru sınıflama oranının en fazla %15'lik test veri setinde olduğu, kappa istatistiğinin ise %15'lik test veri setine kadar artarken daha sonra azaldığı belirlenmiştir. Bunun yanında mutlak hatanın ortalaması monoton olarak artarken diğer hata ölçütlerinin de hem azalıp hem de arttığı görülmektedir. Görsel açıdan daha iyi anlaşılması amacıyla kappa istatistiği, mutlak hatanın ortalaması ve hataların ortalama karekökünün farklı yüzdeliklerdeki test verilerinde nasıl bir değişkenlik gösterdiği Şekil 26'da verilmiştir.



Şekil 26. Lojistik model (lmt) yöntemiyle elde edilen hata ölçütlerinin değişimi.

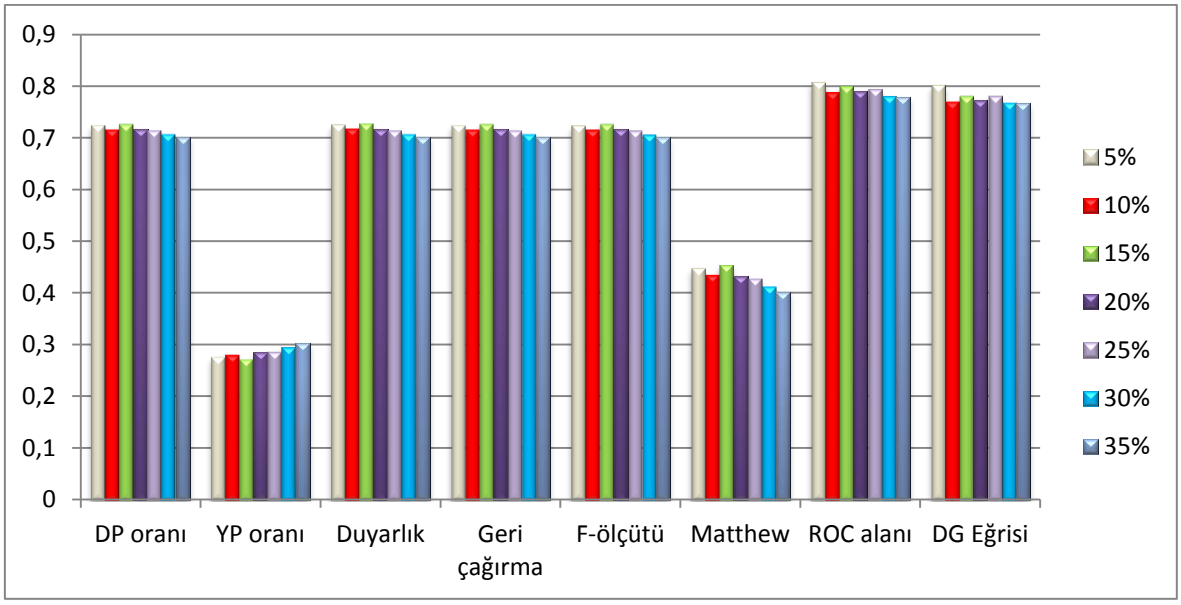
LMT yöntemiyle elde edilen Doğru pozitif, Yanlış pozitif, Duyarlılık, Geri çağırma, F-Ölçütü, Matthews korelasyon katsayısı (MCC), ROC eğrisi altında kalan alan (ROC Area) ve Duyarlık-Hassalık Alanı (PRC Area) ölçütlerine ilişkin sonuçlar Tablo 45'te gösterilmiştir.

Tablo 45

LMT Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri

Oran	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
%5	0,717	0,284	0,718	0,717	0,716	0,435	0,806	0,800
%10	0,718	0,280	0,720	0,718	0,718	0,439	0,789	0,771
%15	0,731	0,270	0,731	0,731	0,730	0,462	0,799	0,781
%20	0,720	0,281	0,720	0,720	0,719	0,439	0,785	0,769
%25	0,711	0,290	0,713	0,711	0,711	0,424	0,792	0,779
%30	0,706	0,297	0,707	0,706	0,705	0,412	0,779	0,770
%35	0,705	0,299	0,706	0,705	0,704	0,409	0,776	0,765

Tablo 45'te görüldüğü üzere test edilen veri setinin oranı arttıkça doğru pozitif ve yanlış pozitif oranlarının hem arttığı hem de azaldığı görülmektedir. Bunun yanında test edilen veri setinin oranı arttıkça duyarlığın ve geri çağırmanın da hem azalıp hem de arttığı belirlenmiştir. Genel olarak bakıldığında test edilen veri setinin %15 olması durumunda değerlendirme ölçütlerinin değişkenlik yönünde farklılaşma olmakta ve ölçütlerin tutarlı sonuçlar üretmediği görülmektedir. Elde edilen sonuçlar Şekil 27'de görselleştirilmiştir.



Şekil 27. Lojistik model yöntemiyle farklı yüzdeliklerdeki geçerlik ölçütleri.

RepTree Yöntemine İlişkin Sonuçlar

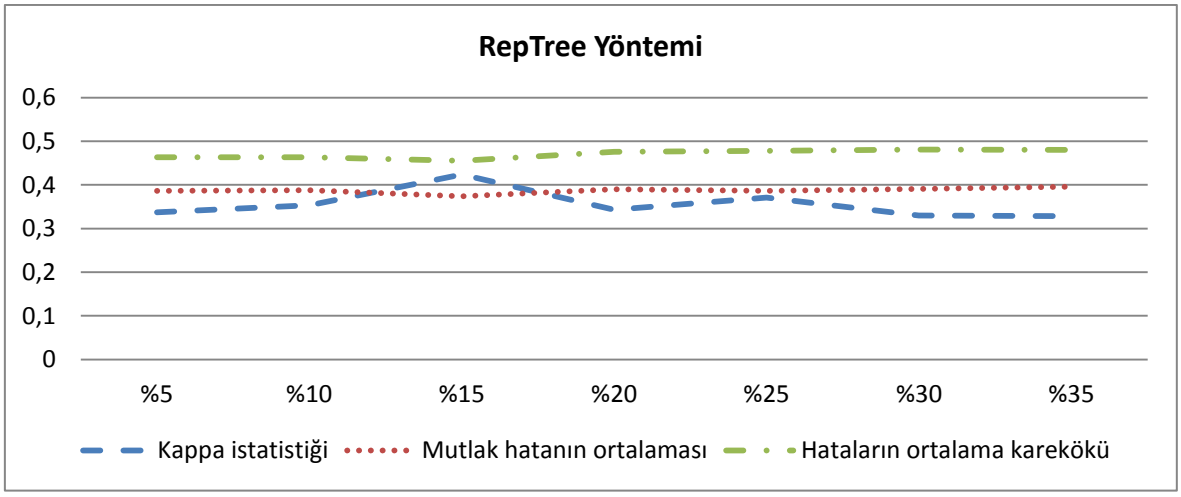
RepTree yöntemiyle test edilen verilerin farklı yüzdeliklerde elde edilen sonuçların güvenirlik değerleri elde edilmiş ve sonuçlar Tablo 46'da gösterilmiştir.

Tablo 46

RepTree Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri

Yöntem	%5	%10	%15	%20	%25	%30	%35
Toplam örnek sayısı	293	586	880	1173	1466	1759	2053
Doğru sınıflanan örnek sayısı	196	396	627	788	1006	1172	1364
Doğru sınıflama oranı (%)	66,89	67,57	71,25	67,17	68,62	66,62	66,43
Kappa istatistiği	0,337	0,353	0,424	0,343	0,371	0,330	0,328
Mutlak hatanın ortalaması	0,386	0,388	0,374	0,390	0,386	0,391	0,396
Hataların ortalama karekökü	0,463	0,463	0,455	0,476	0,478	0,481	0,480
Göreceli mutlak hata	77,27	77,53	74,94	78,10	77,25	78,43	79,58
Göreceli hataların karekökü	92,59	92,30	90,96	95,07	95,45	96,14	96,08

Tablo 46 incelendiğinde test edilen veri setinin oranı arttıkça doğru sınıflama oranının en fazla %15'lik test veri setinde olduğu, kappa istatistiğinin ise %15'lik test veri setine kadar artarken daha sonra azaldığı belirlenmiştir. Bunun yanında mutlak hatanın ortalaması monoton olarak artarken diğer hata ölçütlerinin de hem azalıp hem de arttığı görülmektedir. Görsel açıdan daha iyi anlaşılması amacıyla kappa istatistiği, mutlak hatanın ortalaması ve hataların ortalama karekökünün farklı yüzdeliklerdeki test verilerinde nasıl bir değişkenlik gösterdiği Şekil 28'de verilmiştir.



Şekil 28. RepTree yöntemiyle elde edilen hata ölçütlerinin değişimi.

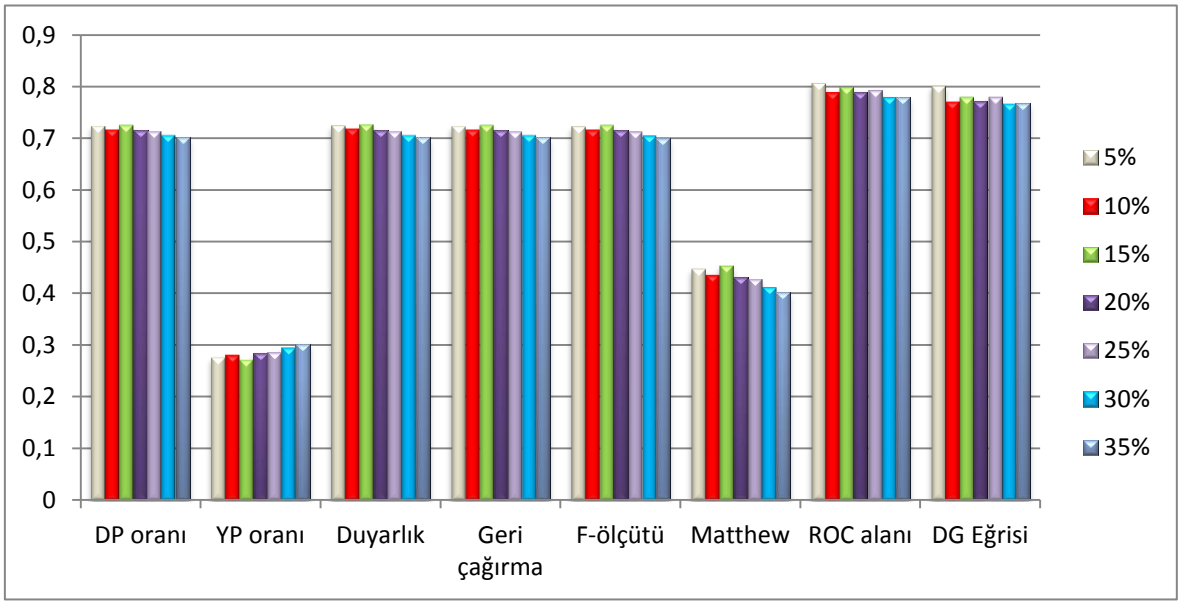
RepTree yöntemiyle elde edilen Doğru pozitif, Yanlış pozitif, Duyarlılık, Geri çağırma, F-Ölçütü, Matthews korelasyon katsayısı (MCC), ROC eğrisi altında kalan alan (ROC Area) ve Duyarlık-Hassalık Alanı (PRC Area) ölçütlerine ilişkin sonuçlar Tablo 47'de gösterilmiştir.

Tablo 47

RepTree Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri

Oran	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
%5	0,669	0,331	0,669	0,669	0,669	0,338	0,746	0,739
%10	0,676	0,321	0,684	0,676	0,673	0,361	0,741	0,707
%15	0,713	0,288	0,714	0,713	0,712	0,426	0,747	0,711
%20	0,672	0,329	0,672	0,672	0,672	0,343	0,711	0,676
%25	0,686	0,315	0,687	0,686	0,686	0,373	0,710	0,678
%30	0,666	0,337	0,667	0,666	0,665	0,332	0,704	0,662
%35	0,664	0,336	0,665	0,664	0,664	0,328	0,701	0,677

Tablo 47'de görüldüğü üzere test edilen veri setinin oranı arttıkça doğru pozitif ve yanlış pozitif oranlarının hem arttığı hem de azaldığı görülmektedir. Bunun yanında test edilen veri setinin oranı arttıkça duyarlığın ve geri çağırmanın da hem azalıp hem de arttığı belirlenmiştir. Genel olarak bakıldığında test edilen veri setinin %15 olması durumunda değerlendirme ölçütlerinin değişkenlik yönünde farklılaşma olmakta ve ölçütlerin tutarlı sonuçlar üretmediği görülmektedir. Elde edilen sonuçlar Şekil 29'da görselleştirilmiştir.



Şekil 29. RepTree yöntemiyle farklı yüzdeliklerdeki geçerlik ölçütlerinin değişimi.

Random Forest Yöntemine İlişkin Sonuçlar

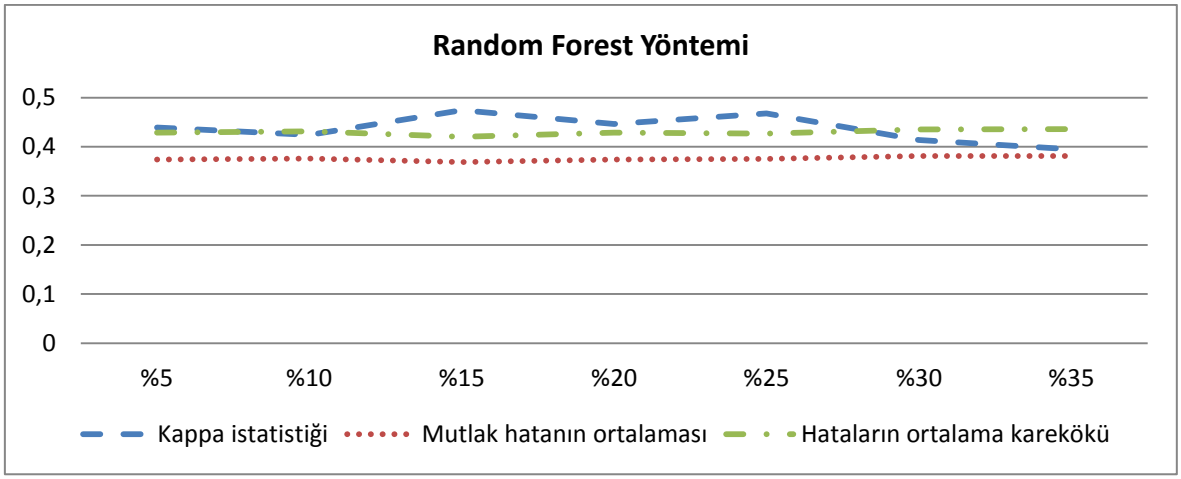
Random Forest yöntemiyle test edilen verilerin farklı yüzdeliklerde elde edilen sonuçların güvenirlik değerleri Tablo 48'de gösterilmiştir.

Tablo 48

Random Forest Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri

Yöntem	%5	%10	%15	%20	%25	%30	%35
Toplam örnek sayısı	293	586	880	1173	1466	1759	2053
Doğru sınıflanan örnek sayısı	211	417	649	849	1077	1246	1437
Doğru sınıflama oranı (%)	72,01	71,16	73,75	72,37	73,46	70,83	69,99
Kappa istatistiği	0,439	0,424	0,474	0,446	0,468	0,414	0,395
Mutlak hatanın ortalaması	0,374	0,376	0,369	0,374	0,375	0,381	0,381
Hataların ortalama karekökü	0,429	0,431	0,420	0,429	0,427	0,435	0,436
Göreceli mutlak hata	74,95	75,26	73,87	74,99	75,20	76,45	76,38
Göreceli hataların karekökü	85,84	85,96	84,03	85,69	85,26	87,10	87,31

Tablo 48 incelendiğinde test edilen veri setinin oranı arttıkça doğru sınıflama oranının en fazla %15'lik test veri setinde olduğu, kappa istatistiğinin ise %25'lik test veri setinden sonra azaldığı belirlenmiştir. Bunun yanında mutlak hatanın ortalaması monoton olarak artarken diğer hata ölçütlerinin de hem azalıp hem de arttığı görülmektedir. Görsel açıdan daha iyi anlaşılması amacıyla kappa istatistiği, mutlak hatanın ortalaması ve hataların ortalama karekökünün farklı yüzdeliklerdeki test verilerinde nasıl bir değişkenlik gösterdiği Şekil 30'da verilmiştir.



Şekil 30. Randon forest yöntemiyle elde edilen hata ölçütlerinin değişimi.

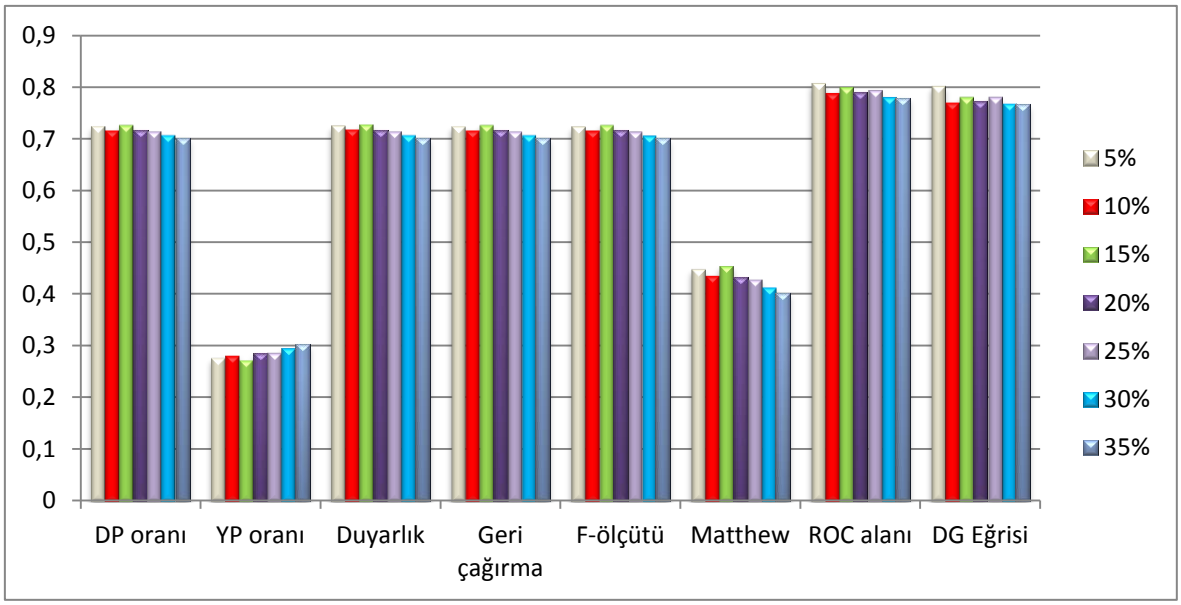
Random Forest yöntemiyle elde edilen Doğru pozitif, Yanlış pozitif, Duyarlılık, Geri çağırma, F-Ölçütü, Matthews korelasyon katsayısı (MCC), ROC eğrisi altında kalan alan (ROC Area) ve Duyarlık-Hassalık Alanı (PRC Area) ölçütlerine ilişkin sonuçlar Tablo 49'da gösterilmiştir.

Tablo 49

Random Forest Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri

Oran	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
%5	0,720	0,282	0,727	0,720	0,718	0,446	0,799	0,799
%10	0,712	0,287	0,715	0,712	0,711	0,427	0,795	0,785
%15	0,738	0,263	0,739	0,738	0,737	0,476	0,814	0,805
%20	0,724	0,278	0,726	0,724	0,723	0,449	0,797	0,788
%25	0,735	0,267	0,739	0,735	0,733	0,473	0,802	0,794
%30	0,708	0,296	0,711	0,708	0,706	0,418	0,783	0,776
%35	0,700	0,306	0,702	0,700	0,698	0,399	0,781	0,774

Tablo 49'da görüldüğü üzere test edilen veri setinin oranı arttıkça doğru pozitif ve yanlış pozitif oranlarının hem arttığı hem de azaldığı görülmektedir. Bunun yanında test edilen veri setinin oranı arttıkça duyarlığın ve geri çağırmanın da hem azalıp hem de arttığı belirlenmiştir. Genel olarak bakıldığında test edilen veri setinin %15 olması durumunda değerlendirme ölçütlerinin değişkenlik yönünde farklılaşma olmakta ve ölçütlerin tutarlı sonuçlar üretmediği görülmektedir. Elde edilen sonuçlar Şekil 31'de görselleştirilmiştir.



Şekil 31. Random forest yöntemiyle farklı yüzdeliklerdeki geçerlik ölçütlerinin değişimi.

Random Tree Yöntemine İlişkin Sonuçlar

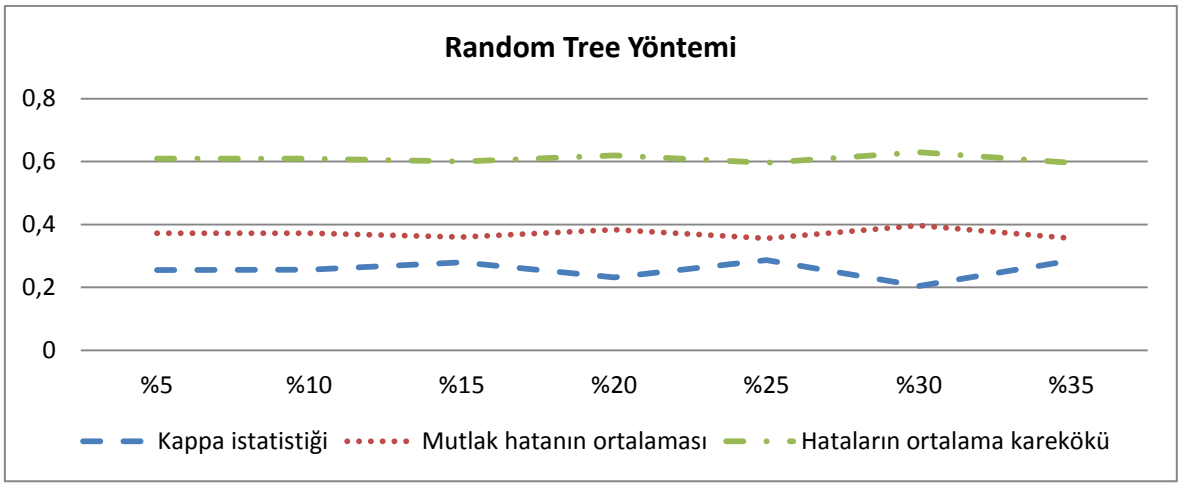
Random Tree yöntemiyle test edilen verilerin farklı yüzdeliklerde elde edilen sonuçların güvenirlik değerleri Tablo 50'de gösterilmiştir.

Tablo 50

Random Tree Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri

Yöntem	%5	%10	%15	%20	%25	%30	%35
Toplam örnek sayısı	293	586	880	1173	1466	1759	2053
Doğru sınıflanan örnek sayısı	184	368	563	723	944	1060	1322
Doğru sınıflama oranı (%)	62,79	62,79	63,97	61,63	64,39	60,26	64,39
Kappa istatistiği	0,255	0,256	0,279	0,232	0,287	0,204	0,286
Mutlak hatanın ortalaması	0,372	0,372	0,360	0,383	0,356	0,397	0,356
Hataların ortalama karekökü	0,609	0,609	0,600	0,619	0,596	0,630	0,596
Göreceli mutlak hata	74,44	74,32	72,08	76,79	71,27	79,62	71,39
Göreceli hataların karekökü	121,85	121,60	119,86	123,70	119,15	126,01	119,39

Tablo 50 incelendiğinde test edilen veri setinin oranı arttıkça doğru sınıflama oranının hem arttığı hem de azaldığı, kappa istatistiğinin ise %25'lik test veri setinden sonra azaldığı belirlenmiştir. Bunun yanında mutlak hatanın ortalaması monoton olarak artarken %35'lik test verisinde azaldığı diğer hata ölçütlerinin de hem azalıp hem de arttığı görülmektedir. Görsel açıdan daha iyi anlaşılması amacıyla kappa istatistiği, mutlak hatanın ortalaması ve hataların ortalama karekökünün farklı yüzdeliklerdeki test verilerinde nasıl bir değişkenlik gösterdiği Şekil 32'de verilmiştir.



Şekil 32. Randon tree yöntemiyle elde edilen hata ölçütlerinin değişimi.

Random Tree yöntemiyle elde edilen Doğru pozitif, Yanlış pozitif, Duyarlılık, Geri çağırma, F-Ölçütü, Matthews korelasyon katsayısı (MCC), ROC eğrisi altında

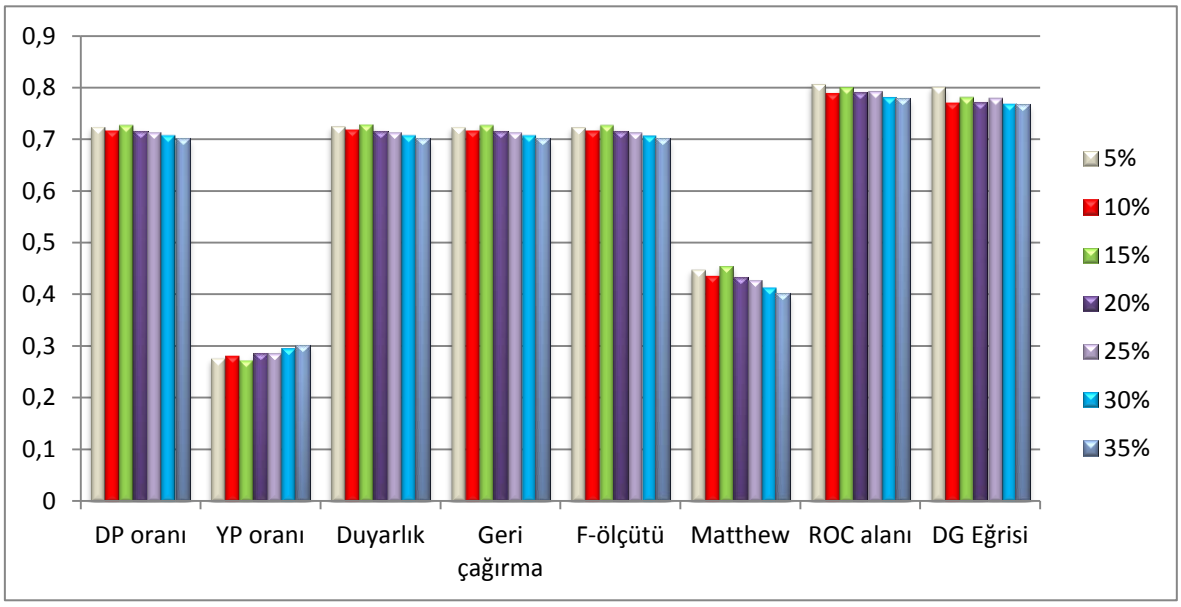
kalan alan (ROC Area) ve Duyarlık-Geri çağırma alanı (DG Eğrisi) ölçütlerine ilişkin sonuçlar Tablo 51'de gösterilmiştir.

Tablo 51

Random Tree Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri

Oran	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
%5	0,628	0,373	0,628	0,628	0,628	0,256	0,628	0,580
%10	0,628	0,372	0,628	0,628	0,628	0,256	0,628	0,581
%15	0,640	0,360	0,640	0,640	0,640	0,280	0,640	0,589
%20	0,616	0,383	0,617	0,616	0,616	0,233	0,617	0,572
%25	0,644	0,357	0,644	0,644	0,644	0,288	0,644	0,593
%30	0,603	0,399	0,602	0,603	0,602	0,204	0,602	0,562
%35	0,644	0,358	0,644	0,644	0,644	0,286	0,643	0,593

Tablo 51'de görüldüğü üzere test edilen veri setinin oranı arttıkça doğru pozitif ve yanlış pozitif oranlarının hem arttığı hem de azaldığı görülmektedir. Bunun yanında test edilen veri setinin oranı arttıkça duyarlığın ve geri çağırmanın da hem azalıp hem de arttığı belirlenmiştir. Genel olarak bakıldığında test edilen veri setinin %15 ve %25 olması durumunda değerlendirme ölçütlerinin değişkenlik yönünde farklılaşma olmakta ve ölçütlerin tutarlı sonuçlar üretmediği görülmektedir. Elde edilen sonuçlar Şekil 33'te görselleştirilmiştir.



Şekil 33. Random tree yöntemiyle farklı yüzdeliklerdeki geçerlik ölçütlerinin değişimi.

Ridge Regresyon Yöntemine İlişkin Sonuçlar

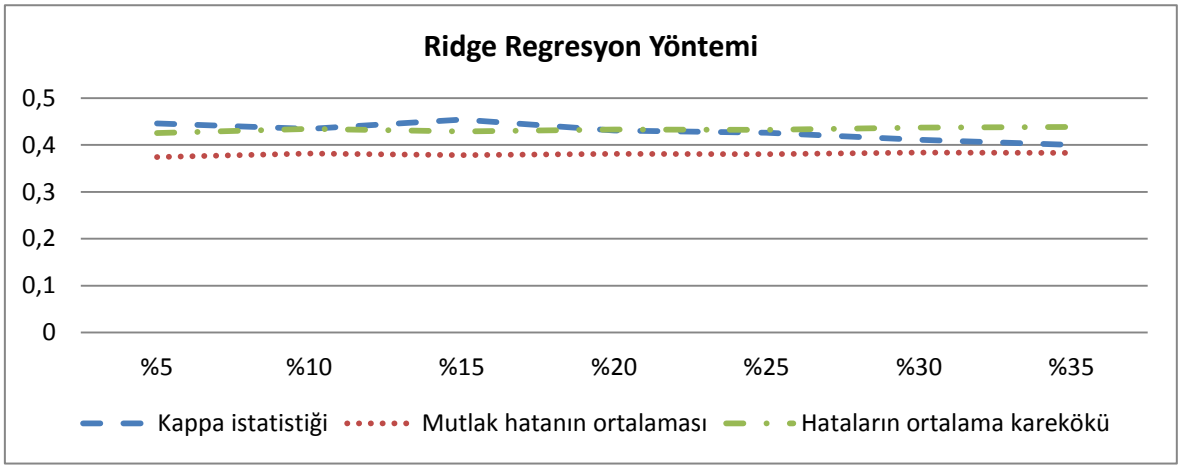
Ridge lojistik regresyon yöntemiyle test edilen verilerin farklı yüzdeliklerde elde edilen sonuçların güvenirlik değerleri Tablo 52'de gösterilmiştir.

Tablo 52

Ridge Regresyon Yöntemiyle Farklı Yüzdeliklerde Elde Edilen Güvenirlik Değerleri

Yöntem	%5	%10	%15	%20	%25	%30	%35
Toplam örnek sayısı	293	586	880	1173	1466	1759	2053
Doğru sınıflanan örnek sayısı	212	420	640	840	1046	1243	1441
Doğru sınıflama oranı (%)	72,35	71,67	72,72	71,61	71,35	70,66	70,19
Kappa istatistiği	0,446	0,434	0,454	0,431	0,426	0,411	0,400
Mutlak hatanın ortalaması	0,374	0,382	0,378	0,381	0,380	0,384	0,383
Hataların ortalama karekökü	0,425	0,434	0,429	0,433	0,432	0,437	0,438
Göreceli mutlak hata	75,02	76,40	75,66	76,28	76,20	77,05	76,93
Göreceli hataların karekökü	84,91	86,65	85,69	86,60	86,28	87,50	87,69

Tablo 52 incelendiğinde test edilen veri setinin oranı arttıkça doğru sınıflama oranının %15'lik test veri setine kadar arttığı sonra azaldığı, kappa istatistiğinin ise %15'lik test veri setinden sonra azaldığı belirlenmiştir. Bunun yanında mutlak hatanın ortalaması monoton olarak artarken diğer hata ölçütlerinin de hem azalıp hem de arttığı görülmektedir. Görsel açıdan daha iyi anlaşılması amacıyla kappa istatistiği, mutlak hatanın ortalaması ve hataların ortalama karekökünün farklı yüzdeliklerdeki test verilerinde nasıl bir değişkenlik gösterdiği Şekil 34'te gösterilmiştir.



Şekil 34. Ridge regresyon yöntemiyle elde edilen hata ölçütlerinin değişimi.

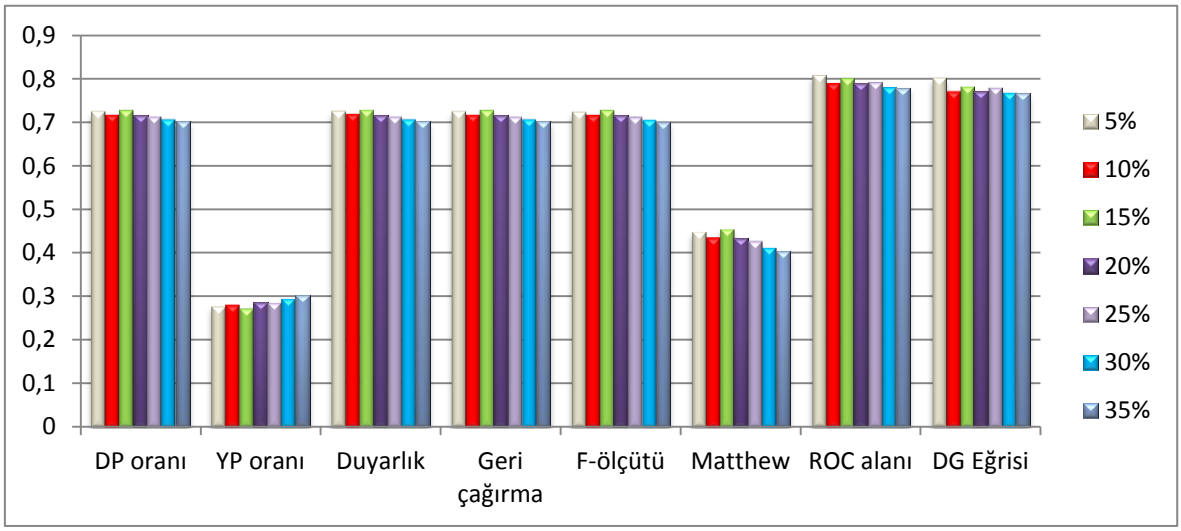
Random Tree yöntemiyle elde edilen Doğru pozitif, Yanlış pozitif, Duyarlılık, Geri çağırma, F-Ölçütü, Matthews korelasyon katsayısı (MCC), ROC eğrisi altında kalan alan (ROC Area) ve Duyarlık-Hassalık Alanı (PRC Area) ölçütlerine ilişkin sonuçlar Tablo 53'de gösterilmiştir.

Tablo 53

Ridge Regresyon Yöntemiyle Farklı Yüzdeliklerdeki Geçerlik Ölçütleri

Oran	DP oranı	YP oranı	Duyarlık	Geri çağırma	F-ölçütü	Matthew	ROC alanı	DG Eğrisi
%5	0,724	0,277	0,725	0,724	0,723	0,448	0,807	0,801
%10	0,717	0,282	0,719	0,717	0,716	0,436	0,789	0,771
%15	0,727	0,273	0,728	0,727	0,727	0,455	0,800	0,781
%20	0,716	0,285	0,716	0,716	0,716	0,432	0,790	0,772
%25	0,714	0,287	0,714	0,714	0,713	0,428	0,793	0,780
%30	0,707	0,296	0,707	0,707	0,706	0,413	0,780	0,768
%35	0,702	0,302	0,702	0,702	0,701	0,402	0,779	0,767

Tablo 53'de görüldüğü üzere test edilen veri setinin oranı arttıkça doğru pozitif oranının önce azaldığı %15'lik veri setinde bir miktar arttığı ve sonradan tekrardan azaldığı belirlenmiştir. Benzer şekilde yanlış pozitif oranının %15'lik veri setinden bir miktar arttığı sonra tekrardan azaldığı görülmektedir. Bunun yanında test edilen veri setinin oranı arttıkça duyarlığın, geri çağırmanın ve F-ölçütünün %15'lik test veri setinden sonra azaldığı belirlenmiştir. Elde edilen sonuçlar Şekil 35'te görselleştirilmiştir.



Şekil 35. Ridge regresyon yöntemiyle farklı yüzdeliklerdeki geçerlik ölçütlerinin değişimi.

Farklı Yöntemlerden Elde Edilen Sonuçların Karşılaştırılması

Genel olarak bakıldığında test edilen veri setinin %15 olması durumunda değerlendirme ölçütlerinin değişkenlik yönünde farklılaşma olmakta ve ölçütlerin tutarlı sonuçlar üretmediği görülmektedir. Öğrenme yöntemleri hata ölçütleri bakımından kendi içinde karşılaştırıldığında en tutarlı sonuçlar Ridge regresyon yöntemi ve Lojistik model tarafından elde edilmektedir. Elde edilen sonuçlardan gerçekte başarılı iken başarılı olarak tahmin edilen öğrenci sayısı anlamıona gelen doğru pozitif oranlarının farklı yüzdelikteki test verilerinde gösterdiği değişimler Tablo 54'de gösterilmiştir.

Tablo 54

Farklı Yöntemler Kullanılarak Elde Edilen Doğru Pozitif Oranları

DP Oranı	%5	%10	%15	%20	%25	%30	%35
1. Decision Stump	71,31	67,91	64,43	64,27	63,50	63,44	63,22
2. Hoefding Tree	70,64	69,96	70,22	69,99	67,80	66,00	63,17
3. J.48	69,96	68,08	66,02	68,28	67,80	65,88	67,65
4. Lojistik Model	**71,67**	**71,84**	**73,06**	**71,95**	**71,14**	**70,60**	**70,53**
5. RepTree	66,89	67,57	71,25	67,17	68,62	66,62	66,43
6. Random Forest	**72,01**	**71,16**	**73,75**	**72,37**	**73,46**	**70,83**	**69,99**
7. Random Tree	62,79	62,79	63,97	61,63	64,39	60,26	64,39
8. Ridge Lojistik Reg.	**72,35**	**71,67**	**72,72**	**71,61**	**71,35**	**70,66**	**70,19**

Tablo 54 incelendiğinde Lojistik model, Random Forest ve Ridge Regresyon yöntemlerinin farklı büyüklükteki test verilerinde en doğru tahmin sonuçlarını verirken test edilen veri seti oranı %15 olduğunda en yüksek değere ulaştığı sonra tekrardan azaldığı görülmektedir. Elde edilen sonuçların görsel olarak daha iyi anlaşılması bakımından doğru pozitif oranlarının farklı yöntemler altındaki değişimi Şekil 36'da gösterilmiştir.

Şekil 36. Farklı algoritmalar kullanılarak elde edilen doğru pozitif oranlarının değişimi.

En çok kullanılan değerlendirme ölçütlerinden biri olan hataların ortalama karekökünün farklı test veri setlerinde farklı yöntemlerde nasıl bir değişim gösterdiği Tablo 55'te verilmiştir.

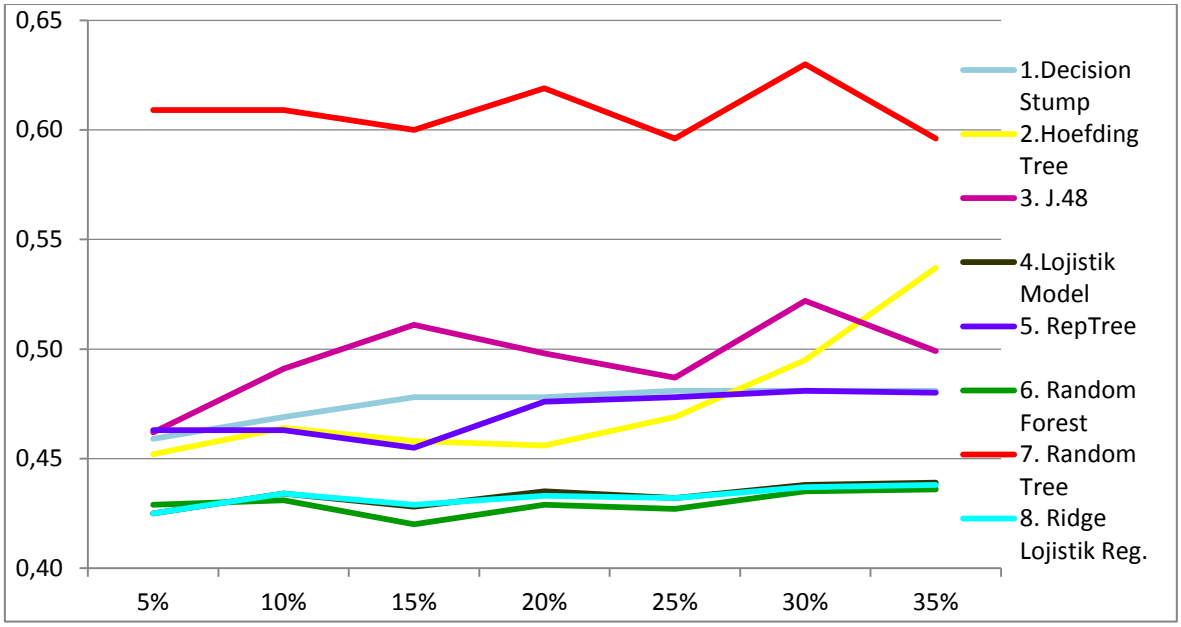
Tablo 55

Farklı Algoritmalar Kullanılarak Elde Edilen Hataların Ortalama Kareköku

Hata	%5	%10	%15	%20	%25	%30	%35
1.Decision Stump	0,459	0,469	0,478	0,478	0,481	0,481	0,481
2.Hoefding Tree	0,452	0,464	0,458	0,456	0,469	0,495	0,537
3. J.48	0,462	0,491	0,511	0,498	0,487	0,522	0,499
4.Lojistik Model	**0,425**	**0,434**	**0,428**	**0,435**	**0,432**	**0,438**	**0,439**
5. RepTree	0,463	0,463	0,455	0,476	0,478	0,481	0,480
6. Random Forest	**0,429**	**0,431**	**0,420**	**0,429**	**0,427**	**0,435**	**0,436**
7. Random Tree	0,609	0,609	0,600	0,619	0,596	0,630	0,596
8. Ridge Lojistik Reg.	**0,425**	**0,434**	**0,429**	**0,433**	**0,432**	**0,437**	**0,438**

Tablo 55 incelendiğinde Lojistik model, Random Forest ve Ridge Regresyon yöntemlerinin farklı büyüklükteki test verilerinde en düşük hata değerlerini verirken en düşük hata değerlerini test edilen veri seti oranı %15

olduğunda elde ettiği ve daha sonra tekrardan hata değerlerinin arttığı görülmektedir. Elde edilen sonuçların görsel olarak daha iyi anlaşılması bakımından Hataların Ortalama Karekökü değerlerinin farklı yöntemler altındaki değişimi Şekil 37'de gösterilmiştir.



Şekil 37. Farklı algoritmalar kullanılarak elde edilen hataların değişimi.

Hataların farklı eğitim setlerinde nasıl bir değişim gösterdiğine ilişkin görsel bir anlam kazanmak amacıyla çizilen Şekil 37 incelendiğinde Random Tree yönteminin en yüksek hata değerine sahip olduğu ve sonrasında ikinci en yüksek hata değerlerinin J.48 yöntemi tarafından elde edildiği görülmektedir. Bunun yanında en düşük hata değerlerinin Ridge regresyon, Random forest ve Lojistik modele ait olduğu görülmektedir. Öte yandan bu yöntemler tarafından elde edilen hata değerlerinin de farklı yüzdelikteki test verilerinde oldukça tutarlı olduğu sonucuna ulaşılmıştır.

Üçüncü Alt Probleme İlişkin Bulgular

PISA öğrenci anketiyle ölçülen değişkenlerden yararlanarak başarıyı tahmin etmede kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge Lojistik Regresyon yöntemleri yardımıyla edilen ölçme sonuçlarının karışıklık matrisiyle hesaplanan doğruluk dereceleri ne düzeydedir?

Çalışmanın üçüncü alt probleminde farklı yöntemler yardımıyla elde edilen ölçme sonuçlarının veri setini test ve eğitim verisi olarak ayırmayıp aynı veri seti üzerinden hem öğrenme yöntemini eğitip hem de test ettiğimizde elde edilen sonuçların doğruluk dereceleri belirlenmeye çalışılmaktadır. Bu sebeple aynı veri seti üzerinden hem eğitilip hem de test edilen öğrenme yöntemlerinin doğruluk oranları ile başarılı ve başarısız olarak tahmin ettiği öğrenci sayıları Tablo 56'da gösterilmiştir.

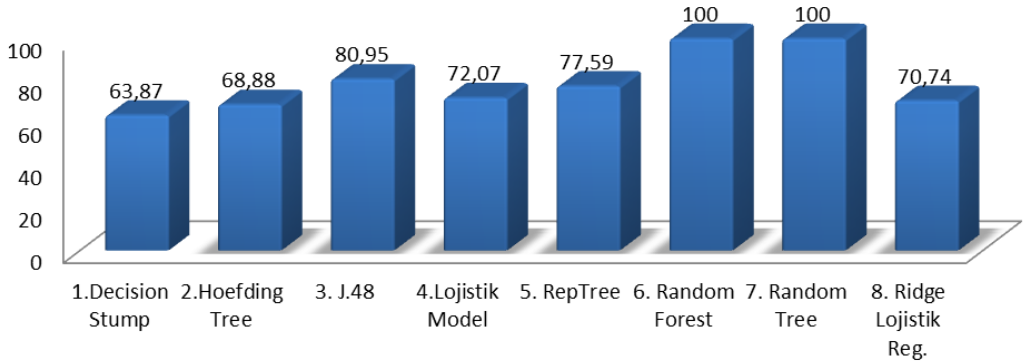
Tablo 56

Farklı Öğrenme Yöntemlerine İlişkin Karışıklık Matrisi ve Doğruluk Oranları

Öğrenme Yöntemi	Karışıklık Matrisi			Genel Başarı (Doğru sınıflama Oranı)
		Başarılı	**Başarısız**	
Decision Stump	**Başarılı**	1959	1137	% 63,87
	Başarısız	982	1787	
		Başarılı	**Başarısız**	
Hoeffding Tree	**Başarılı**	2296	800	% 68,88
	Başarısız	1025	1744	
		Başarılı	**Başarısız**	
J.48	**Başarılı**	2681	415	% 80,95
	Başarısız	702	2067	
		Başarılı	**Başarısız**	
Lojistik Model	**Başarılı**	2352	744	% 72,07
	Başarısız	894	1875	
		Başarılı	**Başarısız**	
RepTree	**Başarılı**	2578	518	% 77,59
	Başarısız	796	1973	
		Başarılı	**Başarısız**	
Random Forest	**Başarılı**	3096	0	% 100
	Başarısız	0	2769	
		Başarılı	**Başarısız**	
Random Tree	**Başarılı**	3096	0	% 100
	Başarısız	0	2769	
		Başarılı	**Başarısız**	
Ridge Regresyon	**Başarılı**	2290	806	% 70,74
	Başarısız	910	1859	

Tablo 56 incelendiğinde aynı veri seti üzerinde hem öğrenme yönteminin eğitilmesi ve yine aynı veri seti üzerinde test edilmesi sonucunda Random Forest ve Random Tree yöntemlerinin başarılı ve başarısız olarak tüm öğrencileri doğru sınıflayarak %100 doğru sınıflama oranına sahip oldukları belirlenmiştir. Sonrasında en iyi doğru sınıflama oranına sahip öğrene yönteminin J.48

algoritması olduğu ve başarı oranının % 80,95 olduğu belirlenmiştir. Bunların haricinde en düşük doğru sınıflama oranına sahip öğrenme yönteminin Decision Stump ve başarı oranının % 63,87 olduğu tespit edilmiştir. Elde edilen sonuçların bir bütün olarak grafik üzerinde gösterimi Şekil 38'de verilmiştir.



Şekil 38. Farklı öğrenme yöntemlerine ilişkin doğruluk oranları

Her ne kadar 10 katlı çapraz geçerleme yöntemiyle başarı oranı %67,28 olan Random tree ve başarı oranı %67,79 olan J.48 yöntemleri tüm veri setinin eğitim ve test verisi oalrak ayrılmaması durumunda sırasıyla %100 ve %80,95 gibi daha yüksek doğru sınıflama oranına sahip oldukları görülmektedir. Elde edilen bu sonucun ortaya çıkmasında aşırı uyum gösterme (*over fitting*) olarak bilinen sorundan kaynaklandığı düşünülmektedir. Nitekim Zhou ve Cervantes (2016) aşırı uyum gösterme sorununun örneklemin evrenin iyi bir temsilcisi olmaması ve bu sebeple eğitim setindeki varyansın yüksek olması ile eğitim veri setindeki yanlış sınıf özellikleri ve yanlış sınıf etiketi gibi gürültülü durumun olmasından kaynaklandığını belirtmektedir. Bu sebeple yeterli delil olmaması durumunda ağaç büyümesinin durdurulması veya daha fazla düğümün oluşmasına engel olunması gibi çözüm yöntemleri önerilmektedir (Bramer, 2013). Araştırma sonucunda rastgele ağaç yönteminin yeterli kanıt olmamasına rağmen büyümeye devam etmesi sonucunda aslında gerçekçi olmayan %100 gibi bir başarı oranına sahip olduğu görülmektedir.

Bölüm 5
Sonuç, Tartışma ve Öneriler

Sınıflandırma ve tahminde bulunma amacıyla kullanılan veri madenciliği yöntemleri, çeşitli bilim dallarında her geçen gün daha popüler hale gelmektedir. Bununla birlikte bu yöntemlerin eğitim alanında çok fazla kullanılmadığı görülmektedir (Sinharay, 2016). Denetimli öğrenme yöntemi olarak tanımlanan VM yöntemleri klasik yöntemlere kıyasla daha iyi bir kestirim gücüne ve daha az soruna sahip olduğu belirtilmektedir (Fernandez-Delgado, Cernadas, Barro ve Amorim, 2014). Bu çalışmada veri madenciliği alanında kullanılan öğrenme yöntemlerinin eğitim alanında kullanılmasıyla elde edilen sonuçlara ilişkin geçerlik ve güvenirlik değerleri farklı koşullar altında incelenmiştir. Buna göre elde edilen araştırma sonuçları alan yazında yapılan çalışmalar ile karşılaştırılarak benzer ve farklı yönleri ortaya çıkarılmış ve ileride yapılacak araştırmalar için önerilerde bulunulmuştur.

5.1. Sonuçlar ve Tartışma

Araştırmadan elde edilen sonuçlar alt problemlere ilişkin başlıklar altında açıklanmıştır. Çalışmanın alt problemlerine geçmeden önce fen okuryazarlığını yordamak amacıyla kullanılan değişken sayısının PISA öğrenci anketi esas alındığında çok fazla olması sebebiyle öncelikle çalışma kapsamında kullanılacak değişkenler belirlenmiştir. Her ne kadar YSA yöntemleri regresyon analizi yöntemine kıyasla modele eklenen değişken sayısının fazla olması durumunda daha iyi tahmin yapacağı belirtilse de (Lykourentzou, Giannoukos, Mpardis, Nikolopoulos ve Loumos, 2009); VM gibi tahmine dayalı yöntemlerde modele daha çok sayıda değişken eklemenin performans kestiriminin doğruluğunda bir artışa sebep olmayacağını belirtmiştir (Huang ve Fang (2013). Bunun yanında Kohavi (1995) tarafından yapılan deneysel çalışmada farklı algoritmalar altında yapılan kestirimlerin katman sayısının 10 ve 20 olması durumunda oldukça iyi olduğu ve neredeyse tamamen yansız olduğu belirtildiği için çalışmada 10 katlı çapraz geçerleme yöntemi kullanılmıştır. Değişken sayısını azaltmak amacıyla öncelikle VM yöntemlerinden Best First Forward, Best First Backward ve Greedy Stepwise yardımıyla 10 katlı çapraz geçerleme yöntemi sonucunda en az %40 ve üzeri

başarılı olan toplam 12 değişken yardımıyla PISA fen okuryazarlığı tahmin edilmeye çalışılmıştır.

Birinci alt probleme ilişkin sonuçlar. Çalışmanın birinci alt probleminde verileri eğitim ve test verisi olarak ayırmadan doğrudan 10 katlı çapraz geçerleme yöntemi yardımıyla kullanılan veri madenciliği tahmin yöntemlerinin güvenirlik ve geçerlik değerleri karşılaştırılmıştır. Büyük veri setlerinde başarının tahmin edilmesinde yoğun bir hesaplama süreci olması sebebiyle veri madenciliği yöntemlerinin kullanıldığı çalışmada elde edilen sonuçlara ilişkin istatistikler incelenmiştir (Hastie, Tibshirani ve Friedman, 2009). Buna göre kullanılan güvenirlik ölçütlerinden doğru sınıflanan örnek sayısı, kappa istatistiği, mutlak hata ve ortalama hataların karekökünе göre en iyi tahmin yönteminin Random forest olduğu belirlenmiştir. Elde edilen bu sonuç Liaw ve Wiener (2002) ile Svetnik, Liaw, Tong ve Wang (2004) tarafından yapılan çalışmalarla benzerlik göstermektedir. Random forest yönteminde bir sınıflandırıcı yerine birden çok sınıflandırıcı kullanılması ve sonrasında onların tahminlerinden elde edilen sonuçlar ile yeni veriyi sınıflandıran öğrenme algoritmaları oluşturulduğu için daha az hatalı ve daha güvenilir sonuçlar elde edildiği düşünülmektedir (Strobl, Malley ve Tutz, 2009). Daha sonra sırasıyla Ridge Lojistik Regresyon, Lojistik model, Hoeffding Tree, J4.8, RepTree, Decision Stump ve son olarak Random Tree yöntemlerinin en az hataya sahip ve en güvenilir yöntemler olduğu belirlenmiştir. Bunun yanında tahminleme yöntemlerinin geçerlik ölçütlerinden doğru pozitif, yanlış pozitif, duyarlık, geri çağırma, f-değeri, MKK, ROC eğrisi altında kalan alan ve Duyarlık-geri getirme (DG) eğrisine göre en iyi yöntemin Random forest olduğu belirlenmiştir. Daha sonra geçerlik ölçütlerine göre sırasıyla Ridge Lojistik Regresyon, Lojistik model, Hoeffding Tree, J4.8, RepTree, Decision Stump ve son olarak Random Tree yöntemlerinin daha başarılı oldukları belirlenmiştir. Elde edilen sonuçlar birlikte değerlendirildiğinde yöntemlerin güvenirlik ve geçerlik ölçütleri bakımından başarı sırasında bir farklılık olmadığı her bir yöntemin kendi içinde kararlı olduğu belirlenmiştir. Alanyazında en çok kullanılan yöntem olarak görülen J4.8 algoritması orta düzeyde bir başarı göstermiştir. Bunun yerine Random forest, Ridge lojistik regresyon, lojistik model ve Hoeffding tree yöntemlerinin hem hataya dayalı güvenirlik değerleri hem de geçerlik ölçütleri bakımından daha başarılı yöntemler olduğu belirlenmiştir.

İkinci alt probleme ilişkin sonuçlar. Çalışmanın ikinci alt probleminde farklı yöntemler yardımıyla test edilen verilerin %5, %10, %15, %20, %25, %30 ve %35 olması durumunda elde edilen sonuçların güvenirlik ve geçerlik değerleri karşılaştırılmıştır. Sinha ve May (2005) bekletme yaklaşımı olarak da bilinen eğitim ve test veri seti yönteminin model performansını tahmin etmede büyük örneklem gruplarında başarılı olurken nispeten daha küçük gruplarda bu yöntemin normalden daha iyimser sonuçlar vereceğini belirtmektedir. Bu sınırlılığı ortadan kaldırmak için tesadüfi olarak belirlenen eğitim ve test veri setlerinin ortalaması alınarak elde edilen çapraz geçerleme yönteminin kullanılması önerilmektedir (Weiss ve Kulikowski, 1991). Bu sebeple çalışmada farklı veri setlerinde öğrenme yöntemlerinin nasıl bir performans gösterdiği incelenmiş ve doğru pozitif oranları bakımından en iyi sonuçların sırasıyla Ridge lojistik regresyon, Random forest, Lojistik model, Decision stump, Hoeffding tree, J4.8, Rep tree ve Random tree yöntemleri tarafından elde edildiği belirlenmiştir. Ortalama hataların karekökleri esas alınarak karşılaştırma yapıldığında en düşük hata değerlerinin tüm örneklem gruplarında sırasıyla Ridge lojistik regresyon, Random forest, Lojistik model, Hoeffding tree, Decision stump, J4.8, Rep tree ve Random tree yöntemleri tarafından elde edildiği belirlenmiştir. Yöntemlerin farklı örneklem gruplarında nasıl bir değişim gösterdiği incelendiğinde Decision stump yönteminin test edilen veri seti %15 ve üzerinde olması durumunda daha kararlı bir kestirim yaptığı tespit edilmiştir. Hoeffding tree yönteminin test edilen veri seti %20'ye kadar olması durumunda kararlı sonuçlar üretirken bu değerden sonra hata değerlerinin artarken kappa istatistiğinin azaldığı belirlenmiştir. J4.8 algoritmasının hata değerlerinin farklı örneklemlerde hem arttığı hem de azaldığı görülmüş ve elde edilen sonuçların kararlı olmadığı tespit edilmiştir. Lojistik modelde test edilen veri setinin hata ve kappa istatistiğinde anlamlı düzeyde bir farklılığa sebep olmadığı ve yöntemin oldukça kararlı kestirim yaptığı belirlenmiştir. Rep tree yönteminin de hata değerlerinin farklı veri setlerinde hem arttığı hem de azaldığı görülmüş ve elde edilen sonuçların kararlı olmadığı tespit edilmiştir. Random forest yöntemi farklı veri setlerinde hata değerleri bakımından kararlı kestirimlerde bulunurken sadece kappa istatistiğinin %15-%25 aralığında bir miktar arttığı sonra tekrardan aynı düzeye indiği belirlenmiştir. En yüksek hata değerlerine sahip olan Random tree yönteminin test edilen veri setinin %15 oluncaya kadar nispeten kararlı kestirimlerde bulunurken sonrasında test edilen ölçütlerde artma ve azalmalar

olduğu görülmüştür. En iyi kestirimde bulunan Ridge regresyon yönteminin tüm test verilerinden oldukça kararlı kestirimlerde bulunduğu belirlenmiştir.

Üçüncü alt problemlere ilişkin bulgular. Çalışmanın üçüncü alt probleminde farklı yöntemler yardımıyla elde edilen ölçme sonuçlarının veri setini test ve eğitim verisi olarak ayırmayıp aynı veri seti üzerinden hem öğrenme yöntemini eğitip hem de test ettiğimizde karışıklık matrisi yardımıyla elde edilen sonuçların doğruluk dereceleri karşılaştırılmıştır. Buna göre Random forest ve Random tree yöntemlerinin tüm örnekleri doğru olarak sınıflayarak en yüksek sınıflama oranına sahip olduğu belirlenmiştir. Üçüncü en yüksek doğru sınıflama oranına sahip J4.8 algoritmasını sırasıyla Rep Tree, Lojistik model, Ridge regresyon, Hoeffding tree ve Decision stump yöntemleri takip etmektedir. Aynı veri seti üzerinde kestirimde bulunulduğu taktirde aslında hatalı ve düşük doğru sınıflama oranlarına sahip olan Random tree ve Rep tree yöntemlerinin %100 gibi oldukça başarılı birer algoritma olarak belirlenmesi oldukça düşündürücüdür. Nitekim bu sorun veri madenciliğinde aşırı uyum gösterme (overfitting) olarak tanımlanan sorundur (Domingos, 2012). Her ne kadar aşırı uyum gösterme sorunu yanlı kestirim ve yüksek varyans gibi farklı şekillerde ortaya çıkabilse de aynı veri seti üzerinde çalışmak bu soruna sebep olabilmektedir (Domingos, 2000). Bu sorunu ortadan kaldırmanın en iyi yöntemlerinde biri de çapraz geçerleme uygulamaktır (Ng, 1997). Araştırmada çapraz geçerleme yöntemi uygulandığında aşırı uyum gösteren Random Tree ve RepTree yöntemlerinin genel başarı oranlarının oldukça düştüğü belirlenmiştir. Bu sonuca göre alanda çalışma yapacak araştırmacıların topladıkları verileri eğitim ve test veri seti olarak ayırmamaları halinde kullanacakları yöntemlerin hatalı sonuçlar vereceği görülmektedir. Bu sebeple araştırmacıların veri madenciliği alanında çalışma yaparken test edilen veri setini en az %15 ve üzerinde tutmaları gerekmektedir. Bunun yanında daha kararlı ve daha doğru kestirimlerde bulunmak için çapraz geçerleme yöntemlerini kullanmaları ve bu aşamada kullanacakları katman sayısının en az 10 ve üzeri olması önerilmektedir.

Fen okuryazarlığının yordanması üzerine yapılan çalışmalar incelendiğinde Kaya (2017) öğrencilerin PISA 2015 sınavından elde edilen sonuçlara göre okuma becerileri ile fen okuryazarlığı arasında pozitif yönde ve yüksek düzeyde istatistiksel olarak anlamlı bir ilişki olduğu belirlenmiştir. Chiu (2007) tarafından

yapılan çalışmada daha çok eğitimsel kaynaklara sahip olan öğrencilerin fen bilimleri okuryazarlığında daha başarılı oldukları sonucuna ulaşmıştır. Usta ve Çıkrıkçı-Demirtaşlı (2014) tarafından yapılan çalışmada öğrencilerin fen bilimleri öğrencisi olarak kendilerini yeterli görmeleri fen bilimleri okuryazarlığındaki değişkenliğin bir kısmını doğrudan açıklarken bir kısmı da bilimsel sorgulamaya verilen önem aracılığı ile açıklandığı belirlenmiştir. Çalışmada bilimsel sorgulamaya verilen önemin fen okuryazarlığı üzerinde anlamlı bir etkide bulunurken fen bilimlerinin toplum için yararlı olduğuna ilişkin görüşlerinin fen bilimlerindeki başarılarında doğrudan bir etkiye sahip olmadığı belirlenmiştir. Her ne kadar bilimsel inançlar özelliği veri madenciliğinde 10 katlı çapraz geçerlemenin her aşamasında fen okuryazarlığı üzerinde anlamlı bir etkiye sahip olması çalışma sonuçlarıyla benzerlik gösterse de fen öz yeterliğinin fen okuryazarlığında etkili olmadığı Usta ve Çıkrıkçı-Demirtaşlı (2014) tarafından yapılan çalışmanın bulgularıyla farklılık göstermektedir. Bu farklılık oluşmasında ilgili çalışmada yapısal eşitlik modeli ile sadece 4 değişken arasındaki doğrudan ve dolaylı ilişkiler incelenirken veri madenciliğinde öğrencilere ilişkin 12 farklı özelliğin ele alınmasının sebep olduğu düşünülmektedir. Benzer bir çalışmada Miller (2006) öğrencilerin kendilerinin yeterli görmeleri ile fen başarıları arasında anlamlı bir ilişki olduğu belirlemiştir. Anıl (2009) yaptığı çalışmada fen okuryazarlığında anne eğitim durumu, baba eğitim durumu, bilgisayar ortamı, tutum ve aile kültür zenliği değişkenleri arasında düşük düzeyde istatistiksel olarak anlamlı ilişkiler bulunmasına rağmen çalışma kapsamında ele alınan beş değişkenin fen bilimleri başarı puanındaki toplam varyansın yaklaşık %20'sini açıkladığı belirlenmiştir. Elde edilen bu sonuç fen okuryazarlığını yordamada anlamlı etkiye sahip olan başka değişkenler olabileceğini göstermektedir. Nitekim veri madenciliği ile fen okuryazarlığının yordanmasında çevreye duyarlık, fen öğrenme süresi, sosyo ekonomik durum indeksi, öğretmenin adil olması, bilimsel inançlar ve evdeki eğitim kaynakları değişkenlerinin özellikle karar ağaçlarında öğrencileri başarılı ve başarısız tahmin etmede en başarılı değişkenler olduğu belirlenmiştir. Ancak çalışmada fene yöntelik tutum ve aile özelliklerinin fen okuryazarlığı üzerinde anlamlı bir etkide bulunmaması Kartal, Doğan ve Yıldırım (2017) tarafından yapılan çalışmanın sonuçlarıyla farklılık göstermektedir.

Model değerlendirme ölçütlerinden ROC eğrisi altında kalan alanların incelenmesinde Elayidom (2012) tarafından yapılan değerlendirme kriterleri esas alınmıştır. Buna göre eğri altındaki alan 1.00-0.90 arası mükemmel, 0.80-0.89 arası iyi, 0.70-0.79 arası normal, 0.60-0.69 arası zayıf ve 0.50-0.59 arası başarısız olarak yorumlanmaktadır. Çalışmada farklı algoritmalar yardımıyla 10 katlı çapraz geçerleme sonucunda sırasıyla Random forest, lojistik model, Ridge lojistik regresyon, Hoeffding tree ve Rep tree yöntemlerinin başarılı ve başarısız olarak tanımlanan öğrencileri normal seviyede tahmin ederken J4.8, Random tree ve Decision stump yöntemlerinin zayıf olarak tanımlanan seviyede kestirimde bulundukları belirlenmiştir. Mehdiyev, Enke, Fettke ve Loos (2016) tarafından yapılan benzer bir çalışmada en doğru kestirimlerin sırasıyla yapay sinir ağları, Random forest, Lojistik regresyon, Radyal temelli ağlar ve C4.5 yöntemleri tarafından gerçekelştirildiği görülmektedir. Elde edilen bu sonuç araştırmanın bulgularıyla benzerlik göstermektedir. Nitekim Weka programında yer alan sekiz farklı öğrenme yönteminin farklı değerlendirme ölçütleri altında en iyi sonuçlar Random forest, Ridge regresyon yöntemi ve lojistik model tarafından elde edilmiştir.

Farklı algoritmalar tarafından başarının yordanması amacıyla elde edilen karar ağaçları incelendiğinde her ne kadar Hoefding tree, J.48, Lojistik model, Reptree ve Random Tree algoritmaları için karar ağaçları program tarafından kolaylıkla oluşturulabilse de diğer algoritmalar için karar ağaçlarının elde edilmemesi bu yöntemlerin bir sınırlılığı olarak görülmektedir. Bu sınırlılığın sebebi karar ağaçlarını grafiksel olarak görselleştirmenin karmaşık olması ve bu alanda çalışmaların halen devam etmesidir (Engel, Charao, Kirsch-Pinheiro ve Steffenel, 2014). Elde edilen ağaç yapıları incelendiğinde Hoefding tree yönteminde fen okuryazarlığını en iyi yordayan değişkenler bilgi ve iletişim teknolojileri kaynakları ile öğrencilerin çevreye duyarlılığıdır. J4.8 yöntemi tarafından oluşturulan karar ağacında fen okuryazarlığını en iyi yordayan değişkenler ilk olarak çevreye duyarlık ve sonrasında sosyo ekonomik durum indeksi ile haftalık fen öğrenme süresi olduğu belirlenmiştir. Lojistik model yöntemi tarafından oluşturulan karar ağacında sırasıyla çevreye duyarlık, fen öğrenme süresi, sosyo ekonomik durum indeksi ve öğretmenin adil olması değişkenlerinin fen okuryazarlığını anlamlı bir şekilde yordadığı belirlenmiştir. Rep tree yöntemi tarafından oluşturulan karar

ağacında fen okuryazarlığını en iyi yordayan değişken fen öğrenme süresi ve sonrasında çevreye duyarlık ve bilgi iletişim teknolojileri kaynakları olduğu tespit edilmiştir. Random tree yönteminde ise sırasıyla fen öğrenme süresi, çevreye duyarlık, evdeki eğitim kaynakları ve toplam öğrenme sürelerinin başarıyı yordayan önemli değişkenler olduğu belirlenmiştir. Karar ağaçları bir bütün olarak incelendiğinde daha önce değişken seçimi aşamasında 10 farklı katmanın her birinde üç farklı yönteme göre %100 başarılı olan çevreye duyarlık, fen öğrenme süresi ve bilgi-iletişim teknolojileri kaynaklarının fen okuryazarlığını tahmin etmede en önemli değişkenler olduğu belirlenmiştir. Nitekim çapraz geçerleme sonucunda yine %100 başarıya sahip bilimsel inançlar özelliği sadece Random tree yönteminde anlamlı bir etkide bulunduğu belirlenmiştir. Bunun yanında 10 katlı çapraz geçerlemenin her bir aşamasında başarılı olan sorgulamaya dayalı fen eğitimi, okul dışı ders çalışma süresi ve öğrenciden beklenen mesleki statü değişkenlerinin karar ağaçlarının dallarının ayrışmasında bir düğüm olarak yer almadığı belirlenmiştir. Yine benzer şekilde %40'lık bir başarıya sahip olan öğretmenin adil olması özelliği lojistik modelde bir düğüm olarak karşımıza çıkarken %80'lik başarıya sahip evdeki eğitim kaynakları ve %60'lık başarıya sahip test kaygısı karar ağaçlarının hiçbirinde yer almamıştır. Elde edilen bu sonuç her bir katmadan öğrencileri başarı ve başarısız olarak sınıflamada başarılı olan yöntemlerin hepsinin karar ağacında birer düğüm olamayacağını göstermektedir. Nitekim ağaçlardaki düğüm noktaları bilgi kazanç fonksiyonları yardımıyla elde edildiğinden (Nowozin, 2012) başarının yordanmasına ilişkin tek bir karar ağacı oluşturmak yerine birden fazla yöntem tarafından elde edilen karar ağaçlarının raporlanarak benzer ve farklı yönlerinin ortaya çıkarılması gerekmektedir. Özellikle fen, matematik ve Türkçe okuryazarlığını yordamaya yönelik çalışma yapacak araştırmacıların tek bir karar ağacına dayalı olarak çalışma yapmak yerine farklı algoritmalar yardımıyla elde ettikleri farklı ağaç yapılarını yorumlamaları önerilmektedir.

Her ne kadar veri madenciliğinde farklı öğrenme yöntemlerini birbirleriyle farklı ölçütlere göre kıyaslamak mümkün olsa da her bir öğrenme yönteminin ne düzeyde kestirim yaptığına ilişkin standart bir değerlendirme ölçütü bulunmamaktadır (Sokolova ve Lapalme, 2009; Cai, Li, Kang ve Liu, 2009; Vihinen, 2011; Tiwari, Jha ve Yadav, 2012; Doreswamy, 2012; Kiranmai ve

Damodaram, 2014; Hossin ve Sulaiman, 2015; Almuniri ve Said, 2017). Ancak doğru pozitif, doğru negatif, duyarlık ve hassaslık gibi farklı değerlendirme ölçütleri yardımıyla Pozitif olabilirlik oranı (*positive likelihood ratio*), Ayırıcılık gücü (*discriminant power*) ve Dengelenmiş doğruluk (*balanced accuracy*) şeklinde üç farklı standart değerlendirme ölçütü geliştirilmiştir (Bekkar, Djemaa ve Alitouche, 2013). Pozitif olabilirlik oranı (L) için 1'e eşit olanlar önemsiz, 1-5 arası düşük, 5-10 arası kayda değer ve 10'dan büyük olanlar iyi olarak tanımlanmaktadır. Ayırıcılık gücü için 1'den küçük olanlar zayıf, 1-2 arası sınırlı, 2-3 arası kayda değer ve 3'ten büyük olanlar iyi olarak sınıflandırılmaktadır. Bunun yanında dengelenmiş doğruluk değerinin duyarlılık ve seçililik toplamının yarısına eşit olduğu düşünüldüğünde alabieceği en yüksek değerin 1'e eşit olacağı bilinmektedir. Her ne kadar diğer ölçütler gibi standart bir değerlendirme ölçütü bulunmasa da dengelenmiş doğruluk oranının 1'e eşit olması beklenmektedir. Buna göre Weka programında 10 katlı çapraz geçerleme yöntemiyle test edilen yöntemlerin olabilirlik oranları, ayırıcılık güçleri ve dengelenmiş doğruluk değerleri Tablo 57'de gösterilmiştir.

Tablo 57

Veri madenciliği yöntemlerini karşılaştırmada kullanılabilecek standart ölçütler

Yöntem	Pozitif olabilirlik oranı	Ayırıcılık gücü	Dengelenmiş doğruluk
Decision stump	1,711	0,273	0,203
Hoeffding tree	2,236	0,396	0,242
J4.8	2,086	0,355	0,229
Lojistik model	2,326	0,411	0,247
RepTree	2,020	0,343	0,226
Random forest	2,468	0,433	0,254
Random tree	1,651	0,251	0,197
Ridge regresyon	2,351	0,418	0,249

Veri madenciliği yöntemlerinin pozitif olabilirlik esas alınarak karşılaştırma yapıldığında Random forest, lojistik model, ridge regresyon, hoeffding tree, J4.8 ve rep tree yöntemlerinin düşük düzeyde; decision stump ve random tree yöntemlerinin ise önemsiz düzeyde kestirim yaptığı belirlenmiştir. Bunun yanında ayırıcılık gücüne göre tüm yöntemlerin zayıf düzeyde kestirim yaptığı belirlenmiştir. Ancak doğru ve yanlış sınıflanan örneklerin sınıf ağırlığına bağlı yüksek doğruluk oranını ortadan kaldıran dengelenmiş doğruluk değerlerinin oldukça düşük olduğu görülmektedir (Bekkar, Djemaa ve Alitouche, 2013). Elde edilen bu ölçütler her ne

kadar alanyazında çok fazla kullanılmasa da alternatif değerlendirme ölçütleri olarak ele alınabilir.

5.2. Öneriler

Araştırmadan elde edilen bulgulara dayalı olarak geliştirilen öneriler araştırmaya ve uygulamaya dönük öneriler olarak iki başlık altında ele alınmıştır.

Uygulamaya dönük öneriler. Araştırmadan elde edilen bulgulara dayalı olarak aşağıdaki önerilerde bulunulmuştur.

1. Değişken sayısını azaltmak ve en çok bilgi veren özellikleri belirleme amacıyla farklı yöntemlerle 10 katlı çapraz geçerleme yaklaşımı ile çalışma kapsamında kullanılacak değişken sayısı 29'dan 12'ye indirilmiştir. Elde edilen bulgulara dayalı olarak fen, matematik ve Türkçe gibi derslerdeki başarının yordanması amacıyla sınırlı sayıda bağımsız değişken almak yerine çok sayıda değişkenin farklı katmanlardaki doğru tahminleme başarıları incelenerek başarısız olan değişkenler analize dahil edilmeyerek daha tutarlı ve güvenilir kestirimler elde edilebilir.

2. Çalışmada farklı yöntemlerle elde edilen karar ağaçlarında çıktı değişkeni üzerinde etkisi olan girdi değişkenlerin farklılık gösterdiği belirlenmiştir. Elde edilen bu bulgulara dayalı olarak akademik başarının yordanmasına ilişkin çalışmalarda veri madenciliğinde tek bir yöntemle çalışmak yerine en az 3 farklı öğrenme yöntemi kullanarak elde edilen sonuçların geçerliğine ilişkin delil sunulmalıdır.

3. Çalışma kapsamında elde edilen sonuçların güvenirliğine ilişkin delil sunmak amacıyla mutlak ve bağıl hata değerlerinin farklı birimlerle ifade edildiği belirlenmiştir. Elde edilen bu bulguya dayalı olarak veri madenciliğinde elde edilen sonuçların güvenirliğine ilişkin bilgi vermek amacıyla ortalama hataların karekökünün yanında göreceli ve mutlak hata değerlerinin de rapor edilmesi önerilmektedir.

4. Veri madenciliğinde örneklemin yeterince büyük olması ve üzerinde çalışılan öğrenme yöntemini aynı veri seti üzerinde eğitmek ve test etmek yerine çapraz geçerleme gibi yöntemlerin kullanılması gerekmektedir. Aksi takdirde aşırı tahminleme sorunuyla karşılaşılacağı ve aslında doğru sınıflama oranı düşük olan

öğrenme yöntemlerinin başarılı olarak değerlendirileceği için bu konuya dikkat edilmesi gerekmektedir.

5. Örneklemin 1000 ve üzerinde olması durumunda test edilen veri setinin en az %15 ve üzeri olması gerekmektedir. Aksi taktirde elde edilen sonuçların tutarlı olmayacağının göz önünde bulundurulması gerekir.

6. Veri madenciliğinde elde edilen sonuçların sadece karışıklık matrisinden elde edilen doğruluk değerlerine göre yorumlamak yerine duyarlık, seçicilik, ROC eğrisi altında kalan alan, Matthew korelasyon katsayısı, F-ölçütü ve Geri çağırma gibi standart ölçütlerin de mutlaka raporlanarak elde edilen sonuçların güvenirlik ve geçerliğine ilişkin birden fazla değerin rapor edilmesi önerilebilir.

7. Elde edilen güvenirlik ve geçerlik değerleri farklı yöntemlerle farklılık gösterse de bunları ortak bazı ölçütlere göre değerlendirmek sonuçları daha iyi yorumlamayı sağlamaktadır. Bu nedenle elde edilen sonuçların güvenirliği için pozitif olabilirlik oranı, ayırıcılık gücü ve dengelenmiş doğruluk gibi standartlaştırılmış değerlendirme ölçütlerinin kullanılması modelleri daha iyi karşılaştırmaya yardım edeceğinden çalışmalarda bu ölçütlerin de rapor edilmesi önerilebilir.

8. Veri madenciliğinde özellikle Weka programında yer alan öğrenme yöntemlerinden Random forest, lojistik model (LMT) ve Ridge regresyon yöntemlerinin farklı koşularda altında genellikle en güvenilir ve geçerli sonuçları elde ettiği belirlendiğinden başarının yordanması veya sınıflama amacıyla yapılacak çalışmalarda bu öğrenme yöntemlerinin kullanılması önerilmektedir.

9. Özellikle lojistik modelde başarılı ve başarısız olarak sınıflanan öğrencilere ilişkin elde edilecek karar ağacının yapraklarında yer alan lojistik denklemler yardımıyla değişkenler arasındaki ilişki tek bir regresyon denklemi ile modellemek yerine farklı özellik değerlerine göre çok sayıda regersyon denklemi kullanarak doğru sınıflama oranı artırılabilir.

10. Bekletme yönteminde test veri setinin en az %15 ve üzeri olması durumunda geçerlik ve güvenirlik ölçütlerinin en yüksek değere ulaşması

sabebiyle bu alanda çalışma yapacak araştırmacıların daha düşük orandaki test veri setlerinden elde edecekleri sonuçlara dikkat etmeleri önerilmektedir.

11. Araştırmacıların elde ettikleri verileri eğitim ve test verisi olarak ayırarak elde ettikleri sonuçları 10-15-20 katmanlı çapraz geçerleme sonuçlarıyla karşılaştırmaları önerilmektedir. Bu sayede çapraz geçerleme için kullanılacak ideal katman sayısının belirlenebileceği düşünülmektedir.

12. Çalışmada PISA verileri kullanarak analiz yapıldığından ileride ÖSYM tarafından yapılan geniş ölçekli sınavlardan elde edilen sonuçların korelasyon ve regresyona dayalı standart istatistiksel teknikler yerine daha zengin kanıtlar sunmak amacıyla veri madenciliğine dayalı yöntemlerle analiz edilmesi önerilmektedir.

Araştırmacılara dönük öneriler. Araştırmadan elde edilen bulgulara dayalı ileride yapılacak olan çalışmalar için aşağıdaki önerilerde bulunulmuştur.

1. Bu çalışmada veri madenciliği alanında en çok kullanılan programlardan biri olan Weka ile farklı öğrenme yöntemlerinin farklı koşullar altında güvenirlik ve geçerlik değerleri karşılaştırılmıştır. Bundan sonra yapılacak çalışmalarda farklı programlar tarafından elde edilen sonuçların karşılaştırılması üzerinde durulabilir.

2. Her ne kadar Weka programının açık kaynak kodlu olması diğer yazılımlara göre bir üstünlük olarak görülse de program yazma konusunda uzmanlaşmış araştırmacıların bu alanda çalışma yapması, programa yeni özellikler eklenmesi önerilebilir.

3. Benzer bir çalışma farklı veri tiplerinin olduğu TIMMS, PIRLS vb. sınavlar üzerinden tekrarlanabilir.

4. MEB öğrencilerin Fen, Matematik, Türkçe vb. gibi derslerindeki başarılarını yordamak için e-okul veri tabanından yararlanarak veri madenciliğine dayalı çalışmalar yapabilir.

5. İlköğretim 8. Sınıfta uygulanan ve diğer sınıflarda da yaygınlaştırılmaya çalışılan Akademik Becerilerin İzlenmesi ve Değerlendirilmesi Projesi (ABİDE)

kapsamında yapılan sınavlardaki başarının yordanmasında VM yöntemlerinden yararlanarak farklı analizler gerçekleştirilebilir.

6. Açıköğretim, açık lise, vb. sınavlardaki başarının yordanması VM yöntemlerinden yararlanarak yapılabilir.

7. Küçük örneklemlerde hangi VM yönteminin daha iyi çalıştığını belirlemek için araştırmalar tasarlanabilir.

Kaynaklar

Abdous, M., He, W., &Yen, C-J. (2012) Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade, *Journal of Educational Technology & Society, 15* (3), 77-88.

Abad, F. M., & Lopez, A. C. (2017) Data-mining techniques in detecting factors linked to academic achievement, School Effectiveness and School Improvement, *School Effectiveness and School Improvement*, *28* (1), 39-55.

Abdeljaber,T.F.(2007) A review of associative classification mining. *Knowledge En gineering Review*, 22 (1), 37-65.

Ahmed, A. B., & 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.

Almuniri, I. & Said, A. M. (2017). School's Performance Evaluation Based on Data Mining. *International Journal of Engineering and Information Systems*, *1* (9), 56-62.

Angeli, C., & Valanides, N. (2013). Using educational data mining methods to assess field-dependent and field-independentlearners' complex problem solving, *Educational Technology Research and Development*, *61* (3), 521-548.

Anıl, D. (2009). Uluslararası Öğrenci Başarılarını Değerlendirme Programı (PISA)'nda Türkiye'deki Öğrencilerin Fen Bilimleri Başarılarını Etkileyen Faktörler, *Eğitim ve Bilim, 34* (152), 87-100.

Aydın, S. (2007). *Veri madenciliği ve Anadolu Üniversitesi uzaktan eğitim sisteminde bir uygulama*. Yayımlanmamış doktora tezi, Anadolu Üniversitesi, Eskişehir.

Barros, R.C. et al., (2015) Automatic Design of Decision-Tree Induction Algorithms, *Springer Briefs in Computer Science*, *176*, 7-45.

Bekkar, M., Dhemaa, H. K., & Alitouche, T. A. (2013) Evaluation Measures for ModelsAssessment over Imbalanced Datasets, *Journal of Information Engineering and Applications*, *3*, 10.

Boss, D. D. (2003). Introduction to the Bootstrap World, *Statistical Science*, *18* (2), 168-174.

Bramer, M. (2013). *Principles of Data Mining (2nd ed.),* London: Springer-Verlag.

Brown, M. S. (2014). *Data Mining For Dummies*, Hoboken, New Jersey: John Wiley & Sons.

Ceci, M. (2005). *Naive Bayesian Learning from Structural Data* (Unpublished Doctoral Thesis), Computer Science and Engineering, University of Bari, Italy.

Chadha, P., & Singh, G. N. (2012). Classification Rules and Genetic Algorithm in Data Mining, *Global Journal of Computer Science and Technology Software & Data Engineering, 12* (15), 50-54.

Chamatkar, A. J., & Butey, P. K. (2014) Importance of data mining with different types of data applications and challenging areas, *Journal of Engineering Research and Applications*, 4 (5), 38-41.

Chen, S. X. and J. S. Liu (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica 7*, 875–892.

Chiu, M. M. (2007). Families, economies, cultures, and science achievement in 41 countries: Country-school and student-level analyses. *Journal of Family Psychology*, *21* (3), 510-519.

Clark, P., & Boswell, R. (1991). Rule induction with CN2: Some recent improvements, *Proc. Fifth Eur. Working Session on Learning*. Springer, Berlin, 151-163.

Dekking, F. M., Kraaikamp, C., Lopuhaa, H. P. & Meester, L. E. (2005) *A modern ıntroduction to probability and statistics: understanding why and how*, United States of America: Springer Science+Business Media.

Domingos, P. (2012), A few useful things to know about machine learning, *Communications of the ACM*, *55* (10), 78–87.

Domingos, P. (2000). *A unified bias-variance decomposition and its applications,* In Proceedings of the Seventeenth International Conference on Machine Learning, 231–238, Stanford, CA: Morgan Kaufmann.

Doreswamy, H. K. (2012). Performance Evaluation of Predictive Classifiers For Knowledge Discovery From Engineering Materials Data Sets. *CIIT International Journal of Artificial Intelligent Systems and Machine Learning*, *3* (3), 162-168.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1–26.

Ehrenberg, R. (2011). Data mining finds new disease links, *Science News*, *180* (8), 15-16.

Elkan, C. (2001). *The foundations of cost-sensitive learning*. Proceedings of the 17th International Joint Conference on Artificial Intelligence, *2*, 973-978.

Engel, T. A., Charao, A. S., Kirsch-Pinheiro, M., & Steffenel, L. (2014). Performance improvement of data mining in Weka through GPU acceleration, *Procedia Computer Science*, *32*, 93–100.

Elayidom, S. M. (2012). *Design and development of data mining models for the prediction of manpower* (Unpublished Doctoral Thesis), Cochin University of Science and Technology Computer Science and Engineering, Kochi, India.

Elhamahmy, M. E., Elmahdy, H. N., & Saroit, I. A. (2010). A new approach for evaluating intrusion detection system, *CiiT International Journal of Artificial Intelligent Systems and Machine Learning*, *2* (11), 290-298.

Elomaa, T., & Kaariainen, M. (2001). An analysis of reduced error pruning, *Journal Of Artificial Intelligence Research, 15*, 163-187.

Evans, M., Hastings, N., & Peacock, B. (2000). "*Bernoulli Distribution*." Ch. 4 in Statistical Distributions, 3rd ed. New York: Wiley.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process of extracting useful knowledge from volumes of data. *Commun. ACM, 39* (11), 27–34.

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*, 3133–3181.

Freitas, A. A. (2003). A survey of evolutionary algorithms for data mining and knowledge discovery, In Tsutsui, S & Ghosh, A. (Ed.), *Advances in evolutionary computing: theory and applications,* New York: Springer-Verlag Inc.

Friedman, J. H., & Fisher N. I. (1999). Bump hunting in high-dimensional data. *Stat Comput, 9*,123–143.

Fürnkranz, J. (1999). Separate-and-Conquer Rule Learning. *Artificial Intelligence Review, 13*, 3-54.

Fürnkranz, J., & Kliegr, T. (2015). A Brief Overview of Rule Learning. In: Bassiliades, N, Gottlob, G, Sadri, F, Paschke, A, Roman, D (Eds.) In *Rule Technologies: Foundations, Tools, and Applications SE - 4. Lecture Notes in Computer Science*, 9202,54–69.

Galdi, P., & Tagliaferri, R. (2017). *Data Mining: Accuracy and Error Measures for Classification and Prediction,* in Reference Module in Life Sciences, Holland: Elsevier

Gschwind, M. (2007). Predicting Late Payments: A Study in Tenant Behavior Using Data Mining Techniques, *The Journal of Real Estate Portfolio Management*, *13* (3), 269-288.

Güldal, H., & Çakıcı, Y. (2017) Ders Yönetim Sistemi Yazılımı Kullanıcı Etkileşimlerinin Sınıflandırma Algoritmaları ile Analizi, *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 21(4),1355-1367.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*, (2nd edition),

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Hayashi, Y., Hsieh, M-H., & Setiono, R. (2009). Predicting Consumer Preference for Fast-Food Franchises: A Data Mining Approach, *The Journal of the Operational Research Society*, *60* (9), 1221-1229.

Heidelberg-Sieber, J. E. (2008). Data Mining: Knowledge Discovery for Human Research Ethics, *Journal of Empirical Research on Human Research Ethics, 3* (3), 1-2.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5, *International Journal of Advanced Computer Science and Applications*, *4* (2), 13-19.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations, *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *5* (2), 1-11.

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education, 61*, 133–145.

Imielinski, T., & Mannila, H. (1996). A database perspective on knowledge discovery. *Communications of the ACM*, *39* (11), 373-408.

Jain, J. K., Tiwari, N., & Ramaiya, M. (2013). A survey: on association rule mining, *International Journal of Engineering Research and Applications (IJERA),* 3 (1), 2065-2069.

Jiang, Y. H., Javaad, S. S., & Golab, L. (2015). Data Mining of Undergraduate Course Evaluations, *Informatics in Education*, *15* (1), 85–102.

Jiang, W., & Zhao, Y. (2014). *Some technical details on confidence intervals for LIFT measures in data mining*, Technical Report 14-02, Department of Statistics, Northwestern University.

Jung, T., Vogiatzian, F., Har-Shemesh, O., Fitzsimons, C., & Quax, R. (2014). Applying information theory to neuronal networks: from theory to experiments, *Entropy* 16, 5721-5737.

Kak, A. (2017). *Decision Trees: How to construct them and how to use them for classifying new data,* The Decision Tree Tutorial, Purdue University.

Kartal, E. E., Dogan, N., & Yıldırım, S. (2017). Exploration of the Factors Influential on the Scientific. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, *11* (1), 320-339.

Kaya, V.H. (2017). Okuma Becerilerinin Fen Bilimleri Okuryazarlığına Etkisi, *Milli Eğitim Dergisi*, *46* (215), 193-207.

Kiranmai, M., & Damodaram, A. (2014). a review on evaluation measures for data mining tasks, *International Journal Of Engineering And Computer Science*, *3* (7), 7217-7220.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, *Appears in the International Joint Conference on Articial Intelligence*, *2*, 1137-1143.

Kuonen, D. (2018). An introduction to bootstrap methods and their application, *WBL in Angewandter Statistik ETHZ 2017/19*, 1-143.

Kurgan, L. A., Cios, K. J., & Dick, S. (2006). Highly scalable and robust rule learner: performance evaluation and comparison. *IEEE Transactions on Systems, Man, and Cybernetics*, 36 (1), 32–53, Feb. 2006.

Kusiak, A. (2001). *Data Analysis: Models and Algorithms*, Proceedings of the SPIE Conference on Intelligent Systems and Advanced Manufacturing, 4191, 1-9.

Landis, J. R., & Koch, G. G. (1977) The measurement of observer aggrement ofr categorical data, *Biometrics*, *31* (3), 159-174.

Larose, D.T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* John Wiley & Sons, New Jersey.

Liaw A. & Wiener M. (2002), Classification and regression by random forest, *R News*, *2* (3), 18-22.

Li, X. B., & Sarkar, S. (2009). Against Classification Attacks: A Decision Tree Pruning Approach to Privacy Protection inData Mining, *Operations Research, 57* (6), 1496-1509.

Liu, Y., & Schumann, M. (2005). Data Mining Feature Selection for Credit Scoring Models, *The Journal of the Operational Research Society*, *56* (9), 1099-1108.

Lykourentzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V., & Loumos, V. (2009). Early and dynamic student achievement prediction in e-learning courses using neural networks, *Journal of the American Society for Information Science and Technology, 60* (2), 372–380.

Maimon, O., & Rokach, L. (2005) *Data mining and knowledge discovery handbook*, Secaucus, NJ: Springer-Verlag Inc.

Mavroforakis, C. (2011). *Data mining with WEKA*, Boston University, Retrived from http://cs-people.bu.edu/cmav/cs105/files/lab12/intro_to_weka.pdf

Mease, D., & Wyner. A. (2008) Evidence contrary to the statistical view of boosting. *The Journal of Machine Learning Research*, *9*,131–156.

MEB (2016). PISA 2015 Ulusal Ön Raporu. Ankara: MEB

Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016). Evaluating Forecasting Methods by Considering Different Accuracy Measures, *Procedia Computer Science, 95*, 264 – 271.

Metzner, J. R., & Barnes, B. H. (2014). Decision Table Languages and Systems, ACM Monograph Series, 159-168. Retrived from https://doi.org/10.1016/B978-0-12-492050-7.50011-1

Miller, M. D. (2006). *Scıence self-effıcacy ın tenth grade Hispanic female hıgh school students* (Unpublished Master's Thesis), MED University of South Florida.

Mishra, T., Kumar, D., & Gupta, S. (2014). *Mining Students' Data for Performance Prediction,* 2014 Fourth International Conference on Advanced Computing & Communication Technologies, ACCT, 255–262.

Nalepa G.J. (2008) Methodologies and Technologies for Rule-Based Systems Design and Implementation. Towards Hybrid Knowledge Engineering. In: Cotta C., Reich S., Schaefer R., Ligęza A. (eds) *Knowledge-Driven Computing. Studies in Computational Intelligence*, *102*, 183-198.

Neelamegam, S., & Ramaraj, E. (2013) Classification algorithm in Data mining: An Overview, *International Journal of P2P Network Trends and Technology (IJPTT),* 4 (8), 369-374.

Ng, A.Y. (1997). Preventing "overfitting" of cross-validation data. In D.H. Fisher (Ed.), Proceedings of the 14th International Conference on Machine Learning, Nashville, TN, USA, July 8–12, San Francisco, CA: Morgan Kaufmann.

North, M. A. (2012). *Data Mining for the Masses*, ABD: A Global Text Project Book.

Nowozin, S. (2012). *Improved information gain estimates for decision tree induction*, ICML'12 Proceedings of the 29th International Coference on International Conference on Machine Learning, July 2012, Edinburgh, Scotland.

OECD (2016). *PISA 2015 Results: Excellence and Equity in Education*, Volume I, OECD, Paris.

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques*, Heidebelrg, Berlin: Springer-Verlag Inc.

Oprea, C. (2014). Performance evaluation of the data mining classification methods, *Information society and sustainable development*, 2344, 249-253.

Paidi, A. N. (2012). Data mining: Future trends and applications*, International Journal of Modern Engineering Research (IJMER)*, *2* (6), 4657-4663.

Podgorelec, V., Kokol, P., & Rozman, I. (2002). Decision Trees: An Overview and Their Use in Medicine, Journal of Medical Systems, *26* (5), 445-463.

Qin, B. Xia, Y., Wang, S., & Du, X. (2011). A novel Bayesian classification for uncertain data, *Knowledge-Based Systems* 24,1151–1158.

Ramageri, M. B. (2010). Data mining techniques and applications, *Indian Journal of Computer Science and Engineering*, *1* (4), 301-305.

Rastogi, R., & Shim, K. (2000). A Decision Tree Classifier that Integrates Building and Pruning, *Data Mining and Knowledge Discovery*, 4, 315–344.

Refaeilzadeh, P., Tang, L., & Liu., H. (2009). *Cross Validation. In Encyclopedia of Database Systems*, (Editors: M. Tamer Ã–zsu and Ling Liu). New York, ABD: Springer publishing.

Schwenke, C., & Schering, A. (2007). *True Positives, True Negatives, False Positives, False Negatives.* New Jersey, ABD: Wiley Encyclopedia of Clinical Trials.

Seth, Y. (2017). *Bootstrapping – A Powerful Resampling Method in Statistics,* A blog on data science, Machine learning and artificial intelligence, Retrived from https://yashuseth.blog/2017/12/02/bootstrapping-a-resampling-method-in-statistics/

Sieber, J. E. (2008). Data mining: knowledge discovery for human research ethics, *J Empir Res Hum Res Ethics*, *3* (3), 1-2.

Sinha, A. P., & May, J. H. (2005) Evaluating and tuning predictive data mining models using receiver operating characteristic curves, *Journal of Management Information Systems*, *21* (3), 249-280.

Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression, *Educational Measurement: Issues and Practice*, *35* (3), 38–54.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks, *Information Processing and Management*, *45*, 427-437.

Souza, J., Matwin, S., & Japkowicz, N. (2002). *Evaluating data mining models: a pattern language.* In: 9th Conference on Pattern Language of Programs (PLOP'02), Monticello, Illinois, 8–12 September 2002.

Stefanowski, J. (2008). *Association rules*, Institute of Computing Sciences Poznan University of Technology Poznan, Poland, Lecture 10, SE Master Course.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*, 323–348.

Svetnik, V., Liaw, A., Tong, C., & Wang, T. (2004). Application of Breimans random forest to modeling structure-activity relationships of pharmaceutical molecules. In F. Roli, J. Kittler, & T. Windeatt (Eds.), *Multiple classifier systems* (vol. 3077, pp. 334–343). Cagliari, Italy: Springer.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson/Allyn & Bacon.

Tadesse, T., Wilhite, D. A., Hayes, M. J., Harms S. K., & Goddard, S. (2005). Discovering Associations between Climatic and Oceanic Parameters to

Monitor Drought in Nebraska Using Data-Mining Techniques, *Journal of Clımate, 18* (10), 1541-1550.

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining.* Boston, USA: Addison-Wesley Longman Publishing Co.

Tan, P. N., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to data mining* (2nd ed.), USA: Pearson.

Thuarisingham, B.M., (2003). *Web Data Mining and Applications in Business Intelligence and Counter Terrorism,* USA: CRC Press LLC, Boca Raton, FL.

Tiwari, M., Jha, M. B., & Yadav, O. (2012). Performance analysis of Data Mining algorithms in Weka, *IOSR Journal of Computer Engineering (IOSRJCE),* 6 (3), 32-41.

Tollenaar, N., & Heijden, P. G. M. (2013) Which method predicts recidivism best?: a comparison of statistical, machine learning anddata mining predictive models, *Journal of the Royal Statistical Society*, *176* (2), 565-584.

Torgo, L. (2016). *Data Mining with R: Learning with Case Studies* (2nd ed.), Florida: Chapman & Hall Book.

Usta, H. G. ve Çıkrıkçı-Demirtaşlı, N. (2014). PISA 2006 Sınavı Sonuçlarına Göre Türkiye'deki Öğrencilerin Fen Bilimleri Okuryazarlığını Etkileyen Duyuşsal Faktörler, *Eğitim Bilimleri Araştırmaları Dergisi*, *4* (2), 93-107.

Vaitsis C, Hervatis V., & Zary N. (2016). *Introduction to Big Data in Education and Its Contribution to the Quality Improvement Processes*. In: Ventura SS (ed). Big Data Real-World Appl. InTech: 41–64.

Vanwinckelen, G., & Blockeel, H. (2012) *On estimating model accuracy with repeated cross-validation*, Belgian-Dutch Conference on Machine Learning (BeneLearn) edition:21 location, 24-25 May 2012.

Weiss, S. M., & Kulikowski, C. A. (1991). *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Mateo, CA: Morgan Kaufmann.

Williams, G. (2010). *Data mining desktop survival guide: Cross validation*, Retrived from https://www.togaware.com/datamining/survivor/Cross_Validation.html

Williams, G. (2011) *Data mining with Rattle and R: The art of excavating data for knowledge discovery*, New York, USA: Springer Science+Business Media

Witten, I. H., & Frank, E. (2005) *Data minig: Practical machine learning tools and techniques*, United States of America: Morgan Kaufmann publications

Witten, I. H., Frank, E., & Hall, M. (2016) *Data minig: Practical machine learning tools and techniques*, United States of America: Morgan Kaufmann publications

Wolf, G. (2016). *Introduction to Data Mining: Bayesian Classification*, Yale University, 1-21.

Wu, X. et al. (2008). Top 10 algorithms in data mining, *Knowl Inf Syst*, 14, 1-37.

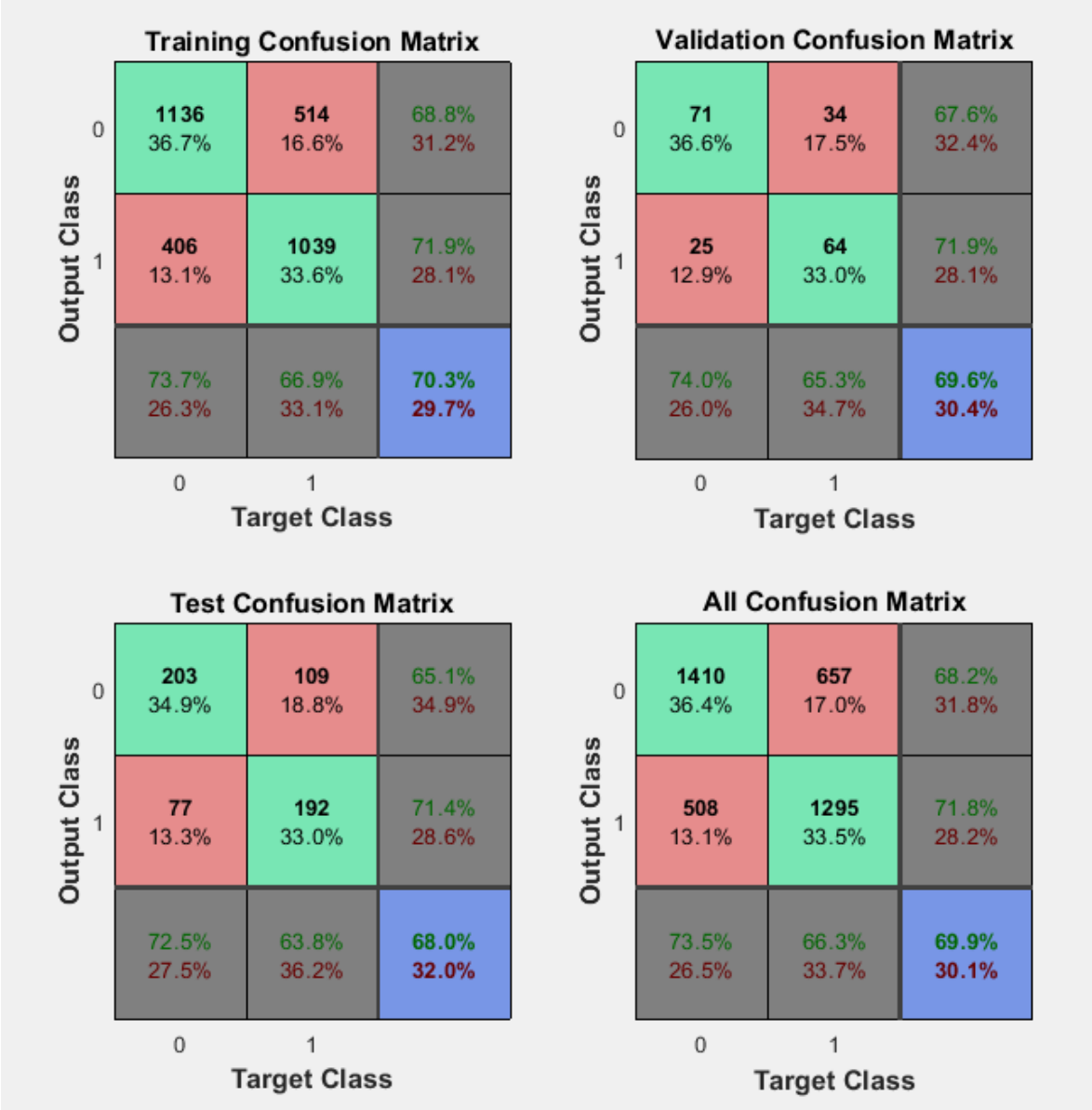
Yin, X., & Han, J. (2003) *CPAR: Classification based on predictive association rule*, In Proceedings of the SIAM International Conference on Data Mining, San Francisco, CA: SIAM Press, 369–376.

Yung, J. L., Hsu, Y. C., & Rice, K. (2012). Integrating Data Mining in Program Evaluation of K-12 Online Education, *Journal of Educational Technology & Society*, *15* (3), 27-41.

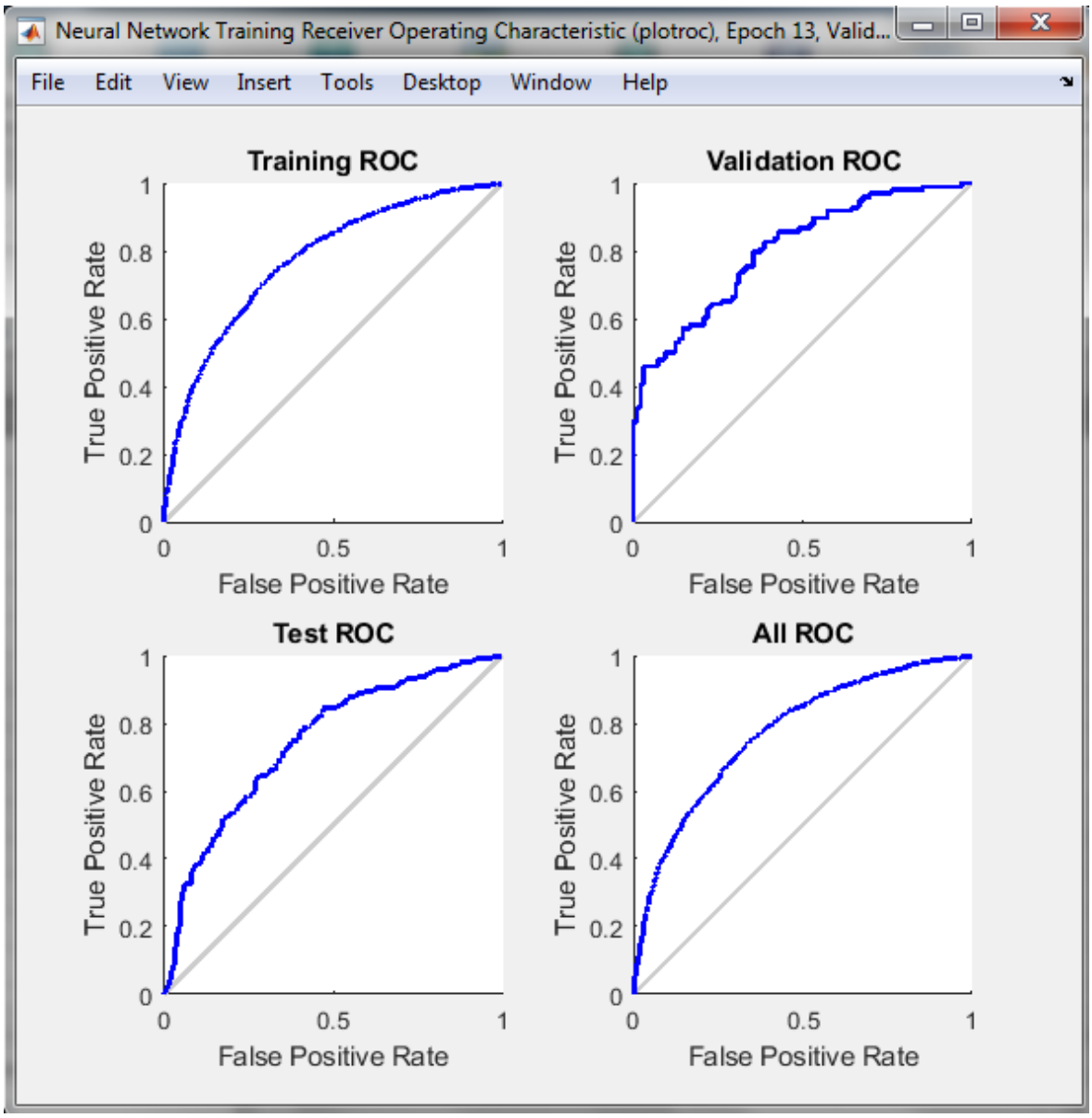
Zhang, M., Hong, Y., & Balakrishnan, N. (2017). *An algorithm for computing the distribution function of the generalized poisson-binomial distribution*, arXiv preprint stat.CO,1-14.

Zhou, B., & Cervantes, C. (2016) Decision Trees, *CS 446 Machine Learning Fall 2016*, 8, 1-15. Retrived from http://l2r.cs.uiuc.edu/Teaching/CS446-17/LectureNotesNew/dtree/main.pdf

EK-A: Matlab'tan Elde Edilen Karışıklık Matrisi



EK-B: Matlab'tan Elde Edilen ROC Eğrisi



EK-C: Etik Komisyonu Onay Bildirimi

Form: 40

Tez Çalışması Etik Komisyon İzin Muafiyeti Formu

09/02/2017

Hacettepe Üniversitesi
Eğitim Bilimleri Enstitüsü
Eğitimde Ölçme ve Değerlendirme Anabilim Dalı Başkanlığı'na

Tez Başlığı / Konusu:	Lise Öğrencilerinin YGS Başarılarını Kestirmede Veri Madenciliği Yöntemlerinin Karşılaştırılması

Yukarıda başlığı/konusu gösterilen tez çalışmam:

1. İnsan ve hayvan üzerinde deney niteliği taşımamaktadır,
2. Biyolojik materyal (kan, idrar vb. biyolojik sıvılar ve numuneler) kullanılmasını gerektirmemektedir.
3. Beden bütünlüğüne müdahale içermemektedir.
4. Gözlemsel ve betimsel araştırma (anket, ölçek/skala çalışmaları, dosya taramaları, veri kaynakları taraması, sistem-model geliştirme çalışmaları) niteliğinde değildir.

Hacettepe Üniversitesi Etik Kurullar ve Komisyonlarının Yönergelerini inceledim ve bunlara göre tez çalışmamın yürütülebilmesi için herhangi bir Etik Komisyondan/Kuruldan izin alınmasına gerek olmadığını; aksi durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Gökhan AKSU
(Öğrencinin Adı Soyadı, İmzası)

Öğrenci Bilgileri

Adı Soyadı	**Gökhan AKSU**
Öğrenci No	N13248472
Anabilim Dalı	Eğitim Bilimleri
Programı	Eğitimde Ölçme ve Değerlendirme
Statüsü	☐ Yüksek Lisans ☒ Doktora ☐ Bütünleşik Dr.

Danışman Görüşü ve Onayı

Uygundur.

Doç. Dr. Nuri DOĞAN
(İmza)
(Danışmanın Ünvanı, Adı ve Soyadı)

EK-Ç: Etik Beyanı

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı bütün bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin bütününü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

13/04/2018

(İmza)
Gökhan AKSU

EK-D: Doktora Tez Çalışması Orijinallik Raporu

13 / 04 / 2018

HACETTEPE ÜNİVERSİTESİ
Eğitim Bilimleri Enstitüsü
Eğitim Bilimleri Ana Bilim Dalı Başkanlığına,

Tez Başlığı : PISA Başarısını Tahmin Etmede Kullanılan Veri Madenciliği Yöntemlerinin İncelenmesi

Yukarıda başlığı verilen tez çalışmamın tamamı (kapak sayfası, özetler, ana bölümler, kaynakça) aşağıdaki filtreler kullanılarak **Turnitin** adlı intihal programı aracılığı ile kontrol edilmiştir. Kontrol sonucunda aşağıdaki veriler elde edilmiştir:

Rapor Tarihi	Sayfa Sayısı	Karakter Sayısı	Savunma Tarihi	Benzerlik Oranı	Gönderim Numarası
05 / 04 / 2018	127	208.372	13 / 04 / 2018	%5	941639800

Uygulanan filtreler:

1. Kaynaklar hariç
2. Alıntılar dâhil
3. 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan eder, gereğini saygılarımla arz ederim.

Ad Soyadı: Gökhan AKSU
Öğrenci No.: N13248472
Ana Bilim Dalı: Eğitim Bilimleri
Programı: Eğitimde Ölçme ve Değerlendirme
Statüsü: ☐ Y.Lisans ☒ Doktora ☐ Bütünleşik Dr.

İmza

DANIŞMAN ONAYI

UYGUNDUR.
Doç. Dr. Nuri DOĞAN

EK-E: Dissertation Originality Report

13 / 04 / 2018

HACETTEPE UNIVERSITY
Graduate School Of Educational Sciences
To The Department Of Educational Sciences

Thesis Title : Investigation of Data Mining Methods Used for Estimating PISA Success

The whole thesis that includes the *title page, introduction, main chapters, conclusions and bibliography section* is checked by using **Turnitin** plagiarism detection software take into the consideration requested filtering options. According to the originality report obtained data are as below.

Time Submitted	Page Count	Character Count	Date of Thesis Defense	Similarity Index	Submission ID
05 / 04 / 2018	127	208.372	13 / 04 / 2018	5%	941639800

Filtering options applied:

1. Bibliography excluded
2. Quotes included
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Educational Sciences Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Name Lastname: Gökhan AKSU
Student No.: N13248472
Department: Educational Sciences
Program: Measurement and Evaluation in Education
Status: ☐ Masters ☒ Ph.D. ☐ Integrated Ph.D.

Signature

ADVISOR APPROVAL

APPROVED
Doç. Dr. Nuri DOĞAN

EK-F: Yayımlama ve Fikrî Mülkiyet Hakları Beyanı

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversite'ye verilen kullanım hakları dışındaki bütün fikrî mülkiyet haklarım bende kalacak, tezimin tamamının veya bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinleri yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversite'ye teslim etmeyi taahhüt ederim.

☐ **Tezimin/Raporumun tamamı dünya çapında erişime açılabilir ve bir kısmı veya tamamının fotokopisi alınabilir.**

(Bu seçenekle teziniz arama motorlarında indekslenebilecek, daha sonra tezinizin erişim statüsünün değiştirilmesini talep etseniz ve kütüphane bu talebinizi yerine getirse bile, teziniz arama motorlarının ön belleklerinde kalmaya devam edebilecektir)

☒ **Tezimin/Raporumun 01.01.2019 tarihine kadar erişime açılmasını ve fotokopi alınmasını (İç Kapak, Özet, İçindekiler ve Kaynakça hariç) istemiyorum.**

(Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir, kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisi alınabilir).

☐ **Tezimin/Raporumun ..... tarihine kadar erişime açılmasını istemiyorum ancak kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisinin alınmasını onaylıyorum.**

☐ **Serbest Seçenek/Yazarın Seçimi:**

.....

.....

.....

13 / 04 / 2018

(İmza)

Gökhan AKSU